# Regulating Cryptocurrencies: A Supervised Machine Learning Approach to De-Anonymizing the Bitcoin Blockchain

HAO HUA SUN YIN, KLAUS LANGENHELDT, MIKKEL HARLEV, RAGHAVA RAO MUKKAMALA, AND RAVI VATRAPU

Hao Hua Sun Yin (awasunyin@gmail.com) is a Research Associate at the Center for Business Data Analytics of the Department of Digitalization at the Copenhagen Business School and a co-founder of Cryptium Labs GmbH, a digital and secure attestations as a service start-up for Proof-of-Stake blockchains such as Tezos, Cøsmos, and Polkadot. She worked previously as a data scientist and software engineer at Chainalysis, as a researcher at Cøsmos, and as a grant manager at the Ethereum Community Fund. Her research interests are in blockchain technology, applied cryptography for privacy and scalability of consensus algorithms, security in distributed and decentralized systems, and hardware, specifically hardware security modules. Her current research focuses on a generic mapping of the Blockchain ecosystem's latest state, in terms of development and research. She holds a Master's in Information Systems from the Copenhagen Business School (CBS).

Klaus Langenheldt (kl.digi@cbs.dk) is an Industrial Ph.D. Student at the Center for Business Data Analytics of the Department of Digitalization at CBS. He holds a M.Sc. in Business Administration and Information Systems (e-business) from that school, where he conducts research on de-anonymizing the Bitcoin Blockchain using supervised machine learning. In addition to his academic work, he brings extensive work experience from the IT sector, working—among others—as Business Developer for Europe's largest incubator Rocket Internet and as Software Engineer for Danske Bank.

Mikkel Harlev (miharlev@gmail.com) is a Research Associate at the Center for Business Data Analytics of the Department of Digitalization at CBS. He holds a M.S.c degree in Business Administration and Information Systems from CBS. He is currently employed as a software Engineer in the largest Danish bank, specializes in Data Science with a focus on its application in the financial sector, and conducts research on Bitcoin blockchain forensics and machine learning.

Raghava Rao Mukkamala (rrm.digi@cbs.dk) is an Associate Professor at the Center for Business Data Analytics of the Department of Digitalization at CBS and an associate professor at the Department of Technology, Kristiania University College, Denmark. He holds a Ph.D. degree from University of Copenhagen. Dr. Mukkamala's research focuses on computational social science using an inter-disciplinary approach, combining formal modelling approaches with data-mining /machine-learning techniques, for modelling of social science phenomena in digital transformation of organizations and society. He works with blockchain-based

*Journal of Management Information Systems* / 2019, Vol. 36, No. 1, pp. 37–73.

ISSN 0742–1222 (print) / ISSN 1557–928X (online)
DOI: https://doi.org/10.1080/07421222.2018.1550550

technologies for social business and Internet of Things. His research work is complemented by over ten years of professional experience in the Danish IT industry as a senior software engineer and consultant.

Ravi Vatrapu (vatrapu@cbs.dk; corresponding author) is a Professor of Computational Social Science at the Department of Digitalization, Copenhagen Business School; professor of applied computing at the Kristiania University College; and director of the Center for Business Data Analytics at CBS. He holds a Ph.D. in Communication and Information Sciences from the University of Hawaii at Manoa. Dr. Vatrapu's research focus is on big social data analytics to design, develop, and evaluate a new holistic approach to computational social science, Social Set Analytics. He is the author of numerous publications in these domains.

Abstract: Bitcoin is a cryptocurrency whose transactions are recorded on a distributed, openly accessible ledger. On the Bitcoin Blockchain, an owning 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 de-anonymizing the Bitcoin Blockchain by using Supervised Machine Learning to predict the type of yet-unidentified entities. We utilized a sample of 957 entities (with ≈385 million transactions), whose identity and type had been revealed, as training set data and built classifiers differentiating among 12 categories. Our main finding is that we can indeed predict the type of a yet-unidentified entity. Using the Gradient Boosting algorithm with default parameters, we achieve a mean cross-validation accuracy of 80.42% and F1-score of ≈79.64%. We show two examples, one where we predict on a set of 22 clusters that are suspected to be related to cybercriminal activities, and another where we classify 153,293 clusters to provide an estimation of the activity on the Bitcoin ecosystem. We discuss the potential applications of our method for organizational regulation and compliance, societal implications, outline study limitations, and propose future research directions. A prototype implementation of our method for organizational use is included in the appendix.

Key words and phrases: cryptocurrencies, Bitcoin, blockchain, cybersecurity, supervised machine learning, online anonymity, cybercrime.

## Introduction

Cryptocurrencies are digital assets that use a decentralized control system and cryptography to facilitate, secure, and verify transactions and create additional assets [84]. Bitcoin is a type of cryptocurrency that was first described in 2008 [83]. Recently, cryptocurrencies in general and Bitcoin in particular have attracted increased attention of the researchers from diverse academic fields [6, 24, 59], as well as practitioners due to its unique characteristics such as the absence of centralized control, safeguards against equivocation and 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 [32] for an analysis of

the Silk Road). Moreover, recent articles and reports [54, 74, 75] have stated that Bitcoin has been used for terror financing, thefts, scams, and ransomware. 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 Bitcoin [6, 24, 59].

Our paper aims to better understand the different kinds of transactions in the Bitcoin ecosystem in order to better inform managerial and organizational aspects of regulation and compliance. We do so by developing a novel approach based on Supervised Machine Learning to de-anonymizing the Bitcoin ecosystem to help identify high-risk counterparties and potential cybercriminal activities. For organizations, 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 could be technically infeasible, according to popular belief about the robustness of anonymity in the Bitcoin ecosystem. However, previous research [78, 92] 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. Furthermore, prior research in the management information systems domain on cyber threat intelligence has proposed several approaches to de-anonymization of different entity types [1, 2, 18, 69, 98].

Our work builds upon and extends this area of research and addresses the call for research that aids better regulation of cryptocurrencies in general and Bitcoin in particular [5, 63, 71]. Knowing that Bitcoin addresses can be clustered, identified, and categorized, we investigate the true level of Bitcoin's anonymity to determine the extent to which it is possible to reveal the identity of users or organizations in the Bitcoin ecosystem by employing a Supervised Machine Learning approach.

## Problem Formulation and Research Question

For this research paper, we collaborated with the blockchain analysis company *Chainalysis* [28], which will be referred to as the *data provider* in the remainder of the paper. The data provider has clustered, identified, and categorized a substantial number of Bitcoin addresses manually or through a variety of clustering techniques (see the section Clustering Concepts). However, the vast majority of clusters on the Bitcoin Blockchain remain uncategorized. 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 recognize the fact that there are additional cluster types participating in the Bitcoin ecosystem, but the scope of this research will be limited to the aforementioned categories as those are the categories provided by the data provider to regulatory

authorities and organizational users. At the time of writing, to the best of our knowledge, there have not yet been any research publications utilizing such rich data set in conjunction with Supervised Machine Learning. Furthermore, alternative data sources, where the labelled data set has as many identified clusters as the data provider's, remain unknown. Based on the aforementioned discussion, our overarching main research question is stated as follows:

> *Research Question: How can one estimate the extent of cybercriminal and illicit activities in the Bitcoin ecosystem by uncovering pseudonymity of Bitcoin blockchain technology?*

The remainder of the paper is organized as follows: in the section to follow, a brief overview of related work is provided and section on Conceptual Framework describes key concepts with regard to cryptocurrencies, Bitcoin, and Blockchain. Methodological details are discussed in Methodology section, and The Results Section presents an overview of the results and provides technical interpretations. Finally, in the Discussion section, we provide a substantive interpretation of our results, discuss organizational implications, address limitation of our current work, and outline future work directions.

## Related Work

In this section, we will review the existing state-of-the-art related to our work on de-anonymizing entities in the context of cryptocurrencies. First, we will provide an overview of related research from the information systems perspective, and then we summarize various initiatives taken up by legal communities toward regulatory framework for cryptocurrencies. Finally, we will also provide a concise description of current state-of-the-art de-anonymizing entities in cryptocurrencies and conclude by stating how our approach is different from the existing research.

## Information Systems Perspective

From an Information Systems (IS) perspective, the state- of-the-art that is relevant to our work is organized into two parts as shown in Appendix-1, Table 1. The first part (Cyber Threat Intelligence) is methodological and discusses the extant literature on cyber threat intelligence; whereas, the second part, which is conceptual, discusses existing research on blockchain and cryptocurrencies and is presented in section after next.

### Cyber Threat Intelligence

In this subsection, we review related research work related to anonymity, identifying the fraudulent traders, identifying the cybercriminal behavior, and financial fraud

Table 1. Raw and Pre-processed Data Size

| | Observations | Raw | Preprocessed |
|---|---|---|---|
| Labelled Data | 957 | 85 GB | 1.4 MB |
| Uncategorized Data | 153,293 | 157 GB | 97 MB |

detection in the context of electronic markets and commerce channels from the management information systems journals and position our work in the context of existing research. Results are presented in part one of Appendix 1, Table 1. First, many studies [1, 2, 18, 69, 98] utilized machine learning approaches and text mining methods on the contextual information. Contextual information consisted of customer reviews, participant messages, organization or industry specific information to perform various kinds of sentiment analysis, and stylometric analysis to derive indicators or clues about fraudulent behavior and cybercriminal activities that identify organizations and individuals who resort to such types of malpractices. Second, identifying fake websites and phishing websites is another important research direction where genre tree kernel methods with fraud cues, statistical learning theory, and classification-based methods were employed to identify the differences in the characteristics between legitimate and phishing/fake websites, such that fraudulent behaviour can be identified among the online web applications [3, 4]. Third, from the Information Systems point of view, several research works [1, 3, 4, 70, 109] predominantly employed Design Science theory as a way to validate or evaluate their researched phenomenon or prototype or methodology. Similar to the aforementioned research, our method also uses machine learning approaches to de-anonymize suspicious entities (such as ransomware or darknet market) who are involved in fraudulent or cybercriminal activities. Finally, in the context of cryptocurrency markets and in relation to cyber threat intelligence, in terms of organizational implications of our research, the focus of banks and financial institutions will be more on preventing or block-listing transactions with suspicious entities. Therefore, the findings of the research work [11] recommending the usage of Trusted Third Party (TTP) in e-commerce markets is quite relevant. For cryptocurrencies, our method of de-anonymizing the suspicious entities or organizations will be quite instrumental in the hands of governmental organizations or some trusted financial institutions if they act as a TTP in the domain of cryptocurrencies.

## Blockchain and Cryptocurrencies

Within the information systems discipline, current research on blockchain technologies and cryptocurrencies is still in the nascent stage, maybe because the blockchain based technologies are still considered as an emergent phenomenon. Therefore, we were not able to find many research articles from the primary information systems journals with respect to blockchain technologies and cryptocurrencies. However, we gathered all the recent research papers from major information systems conferences and journals and

the summary of review results is presented in part II of the Appendix 1, Table 1. First, the most notable research work on blockchain based technologies is Beck and Beck et al. [15, 17], which forecasts that in the near future the blockchain technologies will empower organizations to implement solutions using distributed ledger technologies. These technologies will handle contracts and transactions among the organizations in a decentralized manner without any need of having their own legal entities and will lead to the emergence of decentralized autonomous organizations. Second, since information systems research on the blockchain technology is in the nascent stage, several research frameworks [17, 86, 94] were proposed to study organizational adoption challenges and IT governance, for example, in terms of decision rights, accountability, and incentives for the organizations, which can reap benefits from decentralized solutions using these technologies. Development of the proof of concept prototypes for blockchain technologies using design science guidelines [16, 31, 56, 82] is also an increasing trend in recent years. Finally the research on the cryptocurrencies per se is rather limited [48, 85], when compared to the more general research focus on blockchain based applications for organizations.

## Legal and Regulatory Perspective

Blockchain-based technologies have attracted significant attention from researchers in the area of law, especially on the topics of and aspects in regulation and governance of cryptocurrencies and blockchain based applications such as smart contracts. The relevant extant literature research on the legal aspects of Blockchain regulation, compliance, and governance is summarized in detail in Appendix 1, Table 2.

As shown in Appendix 1, Table 2, we characterize the implications from current research on blockchain regulation into two distinct and opposing research streams: stringent versus open-minded regulation. First, the stream of research studies [5, 63, 71] that is primarily concerned with money laundering and digital crime using cryptocurrencies and their economic and social impacts on societies argues for establishing clear and stringent regulations, compliance protocols, and guidance frameworks for the cryptocurrency industry. On the other hand, a significant amount of research [35, 62, 76, 80, 104, 107, 108] posits that blockchain is a highly disruptive and innovative technology and argues for a more open-minded regulation without attempting to stop or slow its growth with suggestions to amend the existing legal provisions if necessary to create a workable regulatory model. Moreover, it is argued that a proper regulatory model that does not constrain the innovation of cryptocurrencies will allow them to self-regulate within a vaguely defined regulatory framework. At the same time, a proper regulatory model that uncovers the actors in the case of necessities (e.g., money laundering) will help the cryptocurrencies to get rid of their infamous reputation and potentially revolutionize organizations.

Apart from cryptocurrencies such as Bitcoin, prior research also focused is on the regulation and compliance in terms of using blockchain technology for

Table 2. Distribution of Cluster Categories

| Category Label | Cluster observations |
|---|---|
| darknet-market | 46 |
| exchange | 306 |
| gambling | 102 |
| hosted-wallet | 11 |
| merchant-services | 17 |
| mining-pool | 67 |
| mixing | 10 |
| other | 57 |
| personal-wallet | 293 |
| ransomware | 21 |
| scam | 23 |
| stolen-bitcoins | 4 |
| Total observations | 957 |

various applications such as: digital-asset transfers [62], property rights [9], cryptosecurities [67], derivatives markets [106], smart contracts [102], and so on for financial, accounting, and other administrative domains. Unlike the case of cryptocurrencies, the research from the law disciplines argued for usage of blockchain technology for developing applications in these areas, as it would enhance transparency in these application areas by removing hidden secrecy and provides a way for more efficient document and authorship verification, title transfers, and contract enforcement. Finally, just to provide an example of the scope of research regarding blockchain regulation and governance, Young [115] advocated for smart constitution, a blockchain based implementation for governance, which will make government operate in compliance to smart constitution laws in a visible manner and also prohibits operation outside of its mandate.

Positioning this paper in the extant literature on regulation and compliance of cryptocurrencies previously discussed, our method helps unmask fraudulent and criminal actors and could be instrumental for trusted third party providers, such as governmental agencies or regulators, to implement flexible regulatory and compliance measures for cryptocurrencies, without burdening law-abiding citizens who transact in cryptocurrencies.

## Anonymity Perspective

Bitcoin allows end-users to create (pseudo-)anonymous financial transactions without the need for disclosing their personal information. This is done by generating a pseudonym for the user, also called "address." The apparent anonymity and ease to create pseudo-anonymous financial transactions attracted users who value their

privacy on one hand and, on the other hand, it has also attracted cybercriminals who want to use it for ransom-ware and other illegal activities [21, 64, 105]. Therefore, analyzing the pseudo-anonymity and understanding the traceability of Bitcoin flows to investigate the use of it for criminal or fraudulent purposes is of high academic importance as well as practical relevance. Some of the focus areas of research in this direction are applying heuristic approaches or statistical methods [7, 37, 64, 78, 92, 95, 105]. For instance, clustering the Bitcoin addresses by mapping the network, analyzing the traffic and complementing it with external sources of information was explored in Reid and Harrigan [92] using appropriate representation of two networks derived from the transaction history of Bitcoin. In a similar direction, heuristic clustering was employed to group Bitcoin wallets based on evidence of shared spending authority and using categories that are labelled from the interactions with various services to characterize longitudinal changes in the Bitcoin market [78]. Statistical properties of the Bitcoin transaction graph were analyzed in Fleder et al. [37] and Ron and Shamir [95] to identify behavioral patterns of different types of users. One of the interesting findings is that the majority of the minted bitcoins remain inactive and hidden in addresses that never participated in any outgoing transactions.

Privacy guarantees of Bitcoin were analyzed using simulation experiments by replicating the behaviors and transactions of the Bitcoin Blockchain and showed that it is possible to uncover almost 40% of the users' profiles even after adopting the Bitcoin's recommended privacy measures by the users [7]. By observing real-time transaction relay traffic over a period of time, the research in Koshy et al. [64] showed that it is possible to map Bitcoin addresses directly to IP data; thereby, it is possible to reveal the ownership behind the Bitcoin addresses. Alternately, the research in Goldfeder et al. [50] showed that if a user paid on shopping websites using a cryptocurrency then a third party tracker can link the transaction information to the user's cookie and then further to link it to the user's real identity. Furthermore, they also identified that, if a third party tracker is able to link two such online purchases from the same user onto the blockchain, then it possible to identify the entire cluster of addresses and transactions, even if the users employed blockchain anonymity techniques to hide their identity.

Many researchers also focused on the flaws of the Bitcoin Blockchain and explored alternative cryptocurrencies as well as proposals for improvements and/or new methods to bring anonymity to its users. Some of the research explored in-depth investigation on Bitcoin's technological workings, showing its technological flaws and consequent suggestions on how to address them [14], a protocol that enables anonymous payments in Bitcoin and other currencies that relies on technology commonly used by mixing services [25]. In this regard, an important research contribution is the development of an alternative to Bitcoin named Zerocash with zero-knowledge proofs [99] and also privacy-enhancing overlays in Bitcoin from a theoretical perspective [77].

For the majority of previous research, researchers collected the Blockchain data on their own, crafted their own categories, and then attempted to de-anonymize entity types. In contrast to the existing research, since the data provider supplied us with a rich data set that is already clustered, categorized, and identified addresses for forensic purposes, our analysis approach focused on utilizing Supervised Machine Learning methods to categorize the yet-unidentified entities. To the best of our knowledge, there is no other research work that focused on de-anonymizing the Bitcoin Blockchain using Supervised Machine Learning techniques. Even though there is related work where the authors collect raw data, perform their own clustering, and use it as input for a classification problem, our research seminally leverages data that had been previously enriched via co-spend heuristics, intelligence-based and behavioral clustering (see Clustering Methodology), resulting in a dataset that has the highest coverage of labelled entities at the time of writing.

## Conceptual Framework

This section describes the key concepts that inform our empirical wok on de-anonymizing the Bitcoin Blockchain using Supervised Machine Learning. First, we present basic concepts and concise discussion on the anonymity and identity in the context of decentralized networks such as blockchain. Second, we present the basic foundational concepts of blockchain technologies followed by the technological working of the cryptocurrencies. Third, we present an overview of clustering techniques applied to cluster Bitcoin addresses and then, finally, the basic concepts behind different supervised machine-learning algorithms used in this work will be presented.

### Anonymity

In general, the notion of anonymity can be defined as a means to obtain "freedom from identification, secrecy, and lack of distinction" [100, p. 875], which can be further characterized as a phenomenon where one can conceal his identity from other parties [26]. With the advent of the Internet and subsequent developments in electronic commerce, communications, and social media, as well as the innovations such as Web 2.0, there is an ever growing discourse on anonymity, especially both in favor of and against the anonymity on the online environments. The arguments in favor of anonymity perceive the online anonymity as a *necessary tool* to preserve *information privacy*, by protecting confidential information from untrusted platforms and parties [26]. However, anonymity is frequently abused and creates an environment for hate speech and defamatory remarks by individuals who behave irresponsibly with impunity [68, 100].

With the evolution of public-key cryptography and software agents, such as anonymous remailer servers in the 1990s, communication over the Internet can

be created with a high-degree of certainty such that the identity of the originator of the communication can be concealed [40]. These techniques paved the way for creation of pseudonymous entities in the Internet communication, using which messages can be both sent and received by still concealing the identity of the originator. Pseudonymity differs from anonymity in the sense that anonymity requires removal of all identity information, whereas pseudonymity still allows for creation and continuation of a pseudo/alternate identity, which allows for partial concealment of the real identity information [42, 100]. As proposed by Froomkin in the mid-1990s [26, 40, 41], anonymity/pseudonymity in the context of the Internet and electronic communication can be distinguished as four different types.

- **Traceable anonymity**: In traceable anonymity, the recipient of the communication does not have knowledge regarding the identity of the sender. The sender's information is only available to the agent/system that acts as an intermediary in the communication. Traceable anonymity is sufficient for many general-purpose scenarios (such as posting messages to newsgroups), even though it offers the lowest security out of all the four categories.
- **Untraceable anonymity**: The identity of a sender in a communication is not at all identifiable in the case of untraceable anonymity. For example, one could achieve this sort of anonymity by routing the communication through a series of anonymous remailers (e.g., chained remailing [40]) and using the existing encryption techniques. Cryptocurrencies follow a similar approach (e.g., Bitcoin uses *mixing*, see sec. Cluster Categories) to reduce the traceability of the transactions and to complicate the transaction analysis.
- **Untraceable pseudonymity**: In the case of untraceable pseudonymity, the sender of a communication will communicate using a pseudonym, which is not identifiable. In comparison to the untraceable anonymity, due to usage of pseudonym in untraceable pseudonymity, the continuation of this pseudo-identity is maintained over a period of time and, thereby, builds an image and a reputation just like in the case of any other online personality or digital persona. As an example, *Satoshi Nakamoto* is a pseudonym representing a person or a few people, who started Bitcoin as a peer-to-peer electronic cash system in 2013. The real identity of the person/people who own this pseudonym is still not known today, but this pseudonym continues to build/maintain its digital profile as the one who introduced Bitcoin cryptocurrency. Normally, pseudonyms are used consistently over a period of time, whereas anonymous identities are used only once. Asymmetric cryptography is quite instrumental in maintaining the pseudonyms by facilitating transmission of signed messages (signed transactions in case of cryptocurrencies) in the name of pseudonym. The signed messages or digital signatures can neither be forged nor linked to the true identities (as long as the keys are not linked to the real identities).
- **Traceable pseudonymity**: Unlike the untraceable pseudonymity, it is possible to trace the pseudonym of the sender in a communication, not necessarily

> by the receiver of the messages, but by an intermediary or a third party, which might have assigned that pseudonym. Moreover, a distinction can be made in traceable pseudonymity based on whether the pseudonym was issued formally by a third-party provider/intermediary or if the pseudonym is self-chosen by the holder of the pseudonym.

Anonymity in electronic communication is an empowering technology. Being anonymous gives the power to users to do things without being identifiable, which could lead to the development of enhanced societies and flourishing human communications. Among the four types of anonymity previously explained, untraceable anonymity provides the highest level of protection and it allows communication without fear of jail, harm, or other retaliation [41, 43].

> In places that are less free, avoiding retribution for saying the wrong thing may be a matter of life and death. Political dissidents, ethnic minorities, religious splinter groups, people campaigning for women's rights or gay rights, and many others are, or have been, subject to the risk of genuine and very palpable violence. If they wish to speak or write for their causes they need a means to protect themselves. Anonymity is one such tool [43, p. 121].

In Group Support Systems (GSS) research, anonymity is considered a fundamental concept, which is expected to reduce fear of social disapproval and evaluation, as well as enhance participation in group work and, thereby, not only lead to an increase in the number of ideas generated, but also leads to improvement in the quality of decisions [90]. Profiling is another major phenomenon where anonymity is a tool. Profiling became a basis for stereotypical discrimination based on characteristics such as ethnic backgrounds, sexual orientation, political opinion, among others, and especially for commercial entities who profile users for marketing purposes to exercise their market power through price discrimination. Anonymity, however, can be used as a tool to protect against profiling [36]. According to Froomkin,

> … digital anonymity may be a rational response to a world in which the quantity of identifying data on each of us grows daily, and the data become ever easier for government and private parties to access [40, par. 50].

## Cryptocurrencies and the Dark Side of Anonymity

In contrast to the arguments in favor of anonymity, untraceable anonymous/pseudonymous communication opens doors for many illegal and criminal activities. The introduction of anonymous and untraceable communication has also led to a wide range of interpersonal transactions that cannot be easily traced. Especially in the context of cryptocurrencies, cybercriminal activities such as money laundering, extortion, blackmail, ransomware, drug, and other illegal activities have paved the way for the dark side of anonymity [27].

In the context of decentralized peer-to-peer networks, blockchain and cryptocurrencies use traceable pseudonymity techniques to perform communication and transactions. For example, cryptocurrencies use a randomly generated pair of keys to perform transactions to have traceable pseudonymity with self-selection of keys by the users, where one of the keys acts as a pseudonym for the user and the other key is used to sign the transactions. In fact, Bitcoin, as a best practice, advocates their users generate a new key pair (i.e., new pseudonym) for each transaction to protect the privacy of their transactions [20]. As it is difficult to link the pseudonyms to the real identities, the pseudonymity in cryptocurrencies paved the way for criminal and illegal activities on one hand and also made it difficult to implement regulatory measures for anti-money laundering [49, 81]. Particularly during the recent years, Bitcoin offered opportunities for fraud and tax evasion [103], a money laundering route for cybercriminals [49], drug dealing on the dark web and *Silk Roads* [110]. At the same time, Bitcoin and other cryptocurrencies also have legitimate users who wish to transact using the pseudonymity features, and their aspirations should not be bundled with any association with criminal and illegal activities regarding cryptocurrencies [108].

Even though it is not possible to directly link the pseudonyms with real identities, it is possible to profile the pseudonyms and their connected transactions to some extent based on their transactional behavior that is globally visible, using various techniques described in sec. Anonymity Perspective. In the context of anonymity, our work is primarily focused on developing techniques for profiling and de-anonymizing the entities and their connected pseudonyms that are linked to the cybercriminal and illegal activities, rather than identifying the legitimate users of cryptocurrencies, who would like to use the pseudonymity features of cryptocurrencies in a genuine manner.

## Blockchain and Cryptocurrencies

The blockchain concept originated in the development of digital currencies as peer-to-peer versions of electronic cash [12, 83]. Bitcoin is the first known application of the blockchain technology and the Bitcoin Blockchain is basically a distributed database of records of Bitcoin transactions. Blockchain is a chain of digital signatures using public-key cryptographic protocols [79]. Built on the concept of highly distributed storage systems [113], blockchain technology can be considered a distributed data store with state machine replication using peer-to-peer protocol, where the transactions are the atomic changes to the stored data, which are grouped into blocks [73]. The integrity and tamper-resistance of the transactional data is guaranteed through linking of hash values among the blocks. Blockchain uses a distributed ledger as a decentralized data store, which is usually maintained by independent parties or nodes. The consistency of the transactional states of different distributed nodes is achieved through agreement by the consensus of the majority

nodes. Building on the fundamental concepts of the Byzantine fault tolerance algorithm [65], blockchain provides tolerance against arbitrary malicious behavior by adversaries. Blockchain also guarantees a consistent distributed transactional state that leads to identical changes among the independent nodes, even if some of the nodes are compromised or non-responsive due to failures. By using timestamping of its transactions and messages, blockchain provides universally verifiable proofs for existence or absence of a transaction in the distributed transactional state and the underlying cryptographic primitives using hash functions and digital signatures provide guarantee that these proofs are computationally secure and verifiable at any point of time [73]. These timestamped transactions based on cryptographic proofs in a distributed data store eliminate the need to have a central and trusted third party organization to authorize or agree to these transactions; thereby, trust is distributed equally among the participating independent nodes of the blockchain instead of a central authority. Moreover, in terms of accountability, blockchain uses cryptographic schemes that are expensive to compute, such as Proof-of-Work [12, 83]; and anchoring with cryptographic hashes [21], which provides guarantees against the system attacks and risk of system corruption, by making it extremely difficult to falsify the proofs and provide long-term non-repudiation [73].

Blockchain technology came to light when Bitcoin, a decentralized digital cash system, was introduced as a peer-to-peer cryptocurrency in 2009 [83] and as of 2018, Bitcoin is the largest cryptocurrency with a market capital of approximately more than 100 billion USD [34].

Moreover, an important feature of Bitcoin is maintainability of its currency value without any central authority or governmental administration. Bitcoin usage is based on transactions that are stored in the public distributed ledger (datastore) using blockchain technology. Bitcoin uses a Proof-of-Work (PoW) [12, 83] cryptographic scheme to enhance the security of the blockchain network against attacks [22]. As an example, an adversary may broadcast their own version of the blockchain by adding some false transactions to the network. As the security of the blockchain does not rely on any single authority, the independent nodes are not able to determine whether the broadcasted version of blockchain is valid or not. In order to avoid such types of attacks, Bitcoin uses a PoW hashing scheme in the form of mining blocks, where each node that participates in mining is required to solve a computationally challenging task to find a valid header conforming to certain constraints for the newly created block in the form of cryptographic hash. The consensus algorithm makes sure that the blockchain with the greatest cumulative blocks with valid block headers is deemed to be the validated chain. Therefore, even the adversary needs to solve the same computationally challenging tasks to find valid block headers for their newly created blocks, in competition to the rest of the blockchain. Therefore, they will only be successful if they can obtain significant computational resources (e.g., ≥51% blockchain network resources). Independent nodes

that participate in the mining of blocks are incentivized in the form of new coins of cryptocurrency for their work in producing new valid blocks.

### Bitcoin Cryptocurrency

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 [20]). 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 user's account balance and, 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 the same as the original sender address, but it is a best practice to create a new change address for each transaction to have anonymity in the Bitcoin world. Subsequently, to approve a transaction, the sender must use the corresponding *private key* to sign a transaction. The transaction is sent to the network, collected into *blocks* along with other transactions after being verified and, then, accepted into the Blockchain by the consensus of all peers. Finally, the transaction is broadcast 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, if the user chooses to generate a new pair of keys for each transaction. The problem with Bitcoin's architecture is that once a pseudonym is 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 publicizes its own address in order to receive Bitcoin from its customers) or through involuntary disclosure (such as data leakages, addresses taken from court documents, or data exchange partnerships between Bitcoin companies). Such clear-cut identification is, however, seldom possible. However, there are a variety of methods to effectively narrow down the scope of who could own a given Bitcoin address. As Reid et al. [92] found, it is possible to link the *change address* of a transaction back to the initial user. Furthermore, it is possible to cluster individual addresses that are controlled by the same person using different clustering techniques [105]. Moreover, it is even

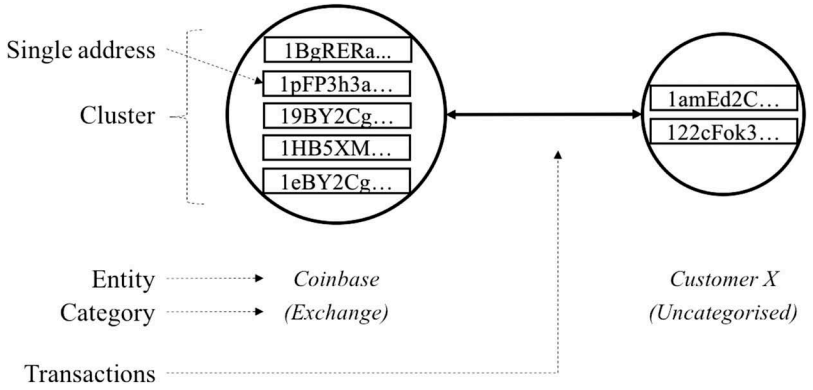


*Figure 1.* Anatomy of a bitcoin cluster.

possible to map IP addresses to Bitcoin addresses as described in sec. Anonymity Perspective. 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.

As shown in Figure 1, an *entity* is defined as a person or organization 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 assigned a *category*. The data provider currently assumes that every entity can only belong to one 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, which are grouped into a cluster, and pertain to an entity. The first entity (*Coinbase*) is labelled with the category label of *Exchange*. An Exchange allows their customers to trade bitcoins for fiat currencies, whereas the other entity is labelled *Uncategorized*, meaning the cluster has not yet been identified (i.e., it has not yet been linked to a real-world identity).

## Clustering Concepts

The transactional data supplied 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 categorized. As previously defined, 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 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 in that transaction, but all previous and future transactions involving those addresses are now linked.

- *Intelligence-based clustering*: In this type of clustering, information is gathered from outside the transaction history to better enable the clustering of data. Data sources for information gathering include but are not limited to: data leaks, court documents, data partnerships, exchanges that share their addresses, and manual merges due to service changing wallets.
- *Behavioral clustering*: As part of this clustering, patterns in the timing or structure of transactions will be utilized 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. Behavioral clustering can be used to cluster and relate the Bitcoin addresses to known hosted services or even to specific wallet software.

The data provider sends at least one transaction to every cluster before categorizing 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 minimize false positives, as errors could cause serious repercussions to innocent users. As discussed in the subsection Anonymity, cryptocurrencies allow their users to use pseudonyms to perform transactions. Since the cryptocurrency users self-select their pseudonyms (i.e., keys) and, in general, they create a new pseudonym for each transaction (as strongly advocated by the cryptocurrencies), it is difficult to identify the real identities behind the pseudonyms. However, the previously discussed clustering concepts can be applied to group the transactions and thereby the underlying pseudonyms as well, to profile these entities (who are the holders of the pseudonyms) based on their transactional behavior. In this context, our first proposition is:

*Proposition 1: De-anonymizing pseudonyms and their connected transactions in Bitcoin blockchain is possible using clustering techniques.*

## Supervised Machine Learning

For the analysis of Bitcoin transaction data, we used Supervised Machine Learning algorithms to detect patterns in the labelled dataset at hand. In this section, we will briefly describe the main idea behind the machine learning algorithm and give a concise introduction to various algorithms used in our method. More specifically,

the underlying function that explains the relationship between explanatory variables, often referred to as predictors, and responsive variables, often referred to as outcomes or targets in statistical literature, are computed to fit a prediction model [38]. Let n be the amount of training examples

$$(x1, .. ., xn, y)$$

where $x$ denotes the feature vector, $xi$ the individual feature component, and y the responsive variable. The Supervised Machine Learning algorithm seeks a function where $x$ is the input and y is the output. The learning problem at hand suggests a prediction model that can classify an entity type of 12 possible classes, which is referred to as a multi-class classification problem. This function can be denoted as:

$$y = f(x) + e$$

where e is the random error considered in the function. For a multi-class classification problem, the quantitative output for y is computed. From this, the algorithm can assign a label to the data point based on the class to which it most likely belongs [38]. The best machine learning algorithm depends on the given learning problem, as proposed by the No Free Lunch Theorem, suggested by Wolpert and Macready, who state that "any two optimization algorithms are equivalent when their performance is averaged across all possible problems" [111, p. 721]. Seven different Supervised Machine Learning algorithms have been tested in this paper. Additionally, the data used for training the prediction model is sparse; whereas, the computational burden associated with training the prediction model is assumed to be low. Therefore, several Machine Learning Algorithms used for hyperparameter tuning and modelling has been tested without meeting the computational budget. Furthermore, this approach presents interesting findings regarding the best machine learning algorithm for classifying a labelled blockchain dataset. Based on the latter, seven different learning algorithms have been applied: Decision Trees, Bagging, Random Forests, Extra Trees, AdaBoost, Gradient Boosting, and k-Nearest Neighbors.

### Decision Tree

One of the basic algorithms utilized is the Decision Tree, which has its name derived from a hierarchical model visually formed like a tree. The algorithm splits the observations into multiple branches, also referred to as subsets, based on a decision node with a given criteria. These criteria are determined from the explanatory variables, as the algorithm seeks to apply the most significant feature to perform the better split between the observations [13]. The best split can be measured by the information gain that is mathematically derived from a decrease in entropy (as explained in the following section). The Entropy describes the homogeneity of a sample distribution [91]. With classes equally divided in a binary subset, the entropy will be 1; however,

as the distribution becomes more homogeneous, the entropy will approach zero. Let G be the information gain and E the entropy:

$$G(y,x) = E(y) - E(y,x)$$

where $E(y)$ denotes the entropy in the parent node (i.e., the class distribution before the split) and $E(y, x)$ the entropy for the proportionally joined child nodes (i.e., after the sample has been split by the decision node, based on a given feature, denoted as x). The individual Entropy is derived from the following formula:

$$E = \sum -pj\ log2\ pj$$

where p is the probability of the $j^{th}$ class in the subset. The total Entropy for the split (i.e., the proportionally joined Entropy) is computed to estimate the information gain. With a feature set larger than one, the entropy for the child subset can be calculated for multiple features to search for the split with the largest information gain. Another way to measure impurity is to compute the Gini Index as evidenced by the following function. Let $G$ be the Gini Index:

$$G = 1 - \sum_{J} pj^{2}$$

Similar to the function for calculating Entropy, the probability distribution among the $j$ number of classes is denoted by $p$ and is derived from the class distribution in the subset. The number of classes corresponds to the 12 classes at hand in the learning problem of this paper. Based on this approach, the algorithm will seek the decision node that lowers the Gini Index. This is undertaken by calculating the weighted branch impurity for the different branches that emerges from a split [91]. This result is compared to other potential splits similar to the aforementioned information gain approach. In other words, the algorithm selects the feature split that leads to the best separation of classes, derived from the lowest Gini Index. Applicable to both the Entropy and Gini Index, the procedures are undertaken recursively until a certain limit is reached, depending on the selection of hyperparameters. There are different versions of the Decision Tree algorithm, such as ID3, C4.5, C.50, and CART. The scikit-learn library utilizes an optimized version of the CART algorithm that supports both criterions.[1]

### Bagging

Bagging, also referred to as Bootstrap Aggregation, is a technique for reducing the variance of an estimated prediction model [38]. A number of weak learners, such as a Decision Tree, is randomly trained on a subset of the training data. The base estimator (e.g., Decision Tree), number of learners, and maximum amount of samples, et cetera, are defined as hyperparameters in the algorithm.

### Random Forests

As an extension of decision trees, the Random Forest algorithm is approaching the classification task by constructing a multitude of Decision Trees to base its prediction upon. It utilizes a modification of the Bagging algorithm, as it de-correlates the trees through the Random Subspace Method. This method constructs multiple trees systematically by pseudo-randomly selecting subsets of the feature vector [53]. It will utilize the reduced variance achieved from the Bagging algorithm and select subsets of features to achieve better generalization.

### Extremely Randomized Trees

Extremely randomized trees are another tree-based ensemble method for classification problems. Geurts et al. [45] state that the cut-point, determined in the node of a decision tree, is associated with high variance in tree-based models, such as CART and C4.5. It is therefore responsible for a significant part of the error rates of tree-based methods [45]. Rather than relying on discriminative thresholds in the decision node, the split is randomly selected in this algorithm to mitigate these problems.

### Adaptive and Gradient Boosting

Boosting algorithms are another type of ensemble learners that fit sequences of weak learners, such as Decision Trees [88]. In other words, it tries to boost the weak learners similar to the Bagging algorithm, as it recursively selects a subset of the training data, which differs among the weak learners. However, AdaBoost (Adaptive Boosting) assigns weights to the samples, based on the ability of the weak learners, in order to predict the individual training sample. Thus, for each iteration, the sample weights are individually computed and the successive learner is applied to the new data subset [88]. Misclassified data points will score a higher weight and, therefore, have a higher chance of being selected for subsequent weak learners, which forces the learners to focus on data instances that are harder to predict. Similarly, the Gradient Boosting algorithm performs a gradient descent procedure that minimizes the loss function by recursively improving weak learners [39]. Thus, the residual loss is reduced by modifying the hyperparameters of the weak learner to minimize the misclassification rate. General to boosting algorithms, a weighted majority vote among the weak learners is utilized to produce a final prediction from the model.[2]

### Nearest Neighbours

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 and, thereby, sensitizing the model to minor changes [61].

Supervised Machine Learning approaches are quite good in identifying the patterns and also making predictions about yet unlabeled categories in the unlabeled dataset. Classifiers using Supervised Machine Learning first predict the probability of each of the categories of a qualitative variable to decide on the most probable label for the unlabeled category [58]. In this research work, we used Supervised Machine Learning algorithms to predict the category for a yet-unidentified entity. Using this approach, we also want to estimate the extent of the cyber-criminal and illicit activities in the Bitcoin blockchain ecosystem. In this context, we propose the exploration of the following propositions:

*Proposition 2: Supervised Machine Learning approaches can be utilized to build classifiers that can predict a category of a yet-unidentified clusters on the Bitcoin blockchain.*

*Proposition 3: Extent of the cybercriminal and illicit activities in the Bitcoin blockchain ecosystem can be estimated with good accuracy.*

## Methodology

In this section, we will first describe the dataset and its primary attributes, and then we will discuss various cluster categories in the context of Bitcoin cryptocurrency. Then, we will discuss the data analysis process, primarily in terms of the accuracies of various algorithms 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.

## Dataset Description

As mentioned before, the dataset used in this research was provided by the company *Chainalysis* [28], which specializes in blockchain data analysis. The dataset primarily contains transactional data, including details about every single transaction an entity has participated in, such as the time stamp, the value sent or received in Bitcoin 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 dataset used in this research contains approximately 395 million transactions pertaining to 957 unique clusters and Table 2 provides distribution of clusters among the predefined categories. The 957 observations are used as training and test sets for the current study, and they have been labeled by the data provider via proprietary domain specific heuristics. This dataset includes five to six categories that are commonly associated with illicit activities, which are darknet-market, mixing, ransomware, scam, stolen-bitcoins (a total of 104 observations), and gambling looking from some jurisdictions' perspective (206 observations in this case).

Table 3. Comparison of Accuracies of Original and Oversampled Datasets

| Supervised Algorithms | Original Dataset Accuracy | Oversampled Minority Accuracy | Oversampled Auto Accuracy |
|---|---|---|---|
| K-Nearest Neighbors (KNN) | 0.4292 | 0.5333 | 0.5806 |
| Classification and Regression Trees (CART) | 0.6917 | 0.7746 | 0.9096 |
| AdaBoost Classifier (ABC) | 0.6125 | 0.2698 | 0.1721 |
| Gradient Boosting Classifier (GBC) | 0.8042 | 0.8127 | 0.9575 |
| Random Forest Classifier (RFC) | 0.7833 | 0.7937 | 0.9564 |
| Extra Trees Classifier (ETC) | 0.7500 | 0.7968 | 0.9641 |
| Bagging Classifier (BGC) | 0.7833 | 0.8159 | 0.9575 |

The set of uncategorized clusters is independent of the previous 957 observations. The uncategorized set are addresses that were grouped by the data provider using the co-spend heuristic algorithm but without any label nor identification, meaning that there is not enough evidence of what the activity of those addresses is. This dataset includes transactional data on 153,293 clusters. In both cases, the transactions included are the ones included up to block height (i.e., number of blocks in the chain) of 512,033, which corresponds to the last block mined on 2018-03-04 23:58:21 (UTC). The number of transactions per cluster varies significantly, ranging from a low number (≥10) to several million transactions. Additionally, Appendix 2, Table 3 provides descriptive statistics of the transactions for every category. More specifically, the *average median, minimum*, and *maximum* number of transactions are shown, based on the number of observations within the respective category.

For each transaction, there are several attributes describing the transaction as shown in Appendix 2, Table 4. To describe the behavior of a cluster in a way that can be passed on to a Supervised Machine Learning algorithm, we extracted a set of features from the original input variables (see Appendix 2, Table 5) for each identified cluster. Apart from the extracted features, we engineered additional features such as the count 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 98 features. Appendix 2, Table 5 shows only a few important extracted features and the full list of 98 features has been omitted from the paper in view of space limitations.

## Cluster Categories

The Bitcoin addresses that were clustered together using the clustering techniques mentioned in sec. Clustering Concepts are further labeled with different category labels assigned by the data provider. These category labels can range from non-

Table 4. Classification Report from GBC with Original Dataset

| Cluster Categories | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| darknet-market | 0.6923 | 0.6923 | 0.6923 | 13 |
| exchange | 0.7391 | 0.9189 | 0.8193 | 74 |
| gambling | 0.8148 | 0.7586 | 0.7857 | 29 |
| hosted-wallet | 0.0000 | 0.0000 | 0.0000 | 2 |
| merchant-services | 0.0000 | 0.0000 | 0.0000 | 4 |
| mining-pool | 1.0000 | 0.7647 | 0.8667 | 17 |
| mixing | 1.0000 | 0.2500 | 0.4000 | 4 |
| other | 0.4286 | 0.4286 | 0.4286 | 14 |
| personal-wallet | 0.9459 | 0.9459 | 0.9459 | 74 |
| ransomware | 1.0000 | 0.5000 | 0.6667 | 6 |
| scam | 1.0000 | 0.6667 | 0.8000 | 3 |
| stolen-bitcoins | 0.0000 | 0.0000 | 0.0000 | 0 |
| **avg/total** | **0.8055** | **0.8083** | **0.7964** | **240** |

Table 5. Prediction Results

| Address | RFC | ETC | BGC | GBC |
|---|---|---|---|---|
| add1 | personal-wallet | personal-wallet | personal-wallet | exchange |
| add2 | ransomware | personal-wallet | gambling | ransomware |
| add3 | ransomware | personal-wallet | ransomware | exchange |
| add4 | personal-wallet | personal-wallet | exchange | ransomware |
| add5 | exchange | personal-wallet | gambling | ransomware |
| add6 | ransomware | ransomware | ransomware | ransomware |
| add7 | exchange | personal-wallet | ransomware | personal-wallet |
| add8 | gambling | gambling | other | gambling |
| add9 | exchange | other | exchange | exchange |
| add10 | gambling | personal-wallet | gambling | ransomware |
| add11 | exchange | other | exchange | exchange |
| add12 | personal-wallet | personal-wallet | personal-wallet | personal-wallet |
| add13 | ransomware | ransomware | ransomware | darknet-market |
| add14 | personal-wallet | exchange | ransomware | ransomware |
| add15 | ransomware | personal-wallet | ransomware | ransomware |
| add16 | ransomware | personal-wallet | ransomware | ransomware |
| add17 | exchange | personal-wallet | gambling | ransomware |
| add18 | exchange | personal-wallet | exchange | exchange |
| add19 | exchange | personal-wallet | gambling | ransomware |
| add20 | exchange | personal-wallet | gambling | ransomware |
| add21 | ransomware | personal-wallet | gambling | ransomware |
| add22 | exchange | personal-wallet | gambling | ransomware |

suspicious activities, such as *exchanges*, to high-risk categories, such as *ransomware*, and so on. The aforementioned list shows the categories used in this work for predicting a 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.
- *Darknet Market*: Marketplaces primarily facilitating trading of illegal goods such as narcotics, stolen credit cards, passports, and so forth. These sites are only accessible on the deep web through, for example, the TOR-browser.
- *Ransomware*: Entities that are utilizing the Bitcoin Blockchain as a medium of exchange to receive ransom fees.
- *Stolen Bitcoins*: Entities that managed to gain access to the private key(s) owned by other entities and committed thievery.
- *Personal Wallets*: Addresses or group of addresses managed by one entity for private uses such as trading, buying goods, gambling, and so forth.
- *Other*: Entities that have been identified but do not belong to any of the nine categories previously mentioned, for example, WikiLeaks' donation address.

Given the aforementioned clustering techniques, the consequently revealed cluster identities, and their corresponding categories, we can illustrate a network of identified clusters as shown in Figure 2.

## Data Analysis Process

As shown in Figure 3, the first phase in the data analysis process is the data preparation, which contains data pre-processing and features extraction as the main processing steps to transform the dataset into the required format. The required format is a matrix where the number of columns is equal to the number of features ($X$), the last column being the label ($y$), and the number of rows is the total number of observations across all labels. The resulting ready-to-feed-in dataset contains (957, 99) rows.

After the data preparation step, the analysis process is composed of three iterations using three different datasets: the original dataset, dataset with minority classes oversampled, and the final one, where all the classes have been oversampled to reach the same number of observations of the most populated class. The distribution of the three datasets is shown in Appendix-2; Figure 1.

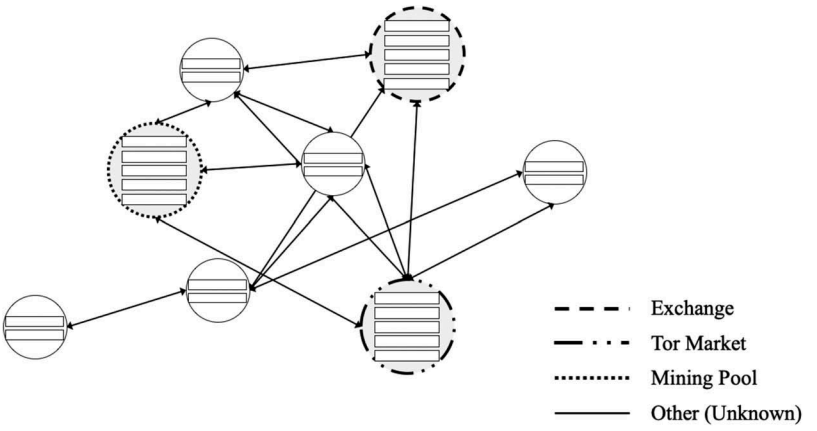


*Figure 2.* Visualization network of different categories

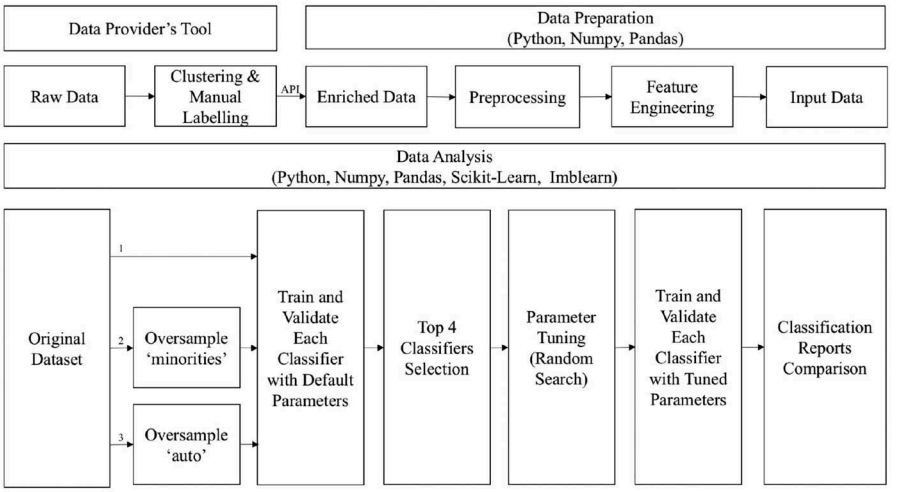


*Figure 3.* Data preparation and analysis process

As our dataset contains three minority and under-represented classes, *hosted-wallet, mixing*, and *stolen-bitcoins*, to compensate for the class imbalance, we used SMOTE (Synthetic Minority Over-Sampling Technique [30]). In each of the iterations, the data is split into training and validation subsets of 75% and 25%, respectively, and used as input for the seven aforementioned supervised learning algorithms with the default parameters as defined by scikit-learn [88]. The performance is measured by the mean cross-validation accuracy, which can be seen in Appendix 2, Figure 2, Figure 3, and Figure 4.

```
GradientBoostingClassifier(loss='deviance', learning_rate=0.1,
↪ n_estimators=100, subsample=1.0, criterion='friedman_mse',
↪ min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0,
↪ max_depth=3, min_impurity_decrease=0.0, min_impurity_split=None,
↪ init=None, random_state=None, max_features=None, verbose=0,
↪ max_leaf_nodes=None, warm_start=False, presort='auto')
```

*Figure 4.* Gradient boosting classifier's hyperparameters.

By comparing the performance across the seven algorithms, the top four were selected: Gradient Boosting, Random Forests, Extra Trees, and Bagging Classifier. For the next step, we tuned the hyperparameters for each model using cross-validated random search. Parameter tuning was undertaken using three-fold cross validation due to the scarcity of known clusters (n = 957). 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 [38]. A random search was utilized with 1,000 iterations, as it is empirically and theoretically more effective than grid search, as it allows the testing of a broader value spectrum for each parameter, and as it is less likely to waste effort on irrelevant hyperparameters, given the same amount of iterations [19]. We assessed the performance (shown in Appendix 2, Figure 5) of each algorithm after training each model with their respective set of optimal parameters.

After comparing the results from the three iterations, the models' oversampled datasets have been discarded. By looking at the cross-validation performance of the models in Appendix 2, Figure 1, Figure 4, where the mean cross-validation scores are ≥90%, indicating that the model is likely to be overfitting the training data and may not be performing well for unseen data. One could argue that the overfitting is caused by a disproportionate increase of synthetic samples as opposed to the original dataset (Appendix 2, Table 7). In the case of SMOTE (minorities), the class of *stolen-bitcoin* went from having 4 observations to 306, meaning that almost 24% of observations are synthetic; and, in the case of SMOTE (auto), almost 74% of observations are synthetic as shown in Appendix 2, Table 7.

After discarding the models that are trained with the oversampled datasets, the possible winning models are one of the top four trained with the original dataset with or without tuned hyperparameters. As shown in Appendix 2, Table 6, the algorithm with the best mean cross-validation accuracy score is Gradient Boosting (GBC) with the default parameters described in Figure 4. For comparison of the Classification Reports 4 and Appendix 2, Table 8, as well as the plotted receiving-operating-characteristic curves (ROC curves) on Figure 5 and Appendix 2, Figure 8, the Gradient Boosting algorithm with default parameters as implemented by Scikit-Learn was used with its tuned hyperparameters version using a random search.

As the classes of interest (darknet-market, scam, ransomware, stolen-bitcoins) are related to illicit activities, it is both important to maximize both precision and recall, since it is as important to minimize false positives as to maximize true positives.

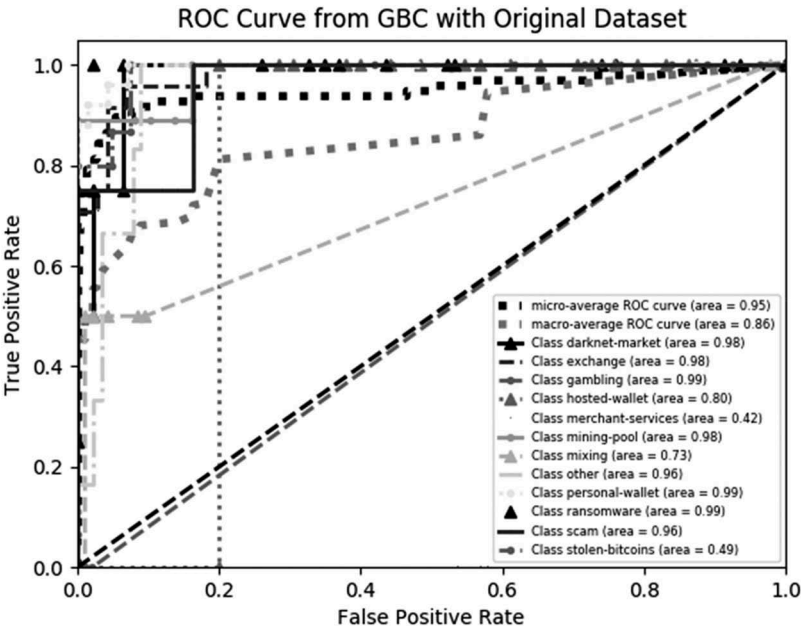


*Figure 5.* ROC curves with GBC and original dataset

Table 6. Summary Statistics of the Uncategorised Dataset

| | Avg. transactions | Median transaction | Min. transaction | Max. transaction | Avg.addresses | Min. address | Max. address |
|---|---|---|---|---|---|---|---|
| uncategorized | 3556.310712 | 1 | 1 | 15746608 | 929.0953729 | 1 | 2250892 |

Therefore, the final model selection will be determined by the *F*1-score. *F*1-scores are provided on the Classification Reports 4 and Appendix 2, Table 8, which indicate that the model with default hyperparameters has a 79.64% *F*1-score, which is 0.36% higher than the model with tuned hyperparameters 79.28%. Therefore, we conclude that tuning with hyperparameters did not help the enhancement of the *F*1-scores of the model and, hence, with choose GBC with an original dataset as our final model.

## Class Imbalance

There are two main 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.

Table 7. Results from Prediction by Classes

| Class | Count | % | Class | Count | % |
|---|---|---|---|---|---|
| exchange | 65102 | 42.4690 | ransomware | 222 | 0.1448 |
| mining-pool | 35767 | 23.3324 | mixing | 162 | 0.1057 |
| personal-wallet | 34887 | 22.7584 | hosted-wallet | 89 | 0.0581 |
| other | 8702 | 5.6767 | gambling | 47 | 0.0307 |
| scam | 7867 | 5.1320 | stolen-bitcoins | 7 | 0.0046 |
| darknet-market | 435 | 0.2838 | merchant-services | 6 | 0.0039 |

This creates dozens or even hundreds of change addresses, obfuscating the actual origin of a transaction, making it hard to identify and cluster addresses. Second, the data provider prioritizes some categories over others, depending on their customers' needs and cybercrime trends, which is why classes such as *hosted-wallet* have fewer observations. Clustering and identifying entities are an ongoing process; hence, the data provider increases the number of categorized entities as time passes.

## Dataset Limitations

Currently, the data used to train the prediction model does not include all of the data that is 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 categorized by the data provider. While we did have more than 900 categorized clusters, a larger sample size could potentially allow the discovery of more categories, as well as an increase in the number of examples for each of the already-defined 10 categories, thereby improving the performance of the model. Finally, with a larger sample, the methodology could be improved by utilizing a test sample, not seen by the classifier, to accurately justify the final model results.

## Results

In the previous section, we extensively described our methodology to identify the most suitable Supervised Machine Learning algorithm that is well suited to our Bitcoin transactions dataset. Based on the comparison of performance measures of various algorithms and various strategies for class imbalance and hyperparameter tuning, we have chosen Gradient Boosting wih default parameters for further analysis. The results presented in this section use the chosen Gradient Boosting algorithm.

## Testing the Model with a List of Suspects

As previously mentioned, the winning model is Gradient Boosting with default parameters, which has a F1-score of 79.64%. This model has been used to make a prediction on a list of 22 uncategorized clusters that are suspected to be associated with illicit activities. The prototype implementation in Python using scikit-learn can be found in Appendix A.4. In short, it is a Jupyter Notebook that loads a list of addresses that the user wishes to analyze. Then, a ready-to-use labeled dataset is fed to Gradient Boosting with default parameters, which focuses on 75% of the labeled set, and performs a prediction on the list of suspects. From the 22 uncategorized clusters, 14 have been labeled as either *ransomware* or *darknet-market*, which are related to illicit and cybercriminal activities. The prediction results of the 22 categories using the top four algorithms is shown in Table 5. As mentioned in research performed prior to this paper, Bitcoin has been widely associated with illicit activities, from payments in the darknet markets to the main payment system of choice for ransom payments [52, 114]. By using this prototype implementation, an analyst or investigator from a regulatory authority could potentially reduce a list of potential suspects and flag addresses for further investigation with an average accuracy of approximately 80% (due to an $F1$-score of 79.64%).

## Estimation of the Bitcoin Ecosystem

Finally, Gradient Boosting has been applied to predict the label of 153,293 uncategorized clusters (a summary of the data is shown in Table 6). These uncategorized clusters were grouped Bitcoin addresses according to co-spend heuristics, but not yet identified by the data provider. After performing the prediction of their labels, by visualizing the predicted results, as shown in Appendix 3, Figure 9, one could make an estimation of what the Bitcoin ecosystem looks like, as well as filtering through a large set of uncategorized clusters to find out which clusters are most likely to be related to cybercriminal entities. However, we do not claim that these predictions are accurate, but these results could provide a list of suspects who might be involved in cybercriminal and illicit activities (with approximately 80% certainty) and can be utilized for further investigation by regulatory authorities as shown in the prototype (results shown in Table 5).

## Discussion

One of the most fundamental design decisions of Bitcoin founder Satoshi Nakamoto was to utilize so-called *pseudonymity*, which presents an inherent tension between the two extremes of *anonymity* and *accountability*. Applications running on such pseudonymous transaction graphs, such as Bitcoin or most Ethereum-based applications, are becoming increasingly prevalent in business and society.[3] It is the characteristic of *pseudonymity* provided by the respective blockchain-based applications that enable their revolutionary attributes, such as decentralization and *trustless*,[4] while simultaneously remaining (publicly) auditable.

This inherent pseudonymity is beneficial as it allows for publicly auditable ledgers that still preserve the individuals' privacy. This enables full accountability in the system while not revealing a user's identity per se. This characteristic is especially powerful for systems where public auditability is beneficial to the public, for example, recording the votes of an election, where each vote is publicly visible. This way, the public can audit an election where individual voters' identities are not revealed. While the pseudonymous architecture of Bitcoin is fertile ground for many potentially revolutionary applications,[5] it also opens up the opportunity for illicit activities.

Pseudonymous blockchain technology such as Bitcoin has long troubled governments, as it is notoriously used for criminal activity, such as terror financing, thefts, money laundering, scams, and ransomware [54, 74, 75]. Not just individuals and organizations, but nation states can be harmed by the aforementioned criminal activities, as they facilitate organized crime, undermine governance, and decrease tax revenues. Hence, governments seek to eliminate such illicit money flows. This is the primary justification provided by several countries that have decided to ban Bitcoin rather than regulate it.[6]

Yet, criminal activity on Bitcoin is not solely a problem for state governance, but also leads consumers to doubt the technology's trustworthiness, effectively hampering technology adoption. As an example, the October 2013 take-down of the Silk Road illustrates how law enforcement action helps the adoption of Bitcoin. Immediately after the Silk Road was shut down, the dollar value of Bitcoin transactions spiked nearly 80% [29]. Therefore, we make the argument that a less anonymous Bitcoin Blockchain is actually *favorable* for its adoption, as it increases consumers' and governments' trust in the technology.

Our paper addresses this issue of accountability by developing and validating a novel method to categorize yet-unidentified clusters on the Bitcoin Blockchain using Supervised Machine Learning algorithms. Our analysis utilized a sample of 957 entities (with 385 million transactions), whose identity and type had been revealed, as the training set data and built classifiers differentiating among 12 categories and showed that we can indeed predict the type of a yet-unidentified entity. Using the Gradient Boosting algorithm with default parameters, we achieved a mean cross-validation accuracy of 80.42% and F1-score of 79.64%. This model is used to predict two uncategorized sets: a list of 22 addresses that are suspected to be related to cybercriminal activities and a list of 153,293 addresses to provide an estimation of the Bitcoin ecosystem. Admittedly, the research is limited by the sample size of observations. As shown previously, our model struggles when predicting classes with considerably low sample sizes such as hosted-wallet and merchant services.

Furthermore, predictive 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 broadening the set of possible algorithms. While the chosen algorithms are deemed computationally expensive, speed is not an issue, as we do not target real-time prediction.

## Implications for Organizations

The results show that it is indeed possible to categorize yet-unidentified clusters. This means, that one can reveal the category of a significant portion of entities on the Bitcoin Blockchain, challenging popular beliefs about Bitcoin's true level of anonymity. With regard to practical applications, our approach could potentially contribute to regulatory compliance and crime investigation irrespective of the two contrasting approaches of stringent versus open-ended regulation regarding work on legal aspects of regulating cryptocurrencies. First, one could flag suspicious entities, such as ransomware or scams, to prevent interaction with high-risk clusters. Also, in accordance to local financial regulations, a company that transacts on the Bitcoin Blockchain might be obliged to prove that the received money had not knowingly been involved in illicit activities.

For such compliance tasks, our research paves the way toward identifying and detecting high-risk transactions, enabling organizations to safeguard their reputation and to comply with local regulations.

## Societal Implications

Our findings spark a discussion on the societal implications of reducing Bitcoin's 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 weak spots of Bitcoin's anonymity, as found in this work, can have positive societal implications. Our research makes users aware of the technology's 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 wider societal trust in the cryptocurrency. Furthermore, heightened accountability can help law enforcement track down criminals and make it less attractive for criminals to turn to cryptocurrencies. When an actor's counter parties can be revealed using the described technique, law enforcement can pinpoint high-risk actors for further investigation and reduce the scope of possible suspects dramatically, which in turn would reduce criminal activity. Anonymity becomes weaker, the more is known about the linking to a subject [89].

As the European Union points out in their proposed cryptocurrency regulation[8]:

> 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.

As such, we believe that our work could inform policy makers rather than outright banning of cryptocurrencies; they could explore the spectrum ranging from

stringent to open-ended regulation. This would allow preservation of the desired features of blockchain-based technologies while addressing the undesired aspects of cybercriminal activities.

## Conclusion

In this paper, a multi-class classification on Bitcoin Blockchain clusters was 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. Using the Gradient Boosting algorithm with default parameters, we achieved an average cross-validation accuracy of 80.42% and a *F*1-score of 79.64% across all labels on 240 observations in the test set. This *F*1-score represents an average precision of 80.55% (the amount of true positives from the total predicted positives) and a recall of 80.83% (the amount of true positives among all actual positives). This model is used to make predictions on two uncategorized sets as demonstrations of its potential applications: a list of 22 addresses that are suspected to be related to cybercriminal activities, specifically related to addresses used to receive ransom payments, and a list of 153,293 addresses to provide an estimation of the Bitcoin ecosystem and the presence of different types of existing entities. Regardless of the accuracy derived from data analytics, our research outcome shows 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. Our paper makes three contributions: (a) it develops and validates a novel-method for de-anonymizing the Bitcoin blockchain transactions; (b) it provides the first estimation of different entity types in the Bitcoin Blockchain ecosystem; and (c) it provides implications for practitioner and regulators in addition to a prototype implementation of our method, which can be used as a tool to assess.

This work paves the way for further research, where 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-tuned differentiation between the clusters. Also, additional data could be utilized by harnessing more of the inherently available data on the Bitcoin Blockchain. Feature engineering processing could be improved, for example, by using automated feature extraction.

## Notes

1. https://en.wikipedia.org/wiki/Silk_Road_(marketplace) [1].
2. Tree algorithms: ID3, C4.5, C5.0 and CART.
3. Ensemble methods.
4. For example, German car manufacturer Daimler AG recently issued a e100 Million Corporate Bond on the blockchain [66]. Also, Bitcoin has reached a trading volume corresponding to more than USD one bn/day [57].
5. The ability to avoid the need for a trusted third party [23].

6. See [112] on the many different ways pseudonymous Blockchain technologies could revolutionize business and society.
7. See https://en.wikipedia.org/wiki/Legality_of_bitcoin_by_country_or_territory
8. Council and Parliament of the European Union: Amendment to directive (EU) 2015/849.

## Supplemental Material

Supplemental data for this article can be accessed on the publisher's website.

## ORCID

Raghava Rao Mukkamala http://orcid.org/0000-0001-9814-3883
Ravi Vatrapu http://orcid.org/0000-0002-9109-5281

## References

1. Abbasi, A.; Albrecht, C.; Vance, A.; and Hansen, J. Metafraud: A meta-learning framework for detecting financial fraud. *MIS Quarterly*, *36*, 4 (2012), 1293–1327.
2. Abbasi, A.; Chen, H.; and Nunamaker, J.F. Stylometric identification in electronic markets: Scalability and robustness. *Journal of Management Information Systems*, *25*, 1 (2008), 49–78.
3. Abbasi, A.; Zahedi, F.M.; Zeng, D.; Chen, Y.; Chen, H.; and Nunamaker Jr, J.F. Enhancing predictive analytics for anti-phishing by exploiting website genre information. *Journal of Management Information Systems*, *31*, 4 (2015), 109–157.
4. Abbasi, A.; Zhang, Z.; Zimbra, D.; Chen, H.; and Nunamaker Jr, J.F. Detecting fake websites: The contribution of statistical learning theory. *MIS Quarterly*, *34*, 3 (2010), 435–461.
5. Ajello, N.J. Fitting a square peg in a round hole: Bitcoin, money laundering, and the fifth amendment privilege against self-incrimination. *Brooklyn Law Review*, *80*, 2 (2014), 435–461.
6. Ali, S.T.; Clarke, D.; and McCorry, P. Bitcoin: Perils of an unregulated global p2p currency. In B. Christianson, P. Švenda, V. Matyáš, J. Malcolm, F. Stajano, and J. Anderson (eds.), *Security Protocols XXIII. Security Protocols 2015. Lecture Notes in Computer Science*, vol. 9379. Cham: Springer, 2015.
7. Androulaki, E.; Karame, G.O.; Roeschlin, M.; Scherer, T.; and Capkun, S. Evaluating user privacy in Bitcoin. In A.R. Sadeghi (ed.), *Financial Cryptography and Data Security. FC 2013. Lecture Notes in Computer Science*, vol. 7859. Berlin: Springer, 2013.
8. Aral, S.; Dellarocas, C.; and Godes, D. Introduction to the special issue-social media and business transformation: A framework for research. *Information Systems Research*, *24*, 1 (2013), 3–13.
9. Arruñada, B. Blockchain's struggle to deliver impersonal exchange. *Minnesota Journal of Law, Science & Technology*, *19*, 1 (2018), 55–105.
10. Atzori, M. (2016). Blockchain technology and decentralized governance: Is the state still necessary? *Social Science Research Network Working Paper Series*, 2015. http://dx.doi.org/10.2139/ssrn.2709713
11. Ba, S.; Stallaert, J.; Whinston, A.B.; and Zhang, H. Choice of transaction channels: The effects of product characteristics on market evolution. *Journal of Management Information Systems*, *21*, 4 (2005), 173–197.
12. Back, A. Hashcash-a denial of service counter-measure, 2002. http://www.hashcash.org/papers/hashcash.pdf. (accessed on January 24, 2019)
13. Baesens, B. *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*. Hoboken: John Wiley & Sons, 2014.

14. Barber, S.; Boyen, X.; Shi, E.; and Uzun, E. Bitter to better — How to make Bitcoin a better currency. In A.D. Keromytis (ed.), *Financial Cryptography and Data Security. FC 2012*. *Lecture Notes in Computer Science*, vol. 7397. Berlin: Springer, 2012.

15. Beck, R. Beyond bitcoin: The rise of blockchain world. *Computer*, *51*, 2 (2018), 54–58.

16. Beck, R.; Czepluch, J.S.; Lollike, N.; and Malone, S. Blockchain-the gateway to trust-free cryptographic transactions. In *Proceedings of ECIS*, ResearchPaper153. İstanbul, Turkey: Springer Publishing Company, 2016, pp. 1–14.

17. Beck, R.; Müller-Bloch, C.; and King, J.L. Governance in the blockchain economy: A framework and research agenda. *Journal of the Association for Information Systems*, *19*, 10 (2018), pp. 1020–1034.

18. Benjamin, V.; Zhang, B.; Nunamaker Jr, J.F.; and Chen, H. Examining hacker participation length in cybercriminal internet-relay-chat communities. *Journal of Management Information Systems*, *33*, 2 (2016), 482–510.

19. Bergstra, J.; and Bengio, Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*, Feb (2012), 281–305.

20. Bitcoin.org. Protect your privacy. 2018. https://bitcoin.org/en/protect-your-privacy. (accessed on January 24, 2019)

21. BitFury Group. On blockchain auditability. 2016. http://bitfury. com/content/5-white-papers-research/bitfury_white_paper_on_blockchain_ auditability.pdf. (accessed on January 24, 2019)

22. BitFury Group. Proof of stake versus proof of work. 2015. http://bitfury.com/content/5-white-papers-research/pos-vs-pow-1.0.2.pdf. (accessed on January 24, 2019)

23. Blundell-Wignall, A. The bitcoin question: Currency versus trust-less transfer technology. *OECD Working Papers on Finance, Insurance and Private Pensions*, 37, 1 (2014), pp. 1–21.

24. Böhme, R.; Christin, N.; Edelman, B.; and Moore, T. Bitcoin: Economics, technology, and governance. *The Journal of Economic Perspectives*, *29*, 2 (2015), 213–238.

25. Bonneau, J.; Narayanan, A.; Miller, A.; Clark, J.; Kroll, J.A.; and Felten, E.W. Mixcoin: anonymity for Bitcoin with accountable mixes. In N. Christin, and R. Safavi-Naini (eds.), *Financial Cryptography and Data Security. FC 2014*. *Lecture Notes in Computer Science*, vol. 8437. Berlin: Springer, 2014.

26. Brazier, F.; Oskamp, A.; Prins, C.; Schellekens, M.; and Wijngaards, N. Anonymity and software agents: An interdisciplinary challenge. *Artificial Intelligence and Law*, *12*, 1–2 (2004), 37–157.

27. Bryans, D. Bitcoin and money laundering: Mining for an effective solution. *Indiana Law Journal*, *89*, 1 (2014), 441–472.

28. Chainanalysis. *Protecting the Integrity of Digital Assets*, 2017. https://www.chainalysis. com/. (accessed on January 24, 2019)

29. Chainanalysis. The Changing Nature of Cryptocrime, 2018. https://www.chainalysis.com/static/Cryptocrime_Report_V2.pdf. (accessed on January 24, 2019)

30. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; and Kegelmeyer, W.P. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*(2002), 321–357.

31. Chen, P., Jiang, B., and Wang, C. Blockchain-based payment collection supervision system using pervasive Bitcoin digital wallet. In *Proceedings of the IEEE 13th International Conference on Wireless and Mobile Computing*, Networking and Communications (WiMob), Rome, (2017), 139–146. doi:10.1109/WiMOB.2017.8115844

32. Christin, N. 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), (2013), 213–224.

33. Christopher, C.M. The bridging model: Exploring the roles of trust and enforcement in banking, bitcoin, and the blockchain. *Nevada Law Journal*, *17*(2016), 139–180.

34. CoinMarketCap. Cryptocurrency market capitalizations, 2018. https://coinmarketcap.com/. (accessed on January 24, 2019)

35. Colombo, R.J. Bitcoin: Hype or harbinger. *Journal of International Business and Law, 16*, 1 (2016),1–5. http://scholarlycommons.law.hofstra.edu/jibl/vol16/iss1/3. (accessed on January 24, 2019)

36. DeLong, J.B.; and Froomkin, A.M. Speculative microeconomics for tomorrow's economy. In B. Kahin, and H. Varian (eds.), *Internet Publishing and Beyond: The Economics of Digital Information and Intellectual Property.* Cambridge: MIT Press, 2000, pp. 6–44.

37. Fleder, M.; Kester, M.S.; and Pillai, S. Bitcoin transaction graph analysis. *arXiv preprint arXiv:1502.01657*, 2015. https://arxiv.org/abs/1502.01657 (accessed on January 24, 2019)

38. Friedman, J.; Hastie, T.; and Tibshirani, R. The elements of statistical learning, vol. 1. *Springer Series in Statistics*. Berlin: Springer, 2001.

39. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*, 5 (2001), 1189–1232.

40. Froomkin, A.M. Anonymity and its enmities. *Journal of Online Law, art. 4*, (1995). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2715621. (accessed on January 24, 2019)

41. Froomkin, A.M. Flood control on the information ocean: Living with anonymity, digital cash, and distributed databases. *Journal of Law and Commerce, 15* (1996), 395–507.

42. Froomkin, A.M. Legal issues in anonymity and pseudonymity. *The Information Society*, *15*, 2 (1999), 113–127.

43. Froomkin, A. M. From anonymity to identification. *Journal of Self-Regulation and Regulation*, *1*, (2015), 120–138.

44. Gabison, G. Policy considerations for the blockchain technology public and private applications. *Science and Technology Law Review*, *19*, (2016), 327–350.

45. Geurts, P.; Ernst, D.; and Wehenkel, L. Extremely randomized trees. *Machine learning*, *63*, 1 (2006), 3–42.

46. Glaser, F. Pervasive decentralisation of digital infrastructures: A framework for blockchain enabled system and use case analysis. In *Proceedings of the 50th Hawaii International Conference on System Sciences*. Honolulu: Hawaii International Conference on System Sciences (HICSS), 2017, pp. 1543–1552.

47. Glaser, F.; and Bezzenberger, L. Beyond cryptocurrencies - a taxonomy of decentralized consensus systems. In *Proceedings of the 23rd European Conference on Information Systems (ECIS 2015)*. Atlanta: Association for Information Systems, 2015, pp. 1–18.

48. Glaser, F.; Zimmermann, K.; Haferkorn, M.; Weber, M.; and Siering, M. Bitcoin- asset or currency? Revealing users' hidden intentions. In *Twenty Second European Conference on Information Systems*. Atlanta: Association for Information Systems, 2014, pp. 1–15.

49. Godsiff, P. Bitcoin: Bubble or blockchain. *Agent and Multi-Agent Systems: Technologies and Applications*, 38 (2015), 191–203.

50. Goldfeder, S.; Kalodner, H.; Reisman, D.; and Narayanan, A. When the cookie meets the blockchain: Privacy risks of web payments via cryptocurrencies. *arXiv preprint arXiv:1708.04748*, (2017). https://arxiv.org/pdf/1708.04748.pdf. (accessed on January 24, 2019)

51. Guo, A. Blockchain receipts: Patentability and admissibility in court. *Chicago-Kent Journal of Intellectual Property*, *16*, 2 (2017), 440–452.

52. Harlev, M.A.; Sun Yin, H.; Langenheldt, K.C.; Mukkamala, R.; and Vatrapu, R. Breaking bad: De-anonymising entity types on the bitcoin blockchain using supervised ma- chine learning. In *Proceedings of the 51st Hawaii International Conference on System Sciences*. Honolulu: Hawaii International Conference on System Sciences (HICSS), 2018, pp. 3497–3506.

53. Ho, T.K. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*, 8 (1998), 832–844.

54. Hout, M.C.V.; and Bingham, T. Silk Road, the virtual drug marketplace: A single case study of user experiences. *International Journal of Drug Policy*, *24*, 5 (2013), 385–391.

55. Hui, K.-L.; Kim, S.H.; and Wang, Q.-H. Cybercrime deterrence and international legislation: Evidence from distributed denial of service attacks. *MIS Quarterly*, *41*, 2 (2017), 497–523.

56. Hyvärinen, H.; Risius, M.; and Friis, G. A blockchain-based approach towards overcoming financial fraud in public sector services. *Business & Information Systems Engineering*, *59*, 6 (2017), 441–456.

57. Investinblockchain. Bitcoin estimated transaction volume usd. 32, (2018). https://www.investinblockchain.com/category/bitcoin/. (accessed on January 24, 2019)

58. James, G.; Witten, D.; Hastie, T.; and Tibshirani, R. *An Introduction to Statistical Learning*. New York: Springer, 2013.

59. Karlstrøm, H. Do libertarians dream of electric coins? the material embeddedness of bitcoin. *Distinktion: Scandinavian Journal of Social Theory*, 15,1(2014), 23–36.

60. Karwatzki, S.; Dytynko, O.; Trenz, M.; and Veit, D. Beyond the personalization–privacy paradox: Privacy valuation, transparency features, and service personalization. *Journal of Management Information Systems*, *34*, 2 (2017), 369–400.

61. Kennedy, P. *A Guide to Econometrics*. Cambridge: MIT Press, 2003.

62. Kiviat, T.I. Beyond bitcoin: Issues in regulating blockchain transactions. *Duke Law Journal*, *65*, (2015), 569–608.

63. Kleiman, J.A. Beyond the silk road: Unregulated decentralized virtual currencies continue to endanger us national security and welfare. *American University National Security Law Brief*, *4*, 1 (2013), 59–78.

64. Koshy, P.; Koshy, D.; and McDaniel, P. An analysis of anonymity in bitcoin using p2p network traffic. *International Conference on Financial Cryptography and Data Security*, 8437 (2014), pp. 469–485.

65. Lamport, L.; Shostak, R.; and Pease, M. The Byzantine Generals problem. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, *4*, 3 (1982), 382–401.

66. LBBW. *Daimler and LBBW: Successful with Blockchain*, 2018. https://www.lbbw.de/articlepage/experience-banking/pilot-project-blockchain-daimler-lbbw_661e61yw9_e.html. (accessed on January 24, 2019)

67. Lee, L. New kids on the blockchain: How bitcoin's technology could reinvent the stock market. *Hastings Business Law Journal*, *12*, 2 (2016), 81–132.

68. Levmore, S. The internet's anonymity problem. In S. Lemore and M. Nussbaum (eds.), *The offensive Internet: Speech, Privacy, and Reputation*. Cambridge: Harvard University Press, 2010.

69. Li, W.; Chen, H.; and Nunamaker Jr, J.F. Identifying and profiling key sellers in cyber carding community: A secure text mining system. *Journal of Management Information Systems*, *33*, 4 (2016), 1059–1086.

70. Li, X.; Sun, S.X.; Chen, K.; Fung, T.; and Wang, H. Design theory for market surveillance systems. *Journal of Management Information Systems*, *32*, 2 (2015), 278–313.

71. Lin, T.C. Compliance, technology, and modern finance. *Brooklyn Journal of Corporate, Financial & Commercial Law*, *11*, 1 (2016), 159–182.

72. Mai, F.; Shan, Z.; Bai, Q.; Wang, X.; and Chiang, R.H. How does social media impact bitcoin value? A test of the silent majority hypothesis. *Journal of Management Information Systems*, *35*, 1 (2018), 19–52.

73. Mamoshina, P.; Ojomoko, L.; Yanovich, Y.; Ostrovski, A.; Botezatu, A.; Prikhodko, P.; Izumchenko, E.; Aliper, A.; Romantsov, K.; Zhebrak, A.; Ogu, I.A.; and Zhavoronkov, A. Converging blockchain and next-generation artificial intelligence technologies to decentralize and accelerate biomedical research and healthcare. *Oncotarget*, *9*, 5 (2018), 5665–5690.

74. Martin, J. *Drugs on the Dark net: How Cryptomarkets are Transforming the Global Trade in Illicit Drugs*. New York: Palgrave Macmillan, 2014.

75. Martin, J. Lost on the Silk Road: Online drug distribution and the cryptomarket. *Criminology & Criminal Justice*, *14*, 3 (2014), 351–367.

76. McLeod, P. Taxing and regulating bitcoin: The government's game of catch up. *CommLaw Conspectus: Journal of Communications Law and Technology Policy*, *22*, (2013), 379–406.

77. Meiklejohn, S.; and Orlandi, C. Privacy-enhancing overlays in bitcoin. In *International Conference on Financial Cryptography and Data Security*. Berlin: Springer, 2015, pp. 27–141.

78. Meiklejohn, S.; Pomarole, M.; Jordan, G.; Levchenko, K.; McCoy, D.; Voelker, G.M.; and Savage, S. A fistful of bitcoins: Characterizing payments among men with no names. In *Proceedings of the 2013 Conference on Internet Measurement Conference*. New York: ACM, 2013, pp. 127–140.

79. Merkle, R. C. Protocols for public key cryptosystems. In *The 1980 IEEE Symposium on Security and Privacy*. New York: IEEE, 1980, pp. 122–134.

80. Morgan, J.S. What I learned trading cryptocurrencies while studying the law. *University of Miami International and Comparative Law Review*, *25*, (2018), 159–226.

81. Moser, M.; Bohme, R.; and Breuker, D. An inquiry into money laundering tools in the bitcoin ecosystem. *eCrime Researchers Summit (eCRS)*, 2013, 1–14. https://ieeexplore.ieee.org/document/6805780. (accessed on January 24, 2019)

82. Naerland, K.; Müller-Bloch, C.; Beck, R.; and Palmund, S. Blockchain to rule the waves nascent design principles for reducing risk and uncertainty in decentralized environments. In *Proceedings International Conference on Information Systems (ICIS)*. Atlanta: Association for Information Systems, 2017, pp. 1–16.

83. Nakamoto, S. Bitcoin: A peer-to-peer electronic cash system, 2008. https://bitcoin.org/en/bitcoin-paper. (accessed on January 24, 2019)

84. Narayanan, A.; Bonneau, J.; Felten, E.; Miller, A.; and Goldfeder, S. *Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction*. Princeton: Princeton University Press, 2016.

85. Nofer, M.; Gomber, P.; Hinz, O.; and Schiereck, D. Blockchain. *Business & Information Systems Engineering*, *59*, 3 (2017), 183–187.

86. Notheisen, B.; Hawlitschek, F.; and Weinhardt, C. Breaking down the blockchain hype towards a blockchain market engineering approach. In *Proceedings of the 25th European Conference on Information Systems (ECIS)*. Atlanta: Association for Information Systems, 2017, pp. 1062–1080.

87. Otjacques, B.; Hitzelberger, P.; and Feltz, F. Interoperability of e-government information systems: Issues of identification and data sharing. *Journal of Management Information Systems*, *23*, 4 (2007), 29–51.

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

89. Pfitzmann, A.; and Köhntopp, M. Anonymity, unobservability, and pseudonymity-a proposal for terminology. *Designing Privacy Enhancing Technologies*, 2009 (2001), 1–9.

90. Pinsonneault, A.; and Heppel, N. Anonymity in group support systems research: A new conceptualization, measure, and contingency framework. *Journal of Management Information Systems*, *14*, 3 (1997), 89–108.

91. Raileanu, L. E.; and Stoffel, K. Theoretical comparison between the Gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, *41*, 1 (2004), 77–93.

92. Reid, F.; and Harrigan, M. An analysis of anonymity in the bitcoin system. *Security and Privacy in Social Networks*, (2013), 197–223. https://link.springer.com/chapter/10.1007/978-1-4614-4139-7_10. (accessed on January 24, 2019)

93. Reyes, C.L. Conceptualizing cryptolaw. *Nebraska Law Review*, *96*, 2 (2017), 384–445.

94. Risius, M.; and Spohrer, K. A blockchain research framework. *Business & Information Systems Engineering*, *59*, 6 (2017), 385–409.

95. Ron, D.; and Shamir, A. Quantitative analysis of the full bitcoin transaction graph. *International Conference on Financial Cryptography and Data Security*, 7859 (2013), 6–24.

96. Rosner, M.T.; and Kang, A. Understanding and regulating twenty-first century payment systems: The ripple case study. *Michigan Law Review*, *114*, 4 (2016), 649–681.

97. Ross, E.S. Nobody puts blockchain in a corner: The disruptive role of blockchain technology in the financial services industry and current regulatory issues. *Catholic University Journal of Law and Technology*, *25*, 2 (2017), 353–386.

98. Samtani, S.; Chinn, R.; Chen, H.; and Nunamaker Jr, J.F. Exploring emerging hacker assets and key hackers for proactive cyber threat intelligence. *Journal of Management Information Systems*, *34*, 4 (2017), 1023–1053.
99. Sasson, E.B.; Chiesa, A.; Garman, C.; Green, M.; Miers, I.; Tromer, E.; and Virza, M. Zerocash: Decentralized anonymous payments from bitcoin. In *2014 IEEE Symposium on Security and Privacy*. New York: IEEE, 2014, pp. 459–474.
100. Scott, S.V.; and Orlikowski, W.J. Entanglements in practice: performing anonymity through social media. *MIS Quarterly*, *38*, 3 (2014), 873–893.
101. Shackelford, S.J.; and Myers, S. Block-by-block: Leveraging the power of blockchain technology to build trust and promote cyber peace. *Yale Journal of Law and Technology*, *19*, 1 (2017), 334–338.
102. Sklaroff, J.M. Smart contracts and the cost of inflexibility. *University of Pennsylvania Law Review*, *166*(2018), 263–303.
103. Smith, A.; and Weismann, M.F. Are you ready for digital currency? *Journal of Corporate Accounting & Finance*, *26*, 1 (2014), 17–21.
104. Sonderegger, D. A regulatory and economic perplexity: Bitcoin needs just a bit of regulation. *Washington University Journal of Law & Policy*, *47*, (2015), 175–216.
105. Spagnuolo, M.; Maggi, F.; and Zanero, S. Bitiodine: Extracting intelligence from the bitcoin network. *International Conference on Financial Cryptography and Data Security*, 8437 (2014), 457–468.
106. Surujnath, R. Off the chain: A guide to blockchain derivatives markets and the implications on systemic risk. *Fordham Journal of Corporate & Financial Law*, *22*, 2 (2017), 257–304.
107. Tsukerman, M. The block is hot: A survey of the state of bitcoin regulation and suggestions for the future. *Berkeley Technology Law Journal*, *30*, 4 (2015), 1127–1169.
108. Turpin, J.B. Bitcoin: The economic case for a global, virtual currency operating in an unexplored legal framework. *Indiana Journal of Global Legal Studies*, *21*, 1 (2014), 335–368.
109. Twyman, N.W.; Lowry, P.B.; Burgoon, J.K.; and Nunamaker Jr, J.F. Autonomous scientifically controlled screening systems for detecting information purposely concealed by individuals. *Journal of Management Information Systems*, *31*, 3 (2014), 106–137.
110. Van Hout, M.C.; and Bingham, T. Responsible vendors, intelligent consumers :Silk road, the online revolution in drug trading.*International Journal of Drug Policy*, *25*, 2 (2014), 183–189.
111. Wolpert, D.H.; and Macready, W.G. Coevolutionary free lunches. *IEEE Transactions on Evolutionary Computation*, *9*, 6 (2005), 721–735.
112. Wright, A.; and De Filippi, P. Decentralized blockchain technology and the rise of lex cryptographia. *SSRN*, (2015). https://ssrn.com/abstract=2580664. (accessed on January 24, 2019)
113. Xu, L. *Highly available distributed storage systems*. PhD thesis, California Institute of Technology, 1999.
114. Yin, H.S.; and Vatrapu, R. A first estimation of the proportion of cybercriminal entities in the bitcoin ecosystem using Supervised Machine Learning. In *2017 IEEE International Conference on Big Data*. New York: IEEE, 2017, pp. 3690–3699.
115. Young, S. Enforcing constitutional rights through computer code. *Catholic University Journal of Law and Technology*, *26*, 1 (2018), 1–20.