Electronic Markets (2023) 33:37
https://doi.org/10.1007/s12525-023-00654-3

RESEARCH PAPER

Detecting anomalous cryptocurrency transactions: An AML/CFT application of machine learning-based forensics

Nadia Pocher[1] · Mirko Zichichi[2] · Fabio Merizzi[2] · Muhammad Zohaib Shafiq[2] · Stefano Ferretti[3]

Received: 29 September 2022 / Accepted: 30 May 2023

Abstract

In shaping the Internet of Money, the application of blockchain and distributed ledger technologies (DLTs) to the financial sector triggered regulatory concerns. Notably, while the user anonymity enabled in this field may safeguard privacy and data protection, the lack of identifiability hinders accountability and challenges the fight against money laundering and the financing of terrorism and proliferation (AML/CFT). As law enforcement agencies and the private sector apply forensics to track crypto transfers across ecosystems that are socio-technical in nature, this paper focuses on the growing relevance of these techniques in a domain where their deployment impacts the traits and evolution of the sphere. In particular, this work offers contextualized insights into the application of methods of machine learning and transaction graph analysis. Namely, it analyzes a real-world dataset of Bitcoin transactions represented as a directed graph network through various techniques. The modeling of blockchain transactions as a complex network suggests that the use of graph-based data analysis methods can help classify transactions and identify illicit ones. Indeed, this work shows that the neural network types known as Graph Convolutional Networks (GCN) and Graph Attention Networks (GAT) are a promising AML/CFT solution. Notably, in this scenario GCN outperform other classic approaches and GAT are applied for the first time to detect anomalies in Bitcoin. Ultimately, the paper upholds the value of public–private synergies to devise forensic strategies conscious of the spirit of explainability and data openness.

Keywords Blockchain technology · Financial technology · Network forensics · Graph analysis · AML/CFT

JEL Classification G18 · O33

Introduction

Over the last 15 years, the application of blockchain and distributed ledger technologies (DLTs) to the financial domain has generated an enthusiastic hype (Ali et al., 2020). Building on years of research in distributed systems and cryptography, the launch of Bitcoin (Nakamoto, 2008) showed it is possible to reliably record information (e.g., transactions) without trusting a centralized party. This opened the way to peer-to-peer transfers and direct participation in a digital global economy. However, the features of disintermediation and perceived anonymity of this Internet of Money (Antonopoulos, 2017) cause regulatory unease.[1] Indeed, they defy accountability and fuel

Responsible Editor: Gilbert Fridgen

✉ Stefano Ferretti
stefano.ferretti@uniurb.it

Nadia Pocher
nadia.pocher@uab.cat

Mirko Zichichi
mirko.zichichi2@unibo.it

Fabio Merizzi
fabio.merizzi@studio.unibo.it

Muhammad Zohaib Shafiq
muhammad.shafiq6@studio.unibo.it

[1] Universitat Autònoma de Barcelona, Bellaterra, Spain

[2] University of Bologna, Bologna, Italy

[3] University of Urbino, Piazza Della Repubblica, 13, 61029 Urbino, Italy

[1] The Internet of Money is neither a legal nor a technical definition; in this work, the term refers to the entire set of cryptocurrency ecosystems, thus including the part of the Internet of Value (Tapscott and Euchner (2019)) that relates to payment tokens.

Published online: 26 July 2023

fears of exploitation for illicit purposes (Chang et al., 2020). As confirmed by industry estimates, in 2022, the volume of crypto-related illicit activity hit USD 20.6 billion and increasingly involves decentralized finance (DeFi) (Chainalysis Team, 2023). This challenges the fight against money laundering and the financing of terrorism and proliferation (AML/CFT).

The AML/CFT framework consists of a set of laws, regulations and procedures that aim to protect the integrity of the financial system mainly by making the concealment of the origin of illicit profits significantly troublesome (Pocher & Zichichi, 2022). Since the identification of customers and counterparties is a key part of AML/CFT compliance for entities such as financial institutions and cryptoasset service providers, some features of the Internet of Money that hinder identifiability emerge as problematic. However, crypto-related laundering appears heavily concentrated: most value originating from illicit addresses is seemingly sent to few services, often built for criminal purposes (Chainalysis Team, 2023). This suggests the key role of effective, and possibly efficient, classification of transactions performed by/received from specific entities to detect and investigate illicit activities in the sphere at hand.

In this context, the picture of untraceable cryptocurrency transfers and individual freedom from governmental control warrants a two-fold interpretation: while user anonymity can safeguard privacy and data protection, lack of identifiability hampers investigation, enforcement, and accountability. Two sets of mutually influencing socio-technical events emerged: law enforcement agencies and private sector providers of RegTech solutions started exploring techniques to "follow the money" across blockchain ecosystems (Bartoletti et al., 2020; Biryukov & Tikhomirov, 2019; Chen et al., 2019; Lischke & Fabian, 2016; Meiklejohn et al., 2016; Moreno-Sanchez et al., 2016), while the unveiled insufficiency in Bitcoin's anonymity spurred altcoin projects (e.g., Monero, ZCash) to implement advanced cryptographic methods that require new analytical tools.[2]

Against this backdrop, in this paper, we focus on the value of intelligence techniques to provide insights into the Internet of Money's ecosystems, with specific regard to machine learning techniques, network, and transaction graph analysis (Fleder et al., 2015; Ober et al., 2013; Weber et al., 2019; Wu et al., 2021). We first provide a background on a notion of anonymity that is specific to the Internet of Money and on the interplay of AML/CFT and blockchain forensics. Consequently, we focus on the anomaly detection approaches that led to our experiments. In particular, we employed a dataset obtained from a set of Bitcoin transactions, represented as a directed graph network (Weber et al., 2019). The modeling of Bitcoin transactions as a complex network fosters the use of specific graph-related analysis techniques, which usually help identify peculiar nodes of a network (Pocher & Zichichi, 2022). As per our central hypothesis, since money laundering involves transaction flow relationships between entities creating a graph structure, AML/CFT analytics could benefit from novel graph analysis techniques in machine learning, namely Graph Convolutional Networks (GCN) and Graph Attention Networks (GAT).

The results of our experiments show that GCN and GAT neural network typologies are promising solutions for AML/CFT. This is in opposition to a state-of-the-art work in which a baseline supervised learning algorithm, i.e., non-graph-based, such as the Random Forest, provided the best performances (Weber et al., 2019). Thus, we underline the value of experimenting with techniques based on machine learning and transaction graph analysis and their combinations. We contextualize our argument vis-`a-vis the amount and complexity of crypto transaction data and the specifics of AML/CFT anomaly indicators. We do so by considering the need to mitigate the shortcomings of rule-based regimes, explainability aspects, and the urgency to engage in research informed by an interdisciplinary methodology.

To summarize, the main contribution of this work is twofold:

- We show how modeling blockchain transactions as complex networks is conducive to the subsequent application of specific graph-based learning approaches for anomaly detection purposes. In particular, our experiments show how the GCN model generates better results than other machine learning methods. Notably, it seems to outperform state-of-the-art implementations in classifying illicit transactions;
- Our work heeds a compound of technical, operational, and regulatory viewpoints when considering the benefits of machine learning for AML/CFT anomaly detection. This allows us to account for the need for interpretability and explainability, as well as the effectiveness and efficiency of the deployed approaches.[3] This methodology displays the value of cross-disciplinary models to improve accuracy, significantly aiding compliance and investigation, reducing false positives and over-reporting.

The remainder is structured as follows. The "Background" section provides a conceptual background on

[2] The term "RegTech", short for "regulatory technology", refers to the use of new technologies to aid regulatory and compliance processes, mainly through FinTech software applications.

[3] Research into the regulatory impacts of the explainability and interpretability of AI applications is vast and detailed; in light of the scope of this work, we perform inevitable simplifications.

Bitcoin's pseudonymity and insights into the relationship between AML/CFT and forensics. The "Related work" section explores related work. The "Anomaly detection approaches" section takes a context-specific approach to outline anomaly detection techniques. In the "Experimenting with machine learning" sections and the "Discussion" sections, we present and discuss our study on machine learning-based AML/CFT classification methods. The "Conclusions" section concludes the paper.

Background

In this section, we discuss key aspects related to pseudonymity and deanoymization, followed by a discussion on AML/CFT and blockchain forensics.

Preliminarily, we point out that in this work, the terms DLT and blockchain are used as synonyms. As is known, while the term DLT refers to a generic idea of distributed ledger, regardless of its implementation, a blockchain is a specific form of DLT in which transactions are stored as a sequence of blocks. Even if the term blockchain is more specific, it is more popular and often used in a wider sense. In this work, the way the DLT stores transactions in the ledger does not influence our model. Indeed, our approach considers the graph generated by transactions, i.e., a direct link from a transaction, say t_1, to another t_2, exists if the money earned in t_1 is spent in t_2. Thus, our study focuses on a layer that is higher than the ledger where transactions are stored.

Pseudonymity and de-anonymization

Untraceability of payments was among the goals of the Internet of Money (Filippi & Wright, 2018). However, in concrete terms, the latter is populated by many socio-technical notions of anonymity and transparency (Pocher & Zichichi, 2022). Following a holistic interpretation, cryptocurrencies are generated and exchanged within socio-technical systems that, as such, comprise of interdependent technology and human systems (Baxter & Sommerville, 2011; Desmond et al., 2019). Hence, their characteristics are influenced by social and technical aspects. Since a literature review (Amarasinghe et al., 2019) falls outside the scope of our work, we heed the specific understanding that anonymity in the Internet of Money "means being able to conduct a financial transaction without anyone besides the sender and the receiver being able to identify the parties involved" (Edmunds, 2020). Indeed, it is a common blockchain goal to combine user anonymity and transparency of operations (i.e., ledger transparency), and a public blockchain is structurally designed to enable anonymous peer-to-peer transfers (Quiniou, 2019).

There is wide agreement that Bitcoin is pseudonymous, and not anonymous (Berg, 2019; Biryukov & Tikhomirov, 2019; Li et al., 2019). Pseudonymity refers to the use of pseudonyms as identifiers, and a pseudonym is a subject's identifier other than the subject's real name (Pfitzmann & Hansen, 2010). In most blockchain systems, public–private key pairs uniquely identify wallet holders (Wang & De Filippi, 2020). Hence, in a crypto transaction, addresses (i.e., public keys) perform the function of usernames. It follows that senders and recipients are pseudonymous, not anonymous, when their address identifies them. However, this is not sufficient from a regulatory perspective because pseudonyms alone do not ensure accountability. Indeed, when AML/CFT rules require identification, they refer to real-world identities.

In principle, a currency scheme aims to prevent that the transaction history of its units can be retraced. If it is possible to associate a coin with its past exchanges, the currency's fungibility is threatened, and its nominal value is affected. Because Bitcoin's features seemed insufficient, new techniques have been embedded into anonymity-enhanced currencies (AECs), also known as "privacy coins", to bypass regulatory constraints and surveillance. They deploy privacy-enhancing technologies such as zero-knowledge proofs (ZKPs). Concurrently, even if the most common way to buy and exchange cryptocurrencies still relies on centralized exchanges, the Internet of Money is witnessing the emergence of DeFi applications,[4] such as stablecoin projects (e.g., DAI), lending platforms (e.g., Aave, Compound), decentralized exchanges (e.g., Uniswap, Pancakeswap) (Amler et al., 2023; Aramonte et al., 2021; Katona, 2021). The total value of DeFi projects reportedly amounted to USD 1 billion in January 2020, USD 27 billion in January 2021, USD 60 billion in April 2021, and USD 40 billion in November 2022 (Chainalysis Team, 2022).

Meanwhile, the private sector and law enforcement professionals have devised strategies to trace transfers in the Internet of Money. The end goal of these intelligence methods is to match users, definitively or statistically, to transactions performed by crypto-addresses—i.e., to connect pseudonyms to real-world identities—leveraging unique identifiers. These techniques were originally labeled "blockchain forensics", as they were informed by the specificities of blockchains and defined as the use of science and technology for the sake of investigation and fact-establishment in a court of law, primarily dealing with recovering and analyzing the evidence on blockchain ledgers (Phan, 2021). Later, analytic solutions started to be requested by regulated entities. Although they have been mostly tested on the Bitcoin

[4] DeFi was defined as an "ecosystem of financial services realized through smart contracts deployed on public distributed ledgers" (Amler et al., 2023), where the role of intermediaries is replaced by self-executing computer code (Katona, 2021). Nonetheless, the levels of decentralization of the relevant projects vary and are debated (Barbereau et al., 2023).

network, data-exploitation strategies have been deployed on Ethereum (Bartoletti et al., 2020; Chen et al., 2019; Li et al., 2021; Moreno-Sanchez et al., 2016), and on non-blockchain DLTs (Ince et al., 2018; Tennant, 2017). Evidently, both obfuscation and traceability are not only endeavors pursued by actors belonging to the crypto sphere, but also activities that influence the overall anonymous or transparent socio-technical character of the domain.

Since identifiers (i.e., addresses/public keys) can be leveraged to connect transactions to their history, Bitcoin's pseudonymity generates an inherent tension between anonymity and accountability (Yin et al., 2019). However, unless they are associated with additional data, identifiers do not reveal personally identifying information (Wang & De Filippi, 2020). Hence, pseudonymity does not imply identifiability, which is subjective: a pseudonymous subject is identifiable only if a specific actor can discover its real-world identity. This is crucial because, in the Internet of Money, there are both (a) actors, such as authorities and cryptoasset service providers, that seek to achieve identification, and (b) strategies employed at various levels to avert it, e.g., advanced cryptography and virtual private networks. This is an example of how technology can both foster new pathways to accountability and disrupt data retrievability. In this context, the transparent nature of (public) blockchains makes them vulnerable to insufficient data privacy, de-anonymization attacks, and surveillance. While de-anonymization is often perceived negatively, it can be applied for investigative purposes and to comply with rules that aim to mitigate specific risks, such as AML/CFT.

AML/CFT and blockchain forensics

The first concern of cryptocurrency misuse originated from transactions on the dark web. While a range of technologies aids darknet operations, cryptocurrencies, mainly Bitcoin and Monero, play a crucial role by facilitating payments (Akhgar et al., 2021). While Bitcoin is still the major player, used by 93% of darknet markets, the adoption of Monero is increasing: 67% of platforms supported it in 2021 vis-`a-vis 45% in 2020, and some support it on an exclusive basis (Chainalysis Team, 2022). Nonetheless, the public perception of crypto-related laundering is likely inflated. Indeed, even if the value of illicit crypto transactions reached an all-time high in 2022, hitting USD 20.6 billion, it accounts only for 0.24% of crypto activity (Chainalysis Team, 2023), and remains small when compared with criminal activities involving fiat currencies (CipherTrace, 2021; Goforth, 2020).[5]

Since the risk-based approach informs AML/CFT obligations, regulated entities must tune compliance efforts: stricter measures if risk factors are higher. The end goal is to draw authorities' attention when suspicions of illicit activities arise by filing a report when the entity knows, suspects, or has reasonable ground to suspect the given funds are the proceeds of a criminal activity or are related to terrorist financing (Directive (EU) 2018/843, 2018; FATF, 2022). Generally, AML/CFT duties apply to crypto-transactions, and cryptoasset service providers are increasingly regulated. In the EU, the 5th AML Directive (EU) 2018/843 (2018) first targeted these activities, and the regime is evolving with the AML Package (European Commission, 2021).

Even if the ledger transparency featured by public blockchains mitigates the risk of fraudulent behavior, the technology is vulnerable to unpredictable exploitation methods (Shayegan et al., 2022; Xu, 2016). This prompted the development of specific techniques of anomaly detection. In this field, the Internet of Money's opaque reputation appears paradoxical since it provides a huge amount of open-source intelligence—e.g., it is possible to extract data from a given transaction and retrieve the history of an address, while methods using networks created by transactions (i.e., "transaction flow analysis") can define patterns to pinpoint suspected addresses (Wu et al., 2021). Different analytic techniques have been refined over time (Yin et al., 2019), and mostly rely on statistical approaches—e.g., the re-use of an account for more transactions or the co-use of more accounts for a single transaction can lead to matching more accounts to the same user (Li et al., 2021).

Starting from 2020, a surge of ransomware attacks highlighted regulatory shortcomings concerning the complex development of the Internet of Money. Indeed, as the latter becomes populated by AECs and other services that increase obfuscation, the risks of fraud increase. More recently, in 2022, hackers stole USD 3.1 billion from DeFi protocols, exploiting their transparency—i.e., typically, DeFi transactions happen on-chain and the smart contract code is publicly viewable. This amount accounts for 82% of all crypto funds stolen by hackers. In the same year, crypto mixers processed USD 7.8 billion, 24% of which originated from illicit addresses (Chainalysis Team, 2023).

To guide regulated entities in the management of their exposures, several authorities publish red flag/risk indicators to guide compliance and supervision. Notably, in the indicators published by the global AML/CFT standard-setter Financial Action Task Force (FATF) there is a section on anonymity risks (FATF, 2020),[6] updated in 2021 (FATF, 2021).

[5] It is worth noting that in 2022 the share of crypto activity associated with illicit activity rose for the first time since 2019. However, 43% of the illicit transaction volume is linked to sanctioned entities. Notably, for the most part, to the crypto exchange Garantex (Chainalysis Team, 2023).

[6] The report targets six types of indicators, relating to (i) transactions, (ii) transaction patterns, (iii) anonymity, (iv) senders/recipients, (v) funding/wealth at source, (vi) geographic risks.

Although a transaction's anonymity level is insufficient to suggest the transfer is suspicious, the FATF underlined inherent issues of privacy-enhancing technologies implemented by privacy coins, such as ZKPs (FATF, 2020). At the same time, a range of institutions highlighted the risks caused by unhosted wallets (Chainalysis Team, 2023; Europol, 2020).

Against this backdrop, forensic methods provide a wide range of information that emerges as pivotal for investigation, compliance, and supervision. Their value is displayed by the debate on the crypto travel rule, pursuant to which regulated entities must identify originators and recipients of cryptotransfers to guarantee traceability. In principle, this is just an expansion of data sharing measures previously applicable only to wire transfers, as required by the FATF Standards and by EU measures part of the AML Package. However, the reactions to the crypto travel rule exemplify the tension between the Internet of Money and an intermediary-based regulatory framework that still has to capture the specifics of peer-to-peer transfers and decentralized platforms. Accordingly, the industry denounces the absence of global standards and technical solutions to underpin effective and affordable compliance.

Related work

While we do not aim to offer a review of the techniques of cryptocurrency forensics, in this section, we describe a few works that provided an application to the concepts introduced above. In blockchain analytics, various methods aim to link pools of addresses and transactions. They can deploy clustering techniques to group addresses owned by the same user (Ince et al., 2018; Neudecker & Hartenstein, 2017; Wu et al., 2021) and also leverage transaction graphs to explore the features of the network (Al Jawaheri et al., 2020; Fleder et al., 2015; Ober et al., 2013; Weber et al., 2019). Some of these approaches aim to identify idioms of use in the network that can erode anonymity (Meiklejohn et al., 2016), while others screen transactions to/from crypto-wallets to classify transactions as licit or illicit (Weber et al., 2019). In principle, these tools do not directly try to link addresses and transactions to real-world identities. However, if one of them is de-anonymized (in other ways), they allow to de-anonymize the whole cluster, as the cluster database allows fast correlation. Likewise, the goal usually is not to identify transaction patterns, but to allow that when an addresses is suspected other addresses of the same cluster can be suspected as well (Wu et al., 2021).

Clustering methodologies are based on heuristic models (Lischke & Fabian, 2016; Reid & Harrigan, 2013), such as: if two/more addresses are inputs to the same transaction, they are controlled by the same user (Meiklejohn et al., 2016). In wallet-closure analysis the heuristics are applied to establish a unique mapping between addresses and an identity (Al Jawaheri et al., 2020). In behavior-based clustering (Yin et al., 2019), addresses are grouped based on patterns such as transaction values (Amarasinghe et al., 2019). Androulaki et al. (2013) showed this could unveil the profiles of 40% of Bitcoin users despite privacy measures.

On the application level, analytic techniques can exploit the possibility to correlate transactions with users' information on social media. Frequently, users post their addresses (e.g., to receive donations) but also reveal personal information (e.g., contact information, age, location) (Al Jawaheri et al., 2020). In this respect, transaction fingerprinting methods can make use of off-network data (Reid & Harrigan, 2013), which is also leveraged by web-scraping and Open Source Intelligence tools. Fleder et al. (2015) annotated the transaction graph by linking user pseudonyms to online identities collected from social media and developed a graph-analysis framework to summarize and cluster users' activity to link identities and transactions.

Specific methods target mixing services (Wu et al., 2020), i.e., the ones that shuffle coins by sending them to different addresses to obfuscate the flow. Although third-party services act as centralization points, thus aiding traceability, new disintermediated methods such as CoinJoin (Al Jawaheri et al., 2020) deploy more sophisticated shuffling approaches. In this context, an important role is played by peer-to-peer cross-chain transfers, and a relatively new subset of analytic efforts aims to trace cross-currency transfers through exchanges such as ShapeShift (Al Jawaheri et al., 2020). Harrigan and Fretter (2016) clustered the addresses of the whole Bitcoin blockchain to show that the methodology remains effective despite mixed transactions.

Another line of forensic research, further discussed in the "Anomaly detection approaches" section, is based on machine learning. Yin et al. (2019) presented a supervised learning-based approach to de-anonymize the Bitcoin blockchain to predict the type of entities yet not identified. They built classifiers concerning 12 categories and concluded that it is possible to predict the type of an entity. To do so, they collaborated with the analytic company Chainalysis that provided the data and had previously clustered, identified, and categorized a considerable number of addresses manually or through clustering techniques. They show two examples, one where they predict a set of 22 clusters suspected to be related to criminal activities, and another where they classify 153,293 clusters to provide an estimation of Bitcoin activity. Furthermore, they concluded it is possible to predict if a cluster belongs to predefined categories such as exchange, gambling, merchant services, mining pool, mixing, ransomware, and scam.

Machine learning solutions benefit from constructing multiple graph types from blockchain data, e.g., a blockchain

Table 1 Summary of the features and comparison of related works with our work

Work	Methodology	Algorithms	Results
Reid and Harrigan (2013)	Network analysis	Flow analysis + off-network information	Associate addresses with each other and with external identifying information
Fleder et al. (2015)	Network analysis	Flow analysis + web scraping	Link illicit activities to online identities
Wu et al. (2021)	Network analysis	Safe Petri Net-based cluster analysis	Find suspected addresses
Al Jawaheri et al. (2020)	Network analysis	Wallet-closure analysis	Infer links between Bitcoin users and hidden services
Harrigan and Fretter (2016)	Network analysis	Address-clustering analysis	Identify super-clusters
Sun et al. (2021)	Graph analysis	Flow-based graphs analysis with coupled tensors	Anomalous transactions detection FAUC metric 0.94
Li et al. (2020)	Graph analysis	Theoretical flow-based multipartite graphs analysis	Anomalous transactions detection FAUC metric 0.96
Yin et al. (2019)	Machine learning	Supervised learning-based (baseline)	Predict type of yet-unidentified entity F1score 0.796 (GradientBoosting)
Weber et al. (2019)	Machine learning + Graph analysis	Supervised learning-based (baseline + GCN)	Predict illicit transactions F1score 0.796 (Random Forest)
Eddin et al. (2021)	Machine learning + Graph analysis	Supervised learning-based (baseline + triage model)	Reduce the number of false positives by 80%
Oliveira et al. (2021)	Machine learning + Graph analysis	Supervised learning-based (baseline + GuiltyWalker)	Predict illicit transactions F1score 0.85 (Random Forest)
Ours	Machine learning + Graph analysis	Supervised learning-based (baseline + GCN + GAT)	Predict illicit transactions F1score 0.844 (GCN)

account (or a group of) is a node, and a single transaction between two accounts is an edge. An edge's weight is then defined as the aggregate transaction volume over a period of time. The latter is the predominant crypto-related forensic method seen in the "AML/CFT and blockchain forensics" section (Weber et al., 2018). Relatedly, Weber et al. (2019) benchmarked GCN against various supervised methods. In contrast, Eddin et al. (2021) extended their work to reduce false alerts through supervised learning methods in a context not related to the Internet of Money. They call the machine learning component the "triage model," tasked to process the rule-generated alerts: the generated score enables alert suppression or prioritization. The GuiltyWalker (Oliveira et al., 2021) leverages random walks on a crypto-transaction graph to characterize distances to previous suspicious activity.

Table 1 shows a summary of the most influential research cited in this section. In this work, we aim to enhance the performance of classifier methods based on machine learning and graph analysis. To this end, we (i) adopt a novel scheme for transaction classification based on GAT; and (ii) resort to an updated implementation of GCN with respect to related works. As pointed out in the results section, this configuration improves state-of-the-art performance. Our methodology is backed up by an analysis of crypto-specific AML/CFT issues and anomaly detection approaches addressed in the next section. In particular, we consider the set of transactions and their inherent characteristics, i.e. the fact that to spend cryptocurrencies, a user needs to have received them from previous transactions. These dependencies allow the creation of a graph whose structure can help identify illicit transactions. However, the need arises to identify the criteria that can inform a proper transactions classification—e.g., defining how it is possible to state that if a transaction is illicit, its neighbor transactions are also illicit, or if any graph-specific patterns represent suspicious activities. To confront these issues, it is essential to have a clear understanding of anomaly detection approaches in the RegTech field.

Anomaly detection approaches

The process of anomaly/outlier detection involves processing data to detect behavior patterns that may indicate a change in system operations. The goal is to single out rare or suspicious events/items—i.e., those significantly different from the dataset (Kamišalić et al., 2021). While collective anomaly detection methods target groups of data points that differ from most of the data, point anomaly detection also considers single data points (Li et al., 2022; Shayegan et al., 2022). AML/CFT-regulated entities, especially in the financial industry, deploy RegTech solutions to screen their operations and detect anomalous activities in an automated

way. Their effort is based on the risk indicators provided by regulators usually in a rulebased format—i.e., templates of sequences of actions that suggest a suspicion in a way that is self-explainable and interpretable. Indeed, compliance decisions must be explainable and traceable for auditing. For this reason, the preliminary review of a flagged account relies on suspiciousness heuristics (e.g., political exposure, geographic location, transaction type, users' behavior) (Weber et al., 2018). This is the case of the mentioned FATF's indicators, developed from analyzing 100 + case studies from 2017 to 2020 (FATF, 2020). Rule-based red flags can pertain to transaction patterns, such as "incoming transactions from many unrelated wallets in relatively small amounts (accumulation of funds) with subsequent transfer to another wallet or full exchange for fiat currency," or to anonymity, such as "moving a VA that operates on a public, transparent blockchain, such as Bitcoin, to a centralized exchange and then immediately trading it for an AEC or privacy coin" (FATF, 2020). In particular, indicators related to anonymity include cases of enhanced obfuscation (e.g., AECs) and disintermediation (e.g., unhosted wallets).

In this context, a lot of time and resources are needed to investigate alerts generated by rule-matching processes and decide when to report a transaction as suspicious. An alert can be a true or a false positive, and arguably the simplicity of rule-based systems, despite guaranteeing interpretability, produces an estimate of around 95–98% false positives (Eddin et al., 2021).

Indeed, classifying entities and discovering patterns in massive time-series transaction datasets that are dynamic, high dimensional, combinatorially complex, non-linear, often fragmented, inaccurate, or inconsistent is a challenging task. Moreover, the difficulty of automating the synthesis of information from multi-modal data streams thrusts the task onto human analysts. This adds to a vicious circle of a compliance approach that stimulates over-reporting due to the cost asymmetry between false positives and false negatives and overburdens law enforcement agencies (Weber et al., 2018). Hence, the automation of an increasing array of processes has been suggested (Oad et al., 2021).

Against this backdrop, in this section, we explore the anomaly detection methods that relate to our experiments. Hence, we focus on machine learning and graph analysis. We take an on-chain data analytic perspective, although we acknowledge the value of tools that target off-chain data, such as Natural Language Processing and sentiment analysis, that also leverage graph methods (Weber et al., 2018). Indeed, while cryptocurrency transactional data is often analyzed through a combination of on-chain and off-chain techniques, thus including information not recorded on the blockchain or recorded on a different blockchain, in this work, we focus on on-chain data.

Machine learning

Machine learning is a part of artificial intelligence that exploits data and algorithms to imitate human learning processes with gradual accuracy improvements. This helps us find solutions to problems in many fields, e.g., vision, speech recognition, robotics (Alpaydin, 2020). In the most diverse contexts, it provides tools that can learn and improve automatically leveraging the vast amount of data available in our age (Kamiˇsali´c et al., 2021). In the compliance domain, advances in these algorithms show great promise, and their deployment in AML/CFT RegTech solutions can improve the efficiency of these applications (Weber et al., 2019). For instance, they can mitigate the shortcomings of rules-based systems and infer patterns from historical data, increasing detection rates and limiting false positives (Lorenz, 2021). In other cases, a more proactive approach is deployed to map and predict illicit transactions (Koshy et al., 2014; Weber et al., 2019).

One of the main distinctions in machine learning is between unsupervised methods, where the model works on its own to discover patterns and information previously undetected, and supervised techniques, where labeled datasets are used to train algorithms. While applying both methods for anomaly detection is possible, most systems deploy unsupervised techniques due to a lack of relevant real-world labeled datasets. In the AML/CFT sphere, this scarcity mainly derives from difficulties in labeling real cases timely and comprehensively. Indeed, manual labels are costly in terms of time and effort, and the nature of the entities involved is complex and ever-evolving (Lorenz, 2021). Hence, analytic companies play a key role in labeling crypto transactions. In order to address the overall lack of data, various strategies have been proposed (Eddin et al., 2021): generate a fully synthetic dataset, simulate only unusual accounts within a real-world dataset, and localize rare events within a peer group. However, better validations of the systems were obtained using analyst feedback or real-labeled data. Parallelly, the dataset shortage has driven the deployment of active learning (i.e., few labels) (Lorenz, 2021).

Supervised baseline techniques

Supervised learning techniques are leveraged for their labeled training data. For instance, they are used to classify anomalies based on association rules to detect suspicious events (Luo, 2014). In the AML/CFT context, the label of each transaction could indicate whether it was identified as money laundering or not (Lorenz, 2021). Recent RegTech solutions deploy widespread supervised learning methods to perform anomaly detection (Yin et al., 2019):

- **Decision tree**—It is one of the base algorithms used in machine learning, with a name derived from a hierarchical model formed visually as a tree where nodes are decisions with specific criteria. The training data is subdivided into subsets following the tree branches. The node decision criteria are determined variables that can be defined as explanatory. The algorithm tries to apply the most significant feature to perform the best division among the training data. The best division can be measured by the information gain, mathematically derived from a decrease in entropy (Alpaydin, 2020).
- **Random forests**—It is an extension of Decision Trees in which an algorithm approaches the classification task by constructing a multitude of trees. Introduced by Breiman (2001), it is an ensemble method applied to sample random subsets of the training data for each Decision Tree. It aims to improve the predictive accuracy of a classifier by combining multiple individual weak learners, i.e., trees.
- **Boosting algorithms**—They are another ensemble method that fits weak learners' sequences. A boosting algorithm tries to boost a Decision Tree by recursively selecting a subset of the training data. AdaBoost (Adaptive Boosting) assigns weights to the data samples based on the weak learners' ability to predict the individual training sample. Thus, the sample weights are individually computed for each iteration, and the successive learner is applied to the new data subset (Yin et al., 2019).
- **Logistic regression**—It is a multiple regression suitable for binary classification, which assesses the relationship between the binary dependent variable (target) and a set of independent categorical or continuous variables (predictors) (Hilbe, 2009). It can be seen as measuring the probability of an event happening, where the probability consists of the ratio between the probability that an event will occur and the probability that it will not.
- **Support vector classification (SVC)**—Given a set of data for training, each labeled with the class to which it belongs among the two possible classes, a training algorithm for Support Vector Machines builds a model that assigns the new data to one of the two classes. This generates a nonprobabilistic binary linear classifier. This model represents data as points in space, mapped in a way that a space separates data belonging to the two categories as ample as possible. New data is then mapped in the same space, and the prediction of the category to which they belong is made based on the side in which they fall (Alpaydin, 2020).
- **K-nearest neighbors (k-NN)**—It is a supervised learning algorithm used in pattern recognition for object classification based on the characteristics of the objects close to the considered one. The model represents data as points in space, i.e., the feature space. Given a notion of distance between data objects, the input is the k nearest training data in the feature space. The underlying idea is that the more similar the instances, the more likely they belong to the same class (Alpaydin, 2020).

Graph analysis

In recent years, a portion of machine learning research focused on real-world datasets that come in graphs or networks—e.g., social networks, knowledge graphs—to generalize learning models to such structured datasets. Graph analytics is becoming increasingly important for AML/CFT, because money laundering involves flow relationships between entities that create graph structures. Some approaches for supervised learning work with graph-structured data based on a variant of neural networks which operate directly on graphs, i.e., graph neural network (You et al., 2020; Kipf & Welling, 2016). Convolutional neural networks, for instance, offer an efficient architecture to extract significant statistical patterns in large-scale and high-dimensional datasets and can be generalized to graphs (Defferrard et al., 2016; Kipf & Welling, 2016). In this work, we use two specific graph-based neural networks, i.e., Graph Convolutional Networks (GCN) and Graph Attention Networks (GAT). These techniques are described in the next section.

Experimenting with machine learning

As contextualized above, AML/CFT analytics benefit from deploying machine learning-based techniques for transaction classification. However, in new techniques, there is the need to balance interpretability and explainability with the reduction of false positives and over-reporting. Accordingly, this section outlines the experimental setup of our study and the relevant results. After describing the dataset, we consider the evaluation method and the implementation of the anomaly detection approaches. Subsequently, we compare the results of our experiments, where state-of-the-art machine learning techniques and graph-based neural networks are employed in an AML/CFT context.

It is worth noting that, in developing this work, we heed several assumptions. Although we have already discussed these throughout the text, we provide the following summary. In our paper, (i) the term Internet of Money refers to the entire set of cryptocurrency ecosystems; (ii) we do not offer a comprehensive review of crypto forensic techniques; (iii) we focus on on-chain data; and (iv) we perform inevitable simplifications when addressing explainability and interpretability of AI applications and relevant legal impacts.

Methodology

Our experimentation is grounded on a seminal work by Weber et al. (2019). Most of the techniques deployed in the study correspond to the standard supervised models mentioned above—i.e., Decision Trees, Logistic Regression, k-NN, SVC, AdaBoost, Random Forests—used as benchmark methods for classification. However, the two graph-based models GCN and GAT deserve close attention in this context. This is for three main reasons: (i) these types of neural networks take into account the graph nature of our dataset; (ii) as the evaluation shows, our application of GCN outperforms benchmark approaches and improves the state of the art; and (iii) to the best of our knowledge, this is the first attempt to deploy the GAT model in the AML/CFT context.

Transaction graph analysis

Graphs represent a typical mathematical tool to model interactions among different entities: humans, elements of a biological system, computing nodes in a distributed system, and others (Pocher & Zichichi, 2022). In a blockchain, transactions are linked by nature since money spent in a transaction originates from previous transfers (Pocher & Zichichi, 2022). This allows the creation of a graph of transactions that can help the classification process. In fact, given a transaction t, it is possible to collect all the connected transactions and recursively search for other ones up to a certain depth level. Given such a connected graph centered at t, an inspection of the neighboring transactions and their classified value can aid the classification of t. Each node of the graph (transaction) has thus a set of neighbors that will influence its classification. Moreover, each node has a set of features associated with the corresponding transaction (see below for the details of the dataset).

An example of this procedure is displayed in Fig. 1, where a connected component—i.e., a subgraph in which each pair of nodes is connected via a path—is obtained from an initial transaction (Fig. 1, top). In the figure, the *red* nodes represent transactions labeled as illicit in the starting dataset, the *green* ones licit transactions and the *grey* ones are still unknown/unlabeled. To show the output of a machine learning classification problem, the bottom part of Fig. 1 shows the output of the process employing a specific classification algorithm, which in this case is Random Forest. In essence, the idea is that knowing the labels of certain transactions aids the classification of the remaining (unknown/unlabeled) ones. Hence, learning methods could pinpoint illicit transactions based on the graph topology and the features of the transactions.

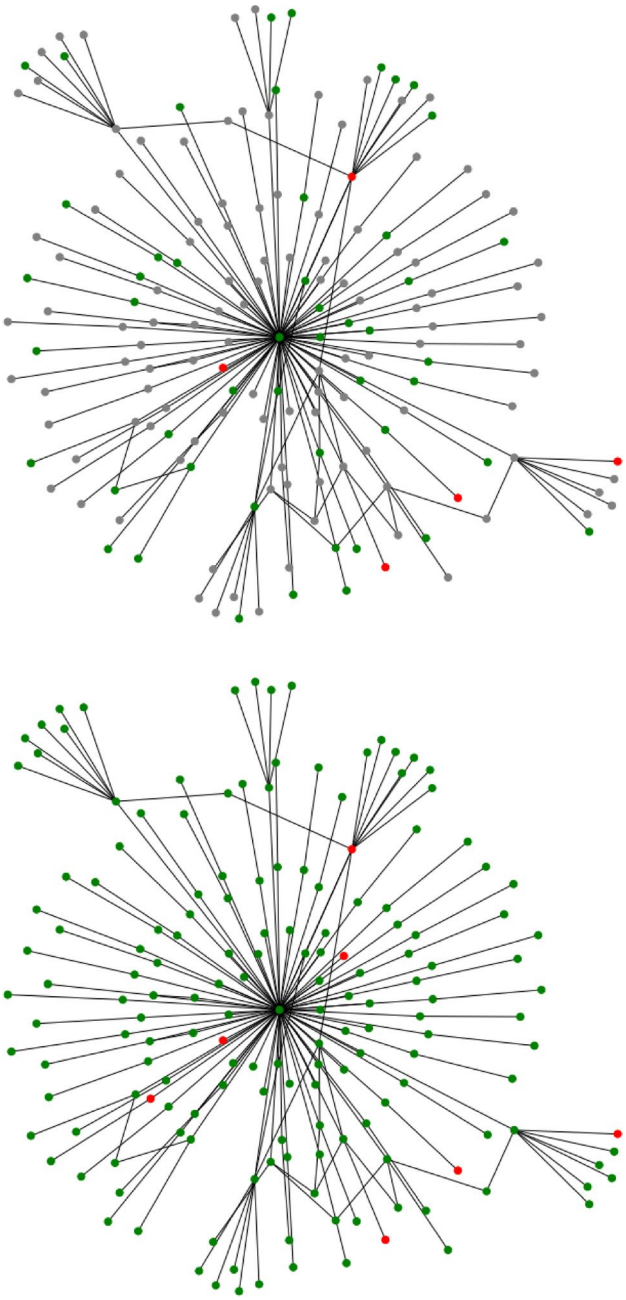

Fig. 1 Connected graph of a considered transaction before and after classification

Dataset

In our work, we experimented with the publicly available Elliptic transactions dataset provided in the context of Weber et al. (2019). For details on the dataset, the reader can refer to the latter and to the description provided on Kaggle

together with the dataset.[7] This dataset contains real Bitcoin transactions represented as a directed graph network, where transactions are nodes, and the directed edges between these transactions represent fund flows from the source address to the destination address. The dataset contains 203,769 transaction nodes connected by 234,355 edges. For each transaction, 167 features are available, of which the first 94 relate to the transaction itself and thus directly extracted from the blockchain—e.g., the number of inputs of a transaction or the number of outputs—while the other 73 features relate to the graph network itself and are extracted from the neighboring transactions of a node. The features do not have any associated descriptions—indeed, Weber et al. (2019) claim that they cannot describe these features due to intellectual property issues. Tests were carried out with transaction features (tx) and transaction features plus aggregated features (tx + agg). Such aggregated features are obtained by aggregating transaction information one-hop backward/forward from the center transaction node. This means obtaining the features of the nodes that share an edge with that transaction node.

Each transaction in the dataset is labeled as illicit, licit, or unknown: 4545 are labeled as illicit, 42,019 are labeled as licit, and the remaining 157205 are unknown. The transactions also contain temporal data. In particular, this is grouped into 49 distinct time steps, evenly spaced the interval of 2 weeks. Each time step contains a connected graph that includes all the transactions verified on the blockchain in the span of 3 h (Weber et al., 2019).

The dataset was pre-processed as follows: (i) the features were merged with the classes; (ii) class values were renamed to integer values; (iii) transaction identifiers were swapped for a sorted index; (iv) only the part of the dataset labeled licit or illicit was selected; and (v) all the edges between unknown transactions were removed. After the pre-processing, our cleaned dataset encompassed 46,564 transactions and 36,624 edges.

Graph convolutional network model architecture

The objective of a GCN model is to learn a function of signals/features on a data set structured as a graph. The model takes as input (i) a graph with nodes and edges between nodes and (ii) a feature description for each node. The key idea is that each node receives and aggregates features from its neighbors to represent and compute its local state. The GCN then usually produces an output feature matrix at the node level (Kipf & Welling, 2016). The GCN model is used for transaction classification because it is a deep neural network that allows capturing the relation among the nodes and their neighborhoods. In other words, it creates a node embedding in a latent vector space that captures the characteristics of the node neighborhood in the graph. This information comes in the form of a look-up table mapping nodes to a vector of numbers. GCNs have been developed using the Keras framework, following the recommendations introduced in You et al. (2020).

The general structure of our graph convolution layer is made of three steps. First, the input node representations are processed using a Feed Forward Network to produce a message. Second, the messages of the neighbors of each node are aggregated using a permutation invariant pooling unsorted segment sum operation. Third, the node representations and aggregated messages are combined and processed to produce the new state of the node representations (node embeddings) via concatenation and Feed Forward Network processing.

Our network architecture consists of a sequential workflow of the model that we display in Table 2 and summarized as follows:

1. Apply pre-processing using Feed Forward Network to the node features to generate initial node representations;
2. Apply two graph convolutional layers, with skip connections, to the node representation to produce node embeddings;
3. Apply post-processing using Feed Forward Network to the node embeddings to generate the final node embeddings;
4. Feed the node embeddings in a Softmax layer to predict the node class.

Graph attention network model architecture

While the GCN model averages the node states from source nodes to the target node, the GAT model gives different importance to each node's edge by using an attention mechanism to aggregate information from neighboring nodes

Table 2 GCN model architecture. Total parameters = 18,774, trainable parameters = 17,756, non-trainable = 1018

Layer (type)	Output shape	Num. parameters
Preprocess (Sequential)	(46,564, 32)	4564
Convolution 1 (GraphConvLayer)	multiple	5888
Convolution 2 (GraphConvLayer)	multiple	5888
Postprocess (Sequental)	(46,564, 32)	2368
Logits (Dense)	multiple	66

[7] Elliptic dataset: https://www.kaggle.com/datasets/ellipticco/elliptic-data-set.

(Veličković et al., 2017). In other words, instead of simply averaging/summing node states from source nodes to the target node, as we do in the GCN model, GAT, on the other hand, first applies normalized attention scores to each source node state and then sums (Veličković et al., 2017).

Our model is built using the Keras framework that, through a graph attention layer that computes pairwise attention scores, aggregates and applies the scores to the node's neighbors. A multi-head attention layer concatenates multiple graph attention layer outputs. Our design choice is to use a single attention layer with multiple heads, enabling the network to jointly attend multiple positions (Liyuan Liu and Liu, 2021). The multi-head layer is then inserted into a general model that implements dense pre-processing/post-processing layers with dropout regularization, as shown in Table 3. The training proved to be subjected to overfitting, and heavy regularization was necessary, which was achieved by dropout layers and using RMSprop optimizer with momentum (Philipp et al., 2017).

Results

For the discussion of the results, we firstly consider the illicit class as, due to the nature of the dataset (less labeled illicit transactions) and of the problem, its classification is more complex. To compare the results, we use the F1-score, a metric obtained from Precision and Recall. These metrics are usually defined for a binary classifier (as in this case) where some special instances need to be identified, e.g., positive cases to a particular test. Precision is the number of true positive (TP) predictions, i.e., how many of the positive predictions made are correct over the sum of TP and false positives (FP). In other words, precision says how many of the identified illicit transactions were illicit.

$$Precision = \frac{TP}{TP + FP.}$$

Recall measures how many positive cases the classifier correctly predicted over all the positive cases in the data, i.e., TP and false negatives (FN). In our context, for instance, it allows us to understand how many illicit transactions the classifier identified over the real considered set of illicit transactions.

$$Recall = \frac{TP}{TP + FN}.$$

The F1-score represents the harmonic mean of Recall and Precision and is thus calculated as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

We also use the Micro Average F1-score for evaluating the methods. It measures the F1-score of the aggregated contributions of all classes.

The final performance results are reported in Fig. 2 and Table 4. It is possible to observe how GCN outperforms other approaches. In particular, the GCN approach provides the best results in terms of recall, i.e., 0.790, and F1-score, i.e., 0.844. In terms of precision, it slightly deviates from the Decision Tree (0.986) and Random Forest (0.981) approaches

Table 3 GAT model architecture. Total parameters = 59,952, trainable parameters = 59,952, non-trainable = 0

Layer (type)	Output shape	Num. parameters
Dense 9 (Dense)	Multiple	10,340
Dropout 6 (Dropout)	Multiple	0
Graph attention (MultiHead-GraphAttention)	Multiple	12,320
Dense 10 (Dense)	Multiple	36,630
Dropout 7 (Dropout)	Multiple	0
Dense 11 (Dense)	Multiple	662

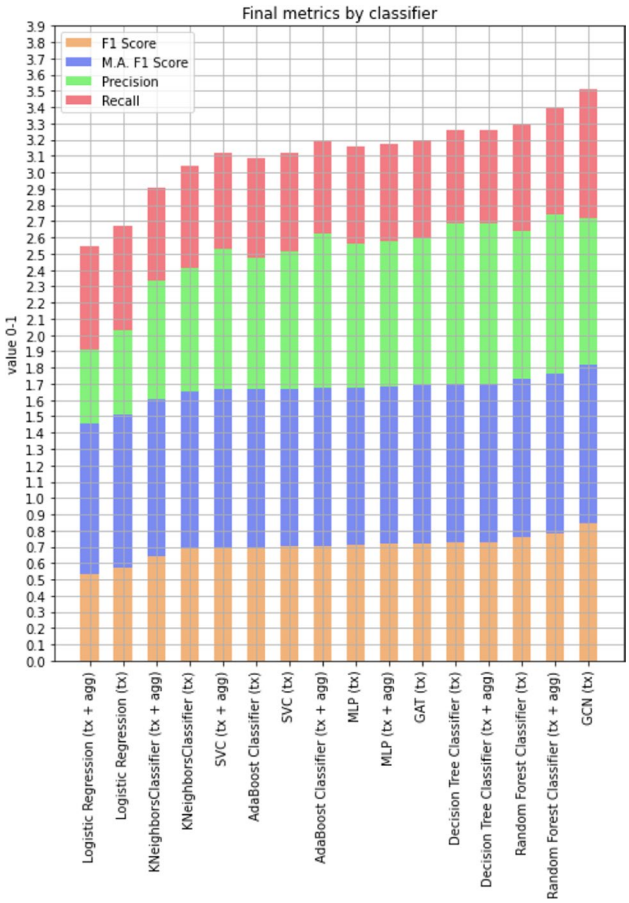

Fig. 2 Barplot aggregating F1-score, micro average F1-score, precision, and recall for all the approaches experimented

Table 4 Table showing the results for **illicit** transaction classification with the F1-score, Micro Average F1-score, precision, and recall metrics for all models

Model	Precision	Recall	F1 score	M.A. F1
Random Forest classifier (tx)	0.909	0.648	0.757	0.974
Random Forest classifier (tx + agg)	0.981	0.651	0.782	**0.977**
Logistic regression (tx)	0.515	0.646	0.573	0.939
Logistic regression (tx + agg)	0.456	0.630	0.529	0.929
MLP (tx)	0.897	0.593	0.714	0.970
MLP (tx + agg)	0.817	0.623	0.707	0.968
k-NN classifier (tx)	0.762	0.629	0.689	0.964
k-NN classifier (tx + agg)	0.730	0.576	0.644	0.960
SVC (tx)	0.842	0.604	0.703	0.968
SVC (tx + agg)	0.862	0.588	0.699	0.968
Decision Tree classifier (tx)	**0.986**	0.573	0.725	0.973
Decision Tree classifier (tx + agg)	**0.986**	0.573	0.725	0.973
AdaBoost classifier (tx)	0.793	0.615	0.693	0.966
AdaBoost classifier (tx + agg)	0.945	0.567	0.708	0.971
GCN (tx)	0.906	**0.790**	**0.844**	0.973
GAT (tx)	0.897	0.605	0.723	0.971

Best results are highlighted in bold

but still provides better performances than all the rests, i.e., 0.906. For what concerns the Micro Average F1-score, all approaches fit in the range of 0.960 to 0.977. These results are in contrast with the results in Weber et al. (2019), where Random Forests provided the best performance. The cause of such improvement might be due to the different architectures we exploited to build the neural network.

Furthermore, the comparison with the other graph-based approach, i.e., GAT, also sees the GCN outperforming. GAT performs better than a simple dense network but cannot reach the results of GCN and Random Forest classifiers. The motivation is probably due to the na¨ıve structure of the neural network, and its optimization is currently under investigation.

So far, we have focused on the classification of illicit transactions. Since the dataset is unbalanced—i.e., it contains more licit than illicit transactions in a ratio of more or less 1 to 10—the problem becomes relatively trivial. For the sake of transparency, however, the final performance results related to licit transactions are reported in Table 5. All the models perform very well and are very similar to each other. In this case, graph-based approaches do not perform better than Random Forest classifiers, but the difference is not outstanding.

Discussion

When interpreted through the lens of the AML/CFT remarks outlined in the previous sections, our findings inspire multi-layered considerations. Accordingly, in this section, our reasoning is threefold. First, we discuss the results of our experiments vis-`a-vis the approaches used as benchmarks. Secondly, we broaden the perspective of the analysis to consider not only the impacts of crypto-related RegTech methodologies on the evolution of the Internet of Money, but also the interplay between the latter and the prospective role of forensics. Finally, we pinpoint a few associated challenges.

From the first perspective, to the best of our knowledge, the experiment described in this paper is the first attempt to implement GAT models to detect anomalies in Bitcoin transactions for AML/CFT purposes. The final results are on par with the state of the art of GCN networks, with GAT marginally worse than GCN. This could be explained by the "simpler" implementation of GAT and the possibility that the dataset responds better to non-spectral methods. Nonetheless, we argue that the novelty of this application could be helpful for general research on GAT anomaly detection techniques. In addition, the results show that the GCN neural network typology is a promising solution for AML/CFT, as it performs better than other approaches.

In this context, it is essential to consider that GCN and GAT classifiers only have access to transaction features, which means that all information about aggregated nodes comes from the graph structure itself. Since the performance of GCN is in line with Random Forest (with aggregated features), we can claim that our graph networks can obtain the same amount of information as the creator of the dataset (Weber et al., 2019). However, choosing one method over another carries additional implications that must be carefully weighed. For example, the performance of Random Forests falls slightly behind GCN's, but there is no sacrifice in explainability because the detectors are derived from Random Forests' rules (Eddin et al., 2021). Given the size and dynamism of real-world information, explainability of the results is challenging to provide, both in

Table 5 Table showing the results for **licit** transaction classification with the F1-score, Micro Average F1-score, precision, and recall metrics for all models

Model	Precision	Recall	F1 Score	M.A. F1
Random Forest classifier (tx)	0.977	0.995	0.986	0.973
Random Forest classifier (tx + agg)	0.977	**0.999**	**0.988**	**0.978**
Logistic regression (tx)	0.976	0.959	0.967	0.939
Logistic regression (tx + agg)	0.975	0.949	0.962	0.929
MLP (tx)	0.973	0.995	0.984	0.970
MLP (tx + agg)	0.974	0.994	0.984	0.970
k-NN (tx)	**0.978**	0.967	0.972	0.949
k-NN (tx + agg)	0.975	0.965	0.970	0.944
SVC (tx)	0.974	0.992	0.983	0.968
SVC (tx + agg)	0.973	0.994	0.983	0.968
Decision Tree classifier (tx)	0.972	**0.999**	0.986	0.973
Decision Tree classifier (tx + agg)	0.972	**0.999**	0.986	0.973
AdaBoost classifier (tx)	0.975	0.989	0.982	0.966
AdaBoost classifier (tx + agg)	0.972	0.998	0.985	0.971
GCN (tx)	0.975	0.994	0.984	0.971
GAT (tx)	0.973	0.992	0.982	0.967

Best results are highlighted in bold

this context and in the broader AI field. Even in our specific narrow instance—i.e., transaction graphs that model illicit activity over time—it is challenging to apply efficient methods whose results can be understood by humans. Although this appears to be a crucial aspect, the literature still lacks some research on the application of explainable AI techniques for AML/CFT anomaly detection (Kute et al., 2021).

From the second perspective, the choice of the forensic approach(es) to deploy must be made taking into consideration the evolution of the Internet of Money, with specific regard to peer-to-peer transfers and DeFi protocols. Indeed, while its developments warrant the application of increasingly sophisticated yet explainable compliance and investigation techniques, we see how the implementation of the crypto travel rule has already prompted the industry to denounce the lack of global standards and technical solutions to underpin effective and affordable compliance. It follows that, while the great quantity and complexity of transaction data to be processed suggests that machine learning will continue to be a part of the solution—with marginal performance differences possibly bearing significant weight when various approaches are combined—it is crucial to back the relevant research with a constructive dialog between the stakeholders involved. In this context, we point out to the increase in the laundering-related use DeFi protocols of 1.964% between 2020 and 2021. In 2021, centralized exchanges received 47% of funds originating from illicit addresses and DeFi protocols 17%, vis-`a-vis 2% in 2020. Likewise, in 2021 funds derived from cryptocurrency thefts were increasingly sent to DeFi platforms (51%) or risky services (25%), while only 15% went to centralized exchanges, possibly due to AML/CFT (Chainalysis Team, 2022). In 2022, almost half of illicit crypto funds passed through a set of intermediary services primarily populated by mixers, illicit services, and DeFi protocols. However, 67% of illicit funds received by exchanges went to only five centralized exchanges, in comparison to 56.7% of 2021 (Chainalysis Team, 2023).

In the near future, regulated entities, law enforcement and supervisors will increasingly need to monitor and analyze crypto transactions to which multilayered obfuscation techniques have been applied. In addition, given the rise in the use of unhosted wallets and decentralized platforms, they will frequently operate without the assistance of centralized counterparty entities. For these reasons, we wish to highlight the value of not only researching innovative machine learning-based forensic applications, but also to adopt an interdisciplinary approach to devise compliance tools that adequately consider the way regulatory regimes are conceived and enforced. For instance, we point to the importance of reconciling the duties placed on regulated entities, the available and prospective intelligence tools, and an AML/CFT regime that is so far inherently and explicitly intermediary-based, with compliance efforts guided by rule-based risk indicators. It is for this reason that our work contextualizes forensic methods into the specifics of risk indicators. Building on these arguments, we emphasize that AML/CFT hurdles cannot be solved by simply resorting to a sophisticated transaction classification scheme. On the contrary, this process needs to be nested into a broader framework to be effective.

Indeed, our analysis of machine learning methods was anchored to the mitigation of the drawbacks of current rule-based systems in terms of false positives and over-reporting. Relatedly, we find that the value of

experimenting with machine learning algorithms for RegTech purposes appears dependent mainly on the relationship between the given approach and the regulatory environment within which it is deployed. In other words, the efficiency of a specific algorithm can be assessed per se, but its effectiveness in an AML/CFT context heavily depends on the extent to which the structure of the model correctly mirrors the regulatory framework—e.g., it generates alerts that are deemed relevant by regulators and mitigates the current trends of over-reporting.

From the third perspective, the supervised classification analysis we conducted could be in theory applied to other types of blockchain and cryptocurrencies, being the analysis constrained on the high-level perspective of the cryptocurrency transactions' graph. However, there is the need for a labeled transaction dataset to build such a transaction graph. And the lack of open data further complicates the task. Indeed, we point out a few challenges identified during our investigation, related to the openness and availability of the datasets being discussed and the explainability of the results. We find it is overarching to confront these open issues and devise appropriate solutions or mitigating measures. On the one hand, our analysis suggests that it is difficult to address efficiency evaluations of machine learning-based AML/CFT tools for anomaly detection and transaction classification, since this feature appears to be increasing to the detriment of interpretability and explainability. On the other hand, it is evident from our studies that the labeled transaction datasets on which supervised learning algorithms are trained are largely proprietary. This does not only impact the development of new methods, but possibly also the transparency of the activity of supervisory bodies. That is, if the activity of the latter, just as the compliance effort of regulated entities, can be based only on the intelligence findings of solutions deploying proprietary algorithms. The interplay between the lack of explainability and the proprietary nature of the datasets suggests worrisome scenarios that call for further research. Hence, it is crucial to foster public–private synergies that can consider the AML/CFT context from a socio-technical, operational, and regulatory viewpoint.

Conclusions

Elaborating on the enthusiasm for the financial application of blockchain and DLTs that surged in the wake of Bitcoin's launch, today the Internet of Money comprises a diverse set of socio-technical systems under constant evolution—e.g., recently, DeFi schemes. Over the years, forensic techniques have been deployed to connect crypto addresses/transactions to real-world identities. This responds to the regulatory quest to ensure accountability through identification, a concept that sits at the core of AML/CFT compliance. Meanwhile, institutions and authorities drafted anomaly indicators to help with the identification of suspicious transfers in compliance with the risk-based approach. In this context, law enforcement agencies and supervisors, often supported by the private sector, began to apply forensic methods to track relevant transfers, as well as regulated entities started benefiting from innovative RegTech solutions that partially automate the detection of anomalous activities.

In this paper, we focused on these techniques from an on-chain data analytic perspective, with a specific focus on approaches based on machine learning and graph analysis. The use of these algorithms in AML/CFT RegTech solutions shows great promise to improve the efficiency of the latter and mitigate the significant drawbacks of current rule-based methodologies. To the best of our knowledge, what we described in this work is the first experiment with GAT models for AML/CFT anomaly detection in Bitcoin. The application of this type of neural network falls in line with the recent focus on deploying machine learning techniques that leverage the inherent structure of many real-world datasets that come in the form of graphs or networks. GCN and GAT models are informed by the idea of creating generalized learning models for these structured datasets, and indeed the one we analyzed consists of (real) Bitcoin transactions represented as a directed graph network.

To conclude, we provide three levels of considerations. From an operational standpoint, our results show that the mentioned graph-based methods perform better than the baseline approaches—e.g., GCN performs better than Random Forests, with GAT being marginally worse than GCN. This encourages further experimentations with the use of GCN neural networks for AML/CFT purposes, while the novelty of our approach could spur further research into GAT-based anomaly detection techniques. From a related methodological perspective, we argue that a constant experimentation with various forensic methods, possibly leveraging the value added by transaction graphs, is crucial to reap the full benefits of analytics in an ever-evolving context of application such as the Internet of Money. These explorations, however, must be backed by serious efforts to foster constructive public–private dialog regarding the openness and the availability of labeled transaction datasets.

From a final conceptual viewpoint, we emphasize that a holistic interpretation of the interplay between AML/CFT measures and the Internet of Money—i.e., one that heeds in a comprehensive fashion socio-technical, operational, and regulatory dynamics when defining the object of the analysis—is crucial to devise effective and possibly efficient

RegTech solutions. Indeed, the efficiency of a specific algorithm may not guarantee its effectiveness in an AML/CFT context, which depends on the extent to which the model responds to regulatory needs and generates relevant alerts. This relevance is influenced by regulatory, compliance, and supervisory needs, as affected by the evolution of the features of the Internet of Money. This holistic approach is especially valuable when it comes to transaction classification and anomaly detection, where a main challenge is the need to balance interpretability and explainability with the goal to reduce the share of false positives and over-reporting.

Funding Open access funding provided by Università degli Studi di Urbino Carlo Bo within the CRUI-CARE Agreement.

Data Availability The software and data is available at the following link: https://github.com/fmerizzi/GCN_detect_bitcoin_money_laundering.

References

Akhgar, B., Gercke, M., Vrochidis, S., & Gibson, H. (2021). *Dark Web Investigation*. Springer. https://doi.org/10.1007/978-3-030-55343-2

Al Jawaheri, H., Al Sabah, M., Boshmaf, Y., Erbad, A. (2020). Deanonymizing Tor hidden service users through Bitcoin transactions analysis. *Computers and Security*, *89*. https://doi.org/10.1016/j.cose.2019.101684.

Ali, O., Ally, M., Dwivedi, Y., et al. (2020). The state of play of blockchain technology in the financial services sector: A systematic literature review. *International Journal of Information Management, 54*, 102199.

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

Amarasinghe, N., Boyen, X., & McKague, M. (2019). A survey of anonymity of cryptocurrencies. *Acm International Conference Proceeding Series. Sydney: Association for Computing Machinery.* https://doi.org/10.1145/3290688.3290693

Amler, H., Eckey, L., Faust, S., Kaiser, M., Schlosser, B. (2023). DeFining DeFi : Challenges and Pathway, 2021–2024. *2021 3rd Conference on Blockchain Research & Applications for Innovative Networks and Services (BRAINS).* https://doi.org/10.1109/BRAINS52497.2021.9569795

Androulaki, E., Karame, G. O., Roeschlin, M., Scherer, T., & Capkun, S. (2013). Evaluating User Privacy in Bitcoin. *LNCS, 7859*, 34–51. https://doi.org/10.1007/978-3-642-39884-14

Antonopoulos, A. M. (2017). *The internet of money - two*. Merkle Boom LLC.

Aramonte, S., Huang, W., Schrimpf, A. (2021). DeFi risks and the decentralisation illusion. *BIS Quarterly Review* (Dec), 21–36.

Barbereau, T., Smethurst, R., Papageorgiou, O., Sedlmeir, J., & Fridgen, G. (2023). Decentralised finance's timocratic governance: The distribution and exercise of tokenised voting rights. *Technology in Society, 73*, 102251.

Bartoletti, M., Carta, S., Cimoli, T., & Saia, R. (2020). Dissecting Ponzi schemes on Ethereum: Identification, analysis, and impact. *Future Generation Computer Systems, 102*, 259–277. https://doi.org/10.1016/j.future.2019.08.014

Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. *Interacting with Computers, 23*(1), 4–17. https://doi.org/10.1016/j.intcom.2010.07.003

Berg, A. (2019). The identity, fungibility and anonymity of money. *Economic Papers*(November), 1–16. https://doi.org/10.1111/1759-3441.12273.

Biryukov, A., & Tikhomirov, S. (2019). Deanonymization and linkability of cryptocurrency transactions based on network analysis. *Proceedings - 4th IEEE European Symposium on Security and Privacy, 2019*, 172–184. https://doi.org/10.1109/EuroSP.2019.00022

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chainalysis Team (2022). *The 2022 Crypto Crime Report.*

Chainalysis Team (2023). *The 2023 Crypto Crime Report.*

Chang, V., Baudier, P., Zhang, H., Xu, Q., Zhang, J., & Arami, M. (2020). How blockchain can impact financial services–The overview, challenges and recommendations from expert interviewees. *Technological Forecasting and Social Change, 158*, 120166. https://doi.org/10.1016/j.techfore.2020.120166

Chen, W., Zheng, Z., Ngai, E. C., Zheng, P., & Zhou, Y. (2019). Exploiting Blockchain Data to Detect Smart Ponzi Schemes on Ethereum. *IEEE Access, 7*(c), 37575–37586. https://doi.org/10.1109/ACCESS.2019.2905769

CipherTrace (2021). *Cryptocurrency crime and anti-money laundering report.* ciphertrace. https://ciphertrace.com/cryptocurrency-crime-and-anti-money-laundering-report-august-2021/

Defferrard, M., Bresson, X., Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, *29*.

Desmond, D. B., Lacey, D., & Salmon, P. (2019). Evaluating cryptocurrency laundering as a complex socio-technical system: A systematic literature review. *Journal of Money Laundering Control, 22*(3), 480–497. https://doi.org/10.1108/JMLC-10-2018-0063

Directive (EU) 2018/843 (2018). *Directive (EU) 2018/843 of the European Parliament and of the Council of 30 May 2018 amending Directive (EU) 2015/849 on the prevention of the use of the financial system for the purposes of money laundering or terrorist financing, and amending Directives 2009/138/EC and 2013/36/EU.*

Eddin, A.N., Bono, J., Aparício, D., Polido, D., Ascensão, J.T., Bizarro, P., & Ribeiro, P. (2021). Anti-money laundering alert optimization using machine learning with graphs. *Arxiv.* https://doi.org/10.48550/ARXIV.2112.07508.

Edmunds, J.C. (2020). *Rogue money and the underground economy. an encyclopedia of alternative and cryptocurrencies*. ABC-CLIO.

European Commission (2021). *Anti-money laundering and countering the financing of terrorism legislative package.* Retrieved from https://ec.europa.eu/. Accessed Nov 2022

Europol (2020). *Internet Organised Crime Threat Assessment 2020.* Retrieved from https://www.europol.europa.eu/. Accessed Nov 2022

FATF (2020). *Virtual assets red flag indicators of money laundering and terrorist financing.* Retrieved from http://www.fatf-gafi.org/. Accessed Nov 2022

FATF (2021). *Second 12-month review of the revised fatf standards on virtual assets and virtual asset service providers.* Retrieved from https://www.fatf-gafi.org/. Accessed Nov 2022

FATF (2022). *International standards on combating money laundering and the financing of terrorism & proliferation: The FATF recommendations.* Retrieved from https://www.fatf-gafi.org/. Accessed Nov 2022

Filippi, P. D., & Wright, A. (2018). *Blockchain and the law: The rule of code*. Harvard University Press.

Fleder, M., Kester, M.S., & Pillai, S. (2015). Bitcoin transaction graph analysis. *Arxiv*. https://arxiv.org/abs/1502.01657. Accessed Nov 2022

Goforth, C.R. (2020). Crypto assets: A Fintech forecast. (September), 5–25.

Harrigan, M., & Fretter, C. (2016). The unreasonable effectiveness of address clustering. *2016 IEEE conferences on ubiquitous intelligence & computing, advanced and trusted computing, scalable computing and communications, cloud and big data computing, internet of people, and smart world congress.* IEEE.

Hilbe, J. M. (2009). *Logistic regression models*. Chapman and hall/CRC.

Ince, P., Liu, J. K., & Zhang, P. (2018). Adding confidential transactions to cryptocurrency IOTA with bulletproofs. *Springer*. https://doi.org/10.1007/978-3-030-02744-53

Kamišalić, A., Kramberger, R., & Fister, I. (2021). Synergy of blockchain technology and data mining techniques for anomaly detection. *Applied Sciences (Switzerland), 11*(17), 7987. https://doi.org/10.3390/app11177987

Katona, T. (2021). Decentralized finance: The possibilities of a blockchain "Money Lego" system. *Financial and Economic Review*, *20*(1), 74–102. https://doi.org/10.33893/fer.20.1.74102.

Kipf, T.N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. https://arxiv.org/abs/1609.02907. Accessed Nov 2022

Koshy, P., Koshy, D., & McDaniel, P. (2014). An analysis of anonymity in Bitcoin using P2P network traffic. *International financial cryptography association, 8437*, 469–485. https://doi.org/10.1007/978-3-662-45472-530

Kute, D.V., Pradhan, B., Shukla, N., & Alamri, A. (2021). Deep learning and explainable artificial intelligence techniques applied for detecting money laundering–a critical review. *IEEE Access*.

Li, X., Liu, S., Li, Z., Han, X., Shi, C., Hooi, B., Huang, H. & Cheng, X. (2020). Flowscope: Spotting money laundering based on graphs. *Proceedings of the AAAI conference on artificial intelligence* 34, 4731–4738. https://doi.org/10.1609/aaai.v34i04.5906

Li, Y., Susilo, W., Yang, G., Yu, Y., Du, X., Liu, D., & Guizani, N. (2019). Toward privacy and regulation in blockchain-based cryptocurrencies. *IEEE Network, 33*(5), 111–117. https://doi.org/10.1109/MNET.2019.1800271

Li, Y., Yang, G., Susilo, W., Yu, Y., Au, M. H., & Liu, D. (2021). Traceable monero: Anonymous cryptocurrency with enhanced accountability. *IEEE Transactions on Dependable and Secure Computing, 18*(2), 679–691. https://doi.org/10.1109/TDSC.2019.2910058

Li, Z., Xiang, Z., Gong, W., & Wang, H. (2022). Unified model for collective and point anomaly detection using stacked temporal convolution networks. *Applied Intelligence, 52*(3), 3118–3131. https://doi.org/10.1007/s10489-021-02559-0

Lischke, M., & Fabian, B. (2016). Analyzing the Bitcoin network: The First Four Years. *Future Internet*, *8*(1). https://doi.org/10.3390/fi8010007.

Liu, L., Liu, J., & Han, J. (2021). Multi-head or single-head? an empirical comparison for transformer training. *Arxiv*. https://arxiv.org/abs/2106.09650.

Lorenz, J.S. (2021). *Machine learning methods to detect money laundering in the Bitcoin blockchain in the presence of label scarcity* (Unpublished doctoral dissertation).

Luo, X. (2014). Suspicious transaction detection for anti-money laundering. *International Journal of Security and Its Applications, 8*(2), 157–166. https://doi.org/10.1016/j.techfore.2020.120166

Meiklejohn, S., Pomarole, M., Jordan, G., Levchenko, K., McCoy, D., Voelker, G. M., & Savage, S. (2016). A fistful of Bitcoins: Characterizing payments among men with no names. *Communications of the ACM, 59*(4), 86–93. https://doi.org/10.1145/2896384

Moreno-Sanchez, P., Zafar, M., & Kate, A. (2016). Listening to whispers of ripple: Linking wallets and deanonymizing transactions in the ripple network. *Proceedings on Privacy Enhancing Technologies, 2016*, 436–453. https://doi.org/10.1515/popets-2016-0049

Nakamoto, S. (2008). *Bitcoin: A peer-to-peer electronic cash system*. www.bitcoin.org/bitcoin.pdf. Accessed Nov 2020

Neudecker, T., & Hartenstein, H. (2017). Could network information facilitate address clustering in Bitcoin? *LNCS, 10323*, 155–169. https://doi.org/10.1007/978-3-319-70278-09

Oad, A., Razaque, A., Tolemyssov, A., Alotaibi, M., Alotaibi, B., & Zhao, C. (2021). Blockchain-enabled transaction scanning method for money laundering detection. *Electronics, 10*(15), 1766. https://doi.org/10.3390/electronics10151766

Ober, M., Katzenbeisser, S., & Hamacher, K. (2013). Structure and anonymity of the Bitcoin transaction graph. *Future Internet, 5*(2), 237–250. https://doi.org/10.3390/fi5020237

Oliveira, C., Torres, J., Silva, M.I., Aparício, D., Ascensão, J.T., & Bizarro, P. (2021). Guiltywalker: Distance to illicit nodes in the Bitcoin network. *Arxiv*. https://arxiv.org/abs/2102.05373. Accessed Nov 2022

Pfitzmann, A., & Hansen, M. (2010). *A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management.* Technical University Dresden, 1–98. 10.1.1.154.635

Phan, T. (2021). *Exploring Blockchain Forensics.*

Philipp, G., Song, D., & Carbonell, J.G. (2017). The exploding gradient problem demystified - Definition, prevalence, impact, origin, tradeoffs, and solutions. *Arxiv*. https://arxiv.org/abs/1712.05577.

Pocher, N. & Zichichi, M. (2022) Towards CBDC-based machine-to-machine payments in consumer IoT. Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing (SAC '22).

Quiniou, M. (2019). *Blockchain: The advent of disintermediation*. ISTE Ltd.

Reid, F., & Harrigan, M. (2013). An analysis of anonymity in the Bitcoin system. In: Altshuler, Y., Elovici, Y., Cremers, A., Aharony, N., Pentland, A. (eds) *Security and Privacy in Social Networks*, 197–223. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-4139-7

Shayegan, M. J., Sabor, H. R., Uddin, M., & Chen, C.-L. (2022). A collective anomaly detection technique to detect crypto wallet frauds on Bitcoin network. *Symmetry, 14*(2), 328. https://doi.org/10.3390/sym14020328

Sun, X., Zhang, J., Zhao, Q., Liu, S., Chen, J., Zhuang, R., Shen, H., & Cheng, X. (2021). Cubeflow: Money laundering detection with coupled tensors. *Pacific-Asia conference on knowledge discovery and data mining.*

Tapscott, D., & Euchner, J. (2019). Blockchain and the internet of value: An interview with Don Tapscott. *Research Technology Management, 62*(1), 12–19. https://doi.org/10.1080/08956308.2019.1541711

Tennant, L. (2017). *Improving the anonymity of the IOTA cryptocurrency*, 1–20. Retrieved from https://laurencetennant.com/. Accessed Nov 2022

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2017). Graph attention networks. *Arxiv*. https://arxiv.org/abs/1710.10903. Accessed Nov 2022

Wang, F., & De Filippi, P. (2020). Self-sovereign identity in a globalized world: Credentials-based identity systems as a driver for economic inclusion. *Frontiers in Blockchain, 2*(January), 1–22. https://doi.org/10.3389/fbloc.2019.00028

Weber, M., Chen, J., Suzumura, T., Pareja, A., Ma, T., Kanezashi, H., Kaler, T., Leiserson. C. E., & Schardl, T. B. (2018). Scalable graph learning for anti-money laundering: A first look. (1970). *Arxiv*. https://arxiv.org/abs/1812.00076. Accessed Nov 2022

Weber, M., Domeniconi, G., Chen, J., Weidele, D.K.I., Bellei, C., Robinson, T., & Leiserson, C.E. (2019). Anti-money laundering in Bitcoin: Experimenting with graph convolutional networks for financial forensics. *Arxiv*(10). https://arxiv.org/abs/1908.02591. Accessed Nov 2022

Wu, J., Liu, J., Chen, W., Huang, H., Zheng, Z., & Zhang, Y. (2020). Detecting mixing services via mining Bitcoin transaction network with hybrid motifs. *Arxiv*. https://arxiv.org/abs/2001.05233. Accessed Nov 2022

Wu, Y., Tao, F., Liu, L., Gu, J., Panneerselvam, J., Zhu, R., & Shahzad, M. N. (2021). A Bitcoin transaction network analytic method for future blockchain forensic investigation. *IEEE Transactions on Network Science and Engineering, 8*(2), 1230–1241. https://doi.org/10.1109/TNSE.2020.2970113

Xu, J. J. (2016). Are blockchains immune to all malicious attacks? *Financial Innovation, 2*(1), 25. https://doi.org/10.1186/s40854-016-0046-5

Yin, H. H. S., Langenheldt, K., Harlev, M., Mukkamala, R. R., & Vatrapu, R. (2019). Regulating cryptocurrencies: A supervised machine learning approach to de-anonymizing the Bitcoin blockchain. *Journal of Management Information Systems, 36*(1), 37–73. https://doi.org/10.1080/07421222.2018.1550550

You, J., Ying, R., & Leskovec, J. (2020). Design space for graph neural networks. *Arxiv*. https://arxiv.org/abs/2011.08843. Accessed Nov 2022