Detection, Ranking and Visualisation of Money Laundering Networks on the Bitcoin Blockchain

A minor thesis submitted in partial fulfilment of the requirements for the degree of Masters of Data Science (Computing Technology)

Jennifer Payne
School of Computer Science and Information Technology
Science, Engineering, and Technology Portfolio,
Royal Melbourne Institute of Technology
Melbourne, Victoria, Australia

October 18, 2025

Declaration

This thesis contains work that has not been submitted previously, in whole or in part, for any other academic award and is solely my original research, except where acknowledged.

This work has been carried out part-time since March 2024, under the supervision of Dr. Son Hoang Dau.

Jennifer Payne
School of Computer Science and Information Technology
Royal Melbourne Institute of Technology
October 18, 2025

Acknowledgements

I extend my thanks to my supervisor, Dr. Son Hoang Dau, for his guidance, mentoring and patience during my studies, supporting a fascinating research topic that is of both professional and personal interest. Above all, I thank my husband, Matúš, who has been my pillar of support and a source of encouragement during my studies. A lot of life happened during this thesis: full-time work, house renovations, the death of a loved one, pregnancy and postpartum, and later, caring for our newborn baby, Maxwell. Completing this thesis was truly a team effort.

Contents

List of Figures

List of Tables

Summary

Cryptocurrency such as Bitcoin can be used to conceal the origins of criminal money due to its speed, global reach, and pseudonymous design. Banks and financial institutions use rule-based systems to detect suspicious activity, but these often create too many false alerts and require large amounts of manual work. Many academic studies have tried to improve this process using artificial intelligence, but most focus only on technical performance rather than on how investigators can use these tools in practice. There is still a major gap between advanced research models and the simple, transparent systems needed by regulators and financial crime teams.

This thesis develops a clear and practical approach that helps investigators identify and understand money-laundering activity on the Bitcoin blockchain. The method combines four key steps: labelling transactions as suspicious or not suspicious, identifying networks of illicit transactions within a specific time frame, ranking the transactions from most to least influential in the network, and visualising and summarising the networks in an easy-to-read format. The entire process runs on cloud infrastructure, making it scalable for large datasets.

The outcome is a usable and transparent system that can help financial crime teams and regulators trace illicit funds more effectively. It balances advanced data science with practical usability—offering a bridge between complex academic models and the real-world tools investigators need to detect and prevent money laundering in digital currencies.

Abstract

Cryptocurrency has become an attractive medium for money laundering due to its pseudo-anonymous design, global accessibility, and rapid transaction speed. Traditional anti-money-laundering (AML) systems in financial institutions rely on rule-based methods that generate high false-positive rates, demand extensive manual review, and require frequent maintenance. Academic research on blockchain-based AML has largely focused on model accuracy in isolation, producing complex, black-box solutions that lack transparency, interpretability, or operational scalability. However, there remains a critical gap between high-performing academic models and investigator-ready systems that can be realistically deployed within regulatory and compliance environments. Such systems are often unsuitable for investigative or regulatory use, where explainability, auditability, and timeliness are critical.

This thesis presents a practical, investigator-centred methodology for detecting, ranking, and visualising money-laundering activity on the Bitcoin blockchain. The framework prioritises usability, interpretability, and computational efficiency through four integrated components:

1. A supervised classification model that identifies suspicious transactions,
2. A graph-based algorithm that constructs illicit-only subnetworks from those transactions,
3. A composite ranking method that prioritises key entities within each subnetwork, and
4. A visualisation module that presents intuitive, dual-perspective (transaction- and address-level) network views to support investigative decision-making.

A supporting cloud-based system architecture enables deployment at scale and integration within existing AML environments. By bridging traditional financial-sector compliance frameworks with blockchain analytics, this research delivers a transparent and scalable end-to-end pipeline designed for investigative teams and regulators. The result is an explainable, auditable, and operationally viable approach that enhances the detection, triage, and understanding of illicit financial flows in decentralised systems.

Code and Data Availability:
All scripts, trained models, and data processing notebooks developed for this research are available at: https://github.com/majorpayne-2021/rmit_master_thesis. The Elliptic++ dataset is available at https://github.com/git-disl/EllipticPlusPlus.

Chapter 1

Introduction

Cryptocurrency has become a major conduit for laundering illicit funds due to its pseudo-anonymous design, borderless nature, and the speed of value transfer on public blockchains. The money-laundering process typically consists of three stages—placement, layering, and integration—where illicit proceeds are first introduced into the system, moved through obfuscating transfers, and finally reintroduced into the legitimate economy [11]. The transparency of blockchain ledgers paradoxically coexists with obfuscation strategies such as address clustering, mixing, and cross-chain bridging, which conceal the source of criminal funds [47, 34, 11]. Between 2019 and 2024, almost USD 100 billion in illicit crypto assets were transferred through conversion services, demonstrating the scale of laundering across predicate crimes including ransomware, fraud, and darknet trade [11].

The regulatory landscape for cryptocurrency has strengthened significantly in recent years. Financial intelligence units and international standard-setters such as the Financial Action Task Force (FATF), AUSTRAC, FinCEN, and the European Union now explicitly require Virtual Asset Service Providers (VASPs) to comply with AML/CTF obligations comparable to those imposed on banks. These include customer due diligence, transaction monitoring, suspicious activity reporting, and implementation of the FATF "travel rule" mandating the exchange of sender and receiver information for crypto transfers [6, 11]. As enforcement intensifies, regulators increasingly expect crypto-based AML systems to be traceable, auditable, and explainable, enabling investigators and compliance officers to justify model outputs and reporting decisions.

Traditional anti-money-laundering (AML) programs in the fiat sector remain dominated by rule-based transaction monitoring systems that rely on thresholding, customer profiling, and typologies derived from regulatory guidance [22, 6]. Although such systems detect well-known behavioural patterns, they fail to identify novel or rapidly evolving laundering typologies. Studies highlight that these approaches are static, generate excessive false positives, and lack contextual network awareness, limiting their usefulness in large-scale investigations [22, 39, 40]. Consequently, the burden on compliance teams remains substantial, with the majority of alerts requiring manual triage and investigation.

Academic research has increasingly explored artificial intelligence (AI) and machine-learning-based AML solutions, ranging from logistic regression and random forests to deep graph neural networks. However, as Han et al. [22] and Deprez et al. [14] observe, much of this research remains at the conceptual or prototype stage. Early deep-learning work such as Weber et al. [46] and the subsequent Elliptic2 dataset by Bellei et al. [8] demonstrated the promise of graph-based methods but also revealed key shortcomings: limited explainability, high computational demands, and limited operational alignment with investigator workflows. Newer studies—Ouyang et al. [34] on subgraph contrastive learning and Elmougy and Liu [16] on interpretable graph analytics—continue to advance technical performance yet remain detached from practical investigative contexts. Most models emphasise accuracy over interpretability, making them unsuitable for compliance environments that require transparency, auditability, and regulatory accountability.

Beyond methodological opacity, another critical gap lies in the lack of investigator-centred design. Most AML research treats transaction detection as a model-centric classification problem, overlooking the need for usable investigative outputs such as prioritised rankings or interpretable network maps. In practice, investigative teams operate under significant resource constraints—limited time, staffing, and capacity to review the high volume of alerts generated by AML systems. Methodologies must therefore not only be transparent and explainable but also help investigators prioritise their efforts, focusing on the transactions, entities, or networks that are most likely to represent serious laundering activity. Traditional rule systems flag individual transactions, while network science—despite its rich history in criminology [42, 18, 30]—has been integrated into operational blockchain investigative processes on a limited basis. There is therefore a clear opportunity to connect the advances of graph-based learning with the practical realities of financial-crime analytics: scalable, explainable, and regulator-auditable systems capable of guiding investigators through vast networks of illicit flows in a way that maximises the impact of limited investigative resources.

This study addresses these gaps by proposing a practical, investigator-ready AML framework tailored for cryptocurrency analysis. The methodology bridges academic research and compliance operations by combining machine-learning classification with graph-based network analysis, ranking, and visual analytics. Investigators are placed at the centre of the design process, ensuring that outputs are intuitive, explainable, and directly aligned with real investigative workflows.

The proposed framework consists of four interconnected components:

1. A classification model to label Bitcoin transactions as suspicious or non-suspicious;
2. A graph algorithm to extract illicit-only subnetworks anchored on highly probable suspicious transactions;
3. A ranking method that integrates structural and financial indicators (e.g., PageRank, value flows, degree metrics) to prioritise key actors; and
4. A visualisation module that presents the resulting networks in an intuitive format, supporting investigative interpretation.

Together, these elements form an end-to-end, explainable system architecture designed for operational scalability. By prioritising transparency, interpretability, and regulator auditability, the framework advances the state of cryptocurrency AML research beyond model-centric paradigms toward applied, human-centred solutions.

In summary, this work contributes to closing three persistent gaps in the literature:

1. The lack of transparent, explainable methods that regulators can audit;
2. The absence of investigator-aligned network-based tools for blockchain AML; and
3. The disconnection between academic innovation and operational implementation in real-time or large-scale investigative contexts.

1.1 Research Questions

This research is guided by the following questions:

1. What classification approaches are most effective for detecting suspicious Bitcoin transactions in an imbalanced dataset while maintaining transparency and auditability?
2. How can illicit-only subnetworks be efficiently constructed from suspicious transactions to mirror investigative workflows and reduce computational overhead?
3. How can a ranking method that integrates structural and financial indicators (e.g., PageRank, value flows, degree measures) be designed to prioritise key actors within illicit subnetworks?
4. How can network visualisations be developed to maximise interpretability and minimise the misidentification of service nodes such as exchanges or mixers?

1.2 Research Contributions

1. Investigator- and regulator-centred methodology — A transparent, explainable framework that aligns model outputs with investigative reasoning, bridging the divide between technical performance and regulatory interpretability.
2. End-to-end AML framework — A fully integrated pipeline spanning classification, subnetwork construction, ranking, and visualisation. Unlike prior work [46, 16, 34], it delivers a cohesive, auditable system rather than isolated experimental models.
3. Enhanced classification performance — A Random-Forest-based classifier achieving higher recall on the Elliptic and Elliptic++ datasets than prior benchmarks [46, 8], balancing interpretability and predictive strength.

4. Efficient illicit-only subnetwork generation — A graph construction method aligned with real-world alert triage, reducing computational load while enhancing investigative usability.

5. Ranking algorithm — A hybrid metric combining PageRank centrality, inbound BTC value, and degree scores to identify key laundering facilitators, extending prior PageRank-based studies [52, 18].

6. Network visualisation — Graph representations that expose laundering typologies while preventing misclassification of exchange or mixer nodes, supporting both analysts and regulators in traceability reporting.

Chapter 2

Related Works

The detection of money laundering has evolved from manual compliance reviews and rule-based systems in traditional banking to data-driven, machine-learning-based methods in cryptocurrency analysis. Across both domains, the persistent challenge lies in distinguishing illicit activity within vast and dynamic transaction networks. This chapter examines the evolution of Anti-Money Laundering (AML) practices—from the regulated financial sector to the emerging cryptocurrency ecosystem—and reviews current machine learning and network analysis approaches. It highlights the growing regulatory emphasis on explainability and traceability, as well as key gaps in scalability, interpretability, and operational integration that motivate the investigator-centred, explainable framework developed in this research.

2.1 AML in the Financial Sector

Money laundering is the process of concealing the origin of illegally obtained funds through legitimate financial channels. It typically unfolds in three stages: placement, layering, and integration [4, 22]. During placement, illicit funds enter the financial system; layering obscures their origin through complex transactions; and integration reintroduces them into the economy as legitimate assets.

Financial institutions play a central role in global AML enforcement, operating within strict regulatory frameworks established by the Financial Action Task Force (FATF), AUSTRAC, and FinCEN. In Australia, the Anti-Money Laundering and Counter-Terrorism Financing Act 2006 (AML/CTF Act) mandates that reporting entities implement Know Your Customer (KYC) processes, ongoing Customer Due Diligence (CDD), and transaction monitoring systems to detect suspicious activity [2]. These programs are designed to ensure that all monitoring outcomes are explainable, traceable, and defensible under audit. Regulatory policy therefore prioritises transparency and human oversight over automation, a stance that has shaped how AML technologies are deployed in practice [22, 26].

Most institutions rely on rule-based transaction monitoring systems, where alerts are generated when thresholds are breached or when customer behaviour matches predefined typologies [6]. This design satisfies legal and compliance obligations by producing interpretable audit

trails, but it also introduces operational inefficiencies. Rule-based systems struggle to detect emerging typologies and generate high false positive volumes that require manual review. Studies by Han et al. [22] and Shaikh et al. [40] highlight that compliance teams spend significant time investigating legitimate alerts, diverting resources from high-risk cases.

The investigative process following an alert typically involves three stages: alert review, case building, and reporting to authorities [26]. This workflow, illustrated in Figure 2.1, begins with automated flagging of transactions that meet typology criteria, progresses to analyst validation of suspicious cases, and concludes with formal reporting to regulatory agencies. While this layered process ensures accountability, it also slows response times and increases analyst workload.

Rule-based systems therefore achieve regulatory compliance and explainability but fall short in analytical adaptability. They excel in documenting decision processes yet fail to capture the relational and behavioural dimensions of laundering that span multiple actors and accounts. As financial crime becomes increasingly networked and cross-border, there is a growing need for AML solutions that preserve auditability while introducing data-driven, pattern-oriented detection capable of linking connected activities across entities.

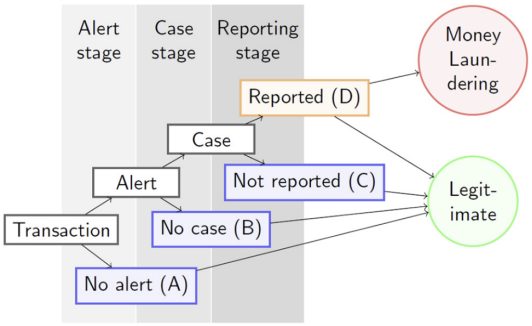


Figure 2.1: Banking investigation process following alert [26]

2.2 AML in the Cryptocurrency Ecosystem

The emergence of cryptocurrency has reshaped the AML landscape, introducing decentralisation, global accessibility, and pseudonymity as both technological innovations and compliance challenges. Unlike traditional financial systems, blockchain networks enable rapid cross-border transfers without intermediaries, reducing regulatory visibility. This openness has made cryptocurrency an attractive medium for laundering proceeds from illicit activities such as fraud, ransomware, darknet trade, and on-chain crimes including hacks, scams, and illegal marketplace transactions [11, 22].

Money laundering in the crypto ecosystem mirrors the classical stages of placement, layering, and integration [4]. During placement, criminal proceeds—whether derived from fiat-based offences or from on-chain activities such as scams, ransomware, or hacks—are introduced into circulation, often through exchanges, peer-to-peer platforms, or over-the-counter brokers [11, 19]. Layering obscures the source of these funds using techniques such as mixing services, tumblers, privacy-focused coins like Monero and Zcash, and cross-chain transfers through decentralised exchanges or bridges. Integration reintroduces the cleaned assets into the legitimate economy, either by conversion to fiat on compliant exchanges or through direct investment into lawful ventures [34, 11]. The layering stage has become particularly

sophisticated, with newer typologies such as chain-hopping, cross-chain swaps, and the use of decentralised finance (DeFi) protocols to conceal illicit flows [11].

Regulators have responded to these evolving laundering typologies by extending AML frameworks to include cryptocurrency-related entities. FATF's "travel rule" requires Virtual Asset Service Providers (VASPs) to collect and share sender and receiver information for crypto transactions, mirroring obligations imposed on banks. National authorities such as FinCEN (USA), AUSTRAC (Australia), and the FCA (UK) have aligned their local regimes accordingly. In Australia, amendments to the AML/CTF Act 2006 extended the law to digital currency exchange providers, requiring registration with AUSTRAC, implementation of KYC, transaction monitoring, and suspicious matter reporting [2]. These amendments formally integrate cryptocurrency businesses into the national AML regime, and recent government proposals seek to apply Australian Financial Services Licence (AFSL) rules to crypto providers [43]. AUSTRAC has also targeted unregulated financial channels such as cryptocurrency ATMs, introducing compliance actions to prevent their use for money laundering [3].

Recent enforcement actions illustrate that regulators are increasingly holding crypto platforms to traditional compliance standards. In the United States, Binance and its CEO pleaded guilty to federal charges including violations of the Bank Secrecy Act and failure to maintain effective AML programs, agreeing to pay over USD 4.3 billion in penalties and implement an independent compliance monitor [44]. In Australia, the Federal Court ruled that Bit Trade Pty Ltd, the operator of Kraken's Australian platform, breached design and distribution obligations, leading to an AUD 8 million penalty and enforcement action by ASIC [7]. These cases underscore the growing regulatory expectation that digital asset exchanges meet the same transparency, reporting, and risk management standards applied to traditional financial institutions.

Academic research increasingly reflects this convergence between technological and regulatory domains. Han et al. [22] emphasise that AI-based AML systems must remain interpretable to meet regulator expectations, while Gruber [21] highlights that cryptocurrency exchanges now function as de facto banks, requiring equivalent standards of disclosure and oversight. Deprez et al. [14] reinforce the importance of regulatory explainability in cryptocurrency AML, noting that while network analytics and deep graph models enhance predictive performance, their lack of interpretability limits real-world deployment. They argue that interpretability and explainability remain the limiting factors for AML adoption in compliance and investigative environments, calling for frameworks that connect analytical advances with regulatory compliance and investigator usability. Chainalysis [12] similarly documents rising enforcement actions against opaque exchanges and mixers, reflecting a global shift toward stricter oversight.

Despite these developments, many AML models in crypto remain technically advanced but practically constrained. Deep learning and graph methods such as those by Weber et al. [46] and Ouyang et al. [34] deliver high predictive accuracy but lack transparency, limiting their acceptance in regulated environments that demand auditable decision-making. Deprez et al. [14] identify this gap as central: performance alone is insufficient without interpretability and traceability.

The ongoing challenge in crypto AML thus lies not in data access but in converting blockchain transparency into regulatory-grade traceability. Current systems frequently function as black

boxes, incapable of explaining how illicit behaviour is detected or prioritised. Bridging this gap demands models that offer analytical rigour while providing outputs that are interpretable, auditable, and usable by investigators under realistic resource constraints.

2.3 Machine Learning and Network-Based Approaches to AML

Machine learning (ML) has become an essential extension of rule-based Anti-Money Laundering (AML) systems, offering improved adaptability and detection capability for complex or evolving laundering behaviours. Over time, academic research has moved from traditional supervised classification models toward network-based and semi-supervised approaches that leverage the relational structure of financial data. Despite these advances, a persistent challenge remains: translating these technical models into transparent, auditable, and operational tools that reflect the investigative reality of AML enforcement.

2.3.1 Supervised and Semi-Supervised Learning for AML Detection

Early research on ML for financial crime detection focused on supervised classification models trained on labelled datasets. Common algorithms included logistic regression, decision trees, random forest, gradient boosting (e.g., XGBoost), and support vector machines [26, 31, 46, 16, 17, 23]. Random forest models, in particular, have remained popular because of their balance between interpretability and predictive power—qualities critical for regulated environments where decisions must be explainable and auditable. Across multiple studies, including Weber et al. [46], Elmougy and Liu [16], Farrugia, Ellul, and Azzopardi [17], Jullum et al. [26], and Silva, Correia, and Maziero [41], random forest consistently outperformed deep or complex neural networks when evaluated on imbalanced or limited AML datasets.

However, the availability of publicly accessible and reliable labelled datasets remains a major limitation in AML research. Illicit activity represents only a small fraction of total transactions, and ground truth labels are difficult to verify in practice. To address this, semi-supervised and hybrid learning methods have emerged. Karim et al. [27] proposed scalable semi-supervised graph learning techniques that reduce dependence on labels while improving generalisation to unseen nodes. Silva, Correia, and Maziero [41] demonstrated that combining node- and edge-level GNNs with XGBoost enhanced performance and stability under severe class imbalance. Samadi, Dong, and Xia [37] introduced a multi-pattern framework that identifies several transaction typologies through off-chain propagation and personalised PageRank scoring. While such typology-based models expand the range of detectable behaviours, their reliance on predefined patterns limits adaptability—criminals quickly modify transaction behaviours to evade recognition. This reinforces the need for explainable and flexible models that capture the structural and value-based dynamics of money laundering rather than static behavioural templates.

Despite these advances, supervised, semi-supervised, and unsupervised models remain largely model-centric. They depend on labelled data that are sparse, biased, or unreliable, limiting reproducibility and validation. Many are optimised for performance metrics rather than usability, interpretability, or auditability, and lack integration into real investigative workflows.

Moreover, unsupervised models cannot be easily verified due to the absence of ground truth labels, which undermines confidence in their operational application. Collectively, these challenges illustrate that while classification accuracy has improved, the bridge between technical performance and investigative practicality remains underdeveloped.

2.3.2 Network Analytics, Visualisation, and Investigative Workflows in AML Detection

Network analysis extends AML detection beyond individual transactions and wallets by examining the relational structure between entities. It enables the identification of hidden dependencies, intermediary actors, and coordinated behaviours across multiple accounts. Within AML research, network analysis represents a diverse methodological domain encompassing several analytical objectives, from feature engineering to subgraph classification and investigative visualisation.

Feature Extraction for Machine Learning

Many studies use graph-derived statistics—such as node degree, clustering coefficient, betweenness, or transaction frequency—as additional features for supervised models. Examples include Hu et al. [23], Farrugia, Ellul, and Azzopardi [17], Weber et al. [46], and Lorenz et al. [31]. In these works, the network primarily functions as a feature generator that improves model accuracy rather than serving as the central analytical framework.

Graph-Based Classification and Subgraph Learning

More recent approaches treat the network itself as the object of analysis. Graph convolutional and attention networks (GCNs, GATs) directly classify nodes or subgraphs by leveraging relational structure [46, 25, 51, 8, 41]. Subgraph-based models such as FlowScope [28] and TRacer [50] aim to improve scalability by limiting computation to localised regions. These models perform well in retrospective analyses but typically assume that a complete, labelled transaction network is available—a premise also seen in Weber et al. [46], Elmougy and Liu [16], Bellei et al. [8], Xiang et al. [51], and Hyun, Lee, and Suh [25].

Network Exploration and Investigative Visualisation

Tools by Phetsouvanh, Oggier, and Datta [36] and Wu et al. [49] focus on tracing fund flows and assessing privacy risks, while Fronzetti Colladon and Remondi [20] demonstrate the practical value of social network analysis in identifying suspicious clusters. These systems assist human interpretation but are designed primarily for exploratory analysis rather than active investigations.

While these methods provide valuable insights into transactional and relational structures, most operate under the assumption that the full transaction graph is known [46, 16, 51, 8, 25]. In operational settings, this assumption is rarely valid. Investigators typically begin with

a single alert, address, or transaction and expand their analysis iteratively, verifying each connection as evidence emerges. Waiting for sufficient time to accumulate a comprehensive background graph allows laundering activity to persist and grow before intervention. Consequently, models developed on static or complete graphs may achieve strong retrospective accuracy but fail to reflect the dynamic, time-sensitive, and incomplete conditions of real AML investigations.

Overall, network-based AML research remains limited by its reliance on complete graph visibility and its retrospective analytical focus. Few approaches consider the temporal and operational constraints of real investigative workflows. There is therefore a clear need for dynamic, forward-tracing network models that can operate on partial data while supporting timely, interpretable insights for investigators.

2.3.3 Ranking and Prioritisation in AML Networks

Ranking algorithms provide a quantitative mechanism for prioritising entities within large transaction networks. Traditional measures such as degree, betweenness, and eigenvector centrality identify structurally important nodes, while iterative algorithms like PageRank and HITS evaluate influence through network connectivity [15, 13]. In AML contexts, early propagation-based approaches by Möser, Böhme, and Breuker [32] applied graph proximity to known illicit addresses to estimate transaction-level risk. More recent methods, including RiskProp [29], extended this to account-level risk scoring on Ethereum using semi-supervised graph propagation.

Privacy-preserving implementations of PageRank [38] illustrate how multiple financial institutions or jurisdictions can collaborate in analysing shared transaction networks without disclosing sensitive information. Through secure multiparty computation, these systems jointly calculate node influence scores while maintaining data confidentiality—a promising capability for regulatory environments that demand both collaboration and privacy compliance.

Despite their analytical strengths, most ranking models remain focused on mathematical influence rather than investigative prioritisation. They often quantify network importance without incorporating transaction value, temporal context, or interpretability. Few frameworks translate influence scores into auditable outputs that investigators can directly act upon. Addressing these gaps requires composite ranking methods that integrate financial flows, structural centrality, and time-based indicators into a transparent scoring system that aligns with investigative priorities.

2.3.4 Summary of Research Gaps

Across the literature, several consistent gaps persist. Most studies remain model-centred, optimising for performance metrics rather than practical integration into investigative workflows. The scarcity of reliable and publicly available labelled data limits validation and reproducibility, while unsupervised models suffer from unverifiable outputs due to the absence of ground truth. Network-based methods typically assume access to complete transaction graphs, ignoring the partial and time-critical nature of real-world investigations; such delays in visibility allow criminal networks to expand and conceal further illicit activity. Moreover, many deep

graph models lack scalability and interpretability at blockchain scale. Finally, the absence of prioritisation and end-to-end integration across classification, network construction, and ranking limits operational applicability.

These gaps underscore the need for AML frameworks that combine technical performance with transparency, auditability, and investigative usability. The following methodology addresses these limitations through an integrated, investigator-centred approach that supports classification, subnetwork construction, ranking, and visualisation in a practical and explainable form.

Chapter 3

Methodology

The methodology presented in this thesis proposes a dynamic and practical framework for detecting money laundering activity on the Bitcoin blockchain. The objective is not to identify the underlying predicate offence, but to establish a repeatable, transparent, and operationally viable process to "follow the funds" and highlight suspicious entities for further investigation. Within an organised crime structure, actors play different roles ranging from low-level money mules to senior coordinators. While mules are often interchangeable and expendable, the greater investigative value lies in tracing transaction flows upwards to reach the more senior, strategic members of a criminal organisation, consistent with the hierarchical model of organised crime described by the United Nations Office on Drugs and Crime[45].

This framework is designed to operate within the constraints faced by real-world anti-money laundering (AML) functions, such as those within cryptocurrency exchanges, financial institutions, or regulatory agencies. It focuses on producing outputs that are explainable, auditable, and interpretable—characteristics that are essential for investigative decision-making and regulatory compliance.

The methodology consists of four interconnected components: classification, subgraph construction, ranking, and visualisation. Each component directly addresses a limitation identified in the literature review. The classification model responds to the limitations of rule-based systems and opaque machine learning models by providing a transparent, auditable classifier for labelling transactions as suspicious or not suspicious. The subgraph construction process addresses the unrealistic reliance on full background graphs found in much of the AML research by instead building dynamic subnetworks from known illicit seeds. The ranking stage responds to the lack of prioritisation frameworks by introducing a structured and explainable way to order transactions according to influence and value within a network. Finally, the visualisation component bridges the gap between data science research and investigative usability, enabling investigators to interpret complex relationships through an intuitive, network-based view.

The overall procedure follows seven main steps:

1. **Retrieve transactions for a specified time window:** Obtain transaction records, their associated features, and the transaction-to-transaction edges that represent Bitcoin

flows between sender and receiver addresses.

2. **Classify transactions:** Use a Random Forest classifier to label each transaction as suspected illicit or licit.
3. **Build subnetworks from illicit seeds:** Starting from each illicit transaction, construct a forward-directed subnetwork using breadth-first search (BFS), following only illicit nodes and stopping expansion at the first licit node. Where illicit nodes share the same transaction path, overlapping subnetworks are produced.
4. **Deduplicate identical subnetworks:** Merge subnetworks that are 100% identical in terms of nodes and edges.
5. **Merge overlapping subnetworks:** Further merge subnetworks that share at least one illicit node, consolidating multiple seeds into a single network where illicit paths converge.
6. **Rank transactions within each subnetwork:** Assign a composite ranking score to each transaction based on structural importance within the network and the relative amount of Bitcoin value consolidated.
7. **Visualise and summarise:** Provide network visualisations and summary tables to investigative teams, including node rankings and key transaction attributes to guide investigative prioritisation.

This design directly addresses the key limitations of existing AML research, which often relies on retrospective, computationally intensive models built on full-graph visibility. In practice, investigators rarely have complete transaction data; rather, they begin with a single alert or entity and iteratively trace related transactions outward. The proposed approach mirrors this investigative process by starting from a single suspicious transaction (seed) and expanding forward through its connected illicit flows. This not only reflects how AML teams operate, but also improves computational efficiency by limiting expansion to relevant areas of the network.

A visual overview of this methodology is provided in Figure A.1 in Appendix A. The framework is designed around three guiding principles: scalability, explainability, and interpretability. Scalability ensures that each stage—from classification to ranking—can handle millions of transactions without sacrificing performance. Explainability ensures that every decision, from classification output to ranking score, can be justified and audited by investigators and regulators. Interpretability ensures that the final outputs—network maps and ranked tables—are immediately usable by non-technical investigators operating under time and resource constraints.

A further distinguishing feature of this framework is its attention to operational constraints. Transaction monitoring in both fiat and cryptocurrency AML functions typically occurs in real time or near-real-time batches. Waiting to accumulate complete background graphs may improve analytical precision, but it also permits criminal activity to continue unchecked. Delayed detection not only reduces the opportunity for intervention but allows illicit actors to consolidate and obscure funds, undermining the traceability that blockchain transparency can offer. The proposed framework therefore prioritises responsiveness, allowing investigators to act quickly with the information available at the time of alert.

Rather than maximising algorithmic complexity, this methodology prioritises transparency, reproducibility, and investigative utility. It provides a framework that cryptocurrency AML teams can implement, investigators can trust, and regulators can audit—balancing innovation with operational practicality.

Each component of this methodology was designed to directly address the research questions posed in this study. The supervised classification stage corresponds to Research Question 1, which investigates how illicit and licit Bitcoin transactions can be distinguished using transparent and auditable machine learning techniques. The subgraph construction process addresses Research Question 2 by exploring how illicit transaction networks can be built and represented in a way that reflects real investigative workflows, starting from known suspicious transactions rather than relying on complete network visibility. The ranking algorithm responds to Research Question 3 by developing a structured method for prioritising key actors within these subnetworks based on their structural importance and financial influence. Finally, the visualisation component contributes to Research Question 4 by translating analytical results into investigator-friendly representations that enhance interpretability and support decision-making. Together, these components form a cohesive framework that links the analytical power of machine learning with the operational needs of financial crime investigation, bridging the gap between academic research and practical AML enforcement.

3.1 Data

3.1.1 Background on Bitcoin Transactions

Bitcoin transactions differ fundamentally from traditional fiat transactions. In conventional off-chain payment systems, such as bank transfers or credit card payments, transactions are generally one-to-one: each transfer involves a single sender and a single receiver, and is recorded in a centralised ledger maintained by a trusted intermediary. Bitcoin, by contrast, operates on the Unspent Transaction Output (UTXO) model [33], which allows for many-to-many transactions. A single Bitcoin transaction may draw funds from multiple input addresses—spending several previous outputs simultaneously—and create multiple outputs that send value to different addresses.

This many-to-many design enables greater flexibility but also introduces significant complexity for financial crime detection. For example, when the total value of a transaction's inputs exceeds the value being sent to the intended recipient, the difference (or "change") is automatically returned to a new output address controlled by the sender. On the blockchain, however, these change outputs are indistinguishable from ordinary payments to third parties. As a result, tracing illicit flows requires careful determination of whether a transaction output represents genuine value transfer or internal fund redirection.

Bitcoin's pseudonymous design further complicates anti-money laundering (AML) detection. While every transaction and address is permanently recorded on the blockchain, addresses themselves are cryptographic identifiers that are not directly linked to real-world identities. Without external data sources such as exchange-based know-your-customer (KYC) records or law enforcement intelligence, ownership attribution remains uncertain. These characteristics contrast sharply with fiat systems, where accounts are typically linked to identifiable entities.

As noted by Wu et al. [48] and Albrecht et al. [1], this pseudonymity allows criminals to exploit blockchain transparency while concealing real-world identities, creating new challenges for regulators and investigators. As a result, Bitcoin transaction analysis requires an approach that goes beyond individual transactions and wallets, instead modelling the broader network of interactions where the flow of value must be inferred from structural and temporal patterns rather than explicit identity information.

3.1.2 The Elliptic and Elliptic++ Datasets

The Elliptic dataset [46] captures these dynamics through a directed transaction graph consisting of 203,769 Bitcoin transactions (nodes) and 234,355 edges representing BTC flows. Each transaction is labelled as licit, illicit, or unknown, though the data is highly imbalanced: only around 2% of transactions (4,545) are illicit, 21% (42,019) are licit, and the remaining 77% (157,205) are unlabelled. Each transaction is represented by 166 features, including 94 local features (such as number of inputs, outputs, transaction size, and fees) and 72 aggregated features derived from neighbouring transactions. To protect the authors' proprietary data and intellectual property, transaction identifiers were masked and all features were normalised. The labelling methodology used by Elliptic is only described at a high level, with limited disclosure of the heuristics behind the classification process.

The Elliptic++ dataset [16] builds upon the Elliptic dataset through two major enhancements. First, it deanonymises 99.5% of the masked transactions by linking them to wallet addresses, expanding the dataset to include over 822,000 wallet nodes while retaining the same 203k labelled transactions. Second, it augments the feature set with 17 new attributes—raising the total to 183—by incorporating raw, unmasked blockchain information such as total inbound and outbound BTC value, number of input and output addresses, and transaction size. Elliptic++ also introduces address-level and user-entity-level datasets produced via clustering algorithms, although this study focuses exclusively on the transaction-level data to maintain consistency with the original Elliptic framework.

Elliptic [46] and Elliptic++ [16] remain the most widely used and reliable labelled Bitcoin transaction datasets in AML research. The latter represents a significant step forward by integrating raw blockchain features and address linkages while maintaining compatibility with the original data structure. It therefore provides the most comprehensive and realistic foundation available for developing AML detection methodologies that combine machine learning and network analytics. Despite certain limitations, Elliptic++ was selected for this research as it remains the largest and most reliable public dataset for studying illicit Bitcoin activity. Its feature-rich structure and deanonymised address information make it particularly well suited for the four-stage methodology proposed in this thesis, supporting both supervised classification and graph-based subnetwork construction.

While the dataset provides an invaluable resource, several constraints influence how downstream models can be designed and interpreted:

- The data is severely imbalanced, with illicit transactions representing only around 2% of all records.

- A large proportion of transactions (77%) are unlabelled, limiting confidence in model training and ground-truth validation.
- Labels exist only at the transaction level, with no verified wallet- or entity-level classifications. The dataset does not specify the underlying predicate offences (e.g., darknet market, ransomware, scam).
- Wallet type information (such as exchanges, merchants, or individuals) is not provided, reducing contextual interpretability.
- Although Elliptic++ deanonymises most transactions, the illicit/licit labels remain derived from Elliptic's proprietary heuristics, which are not publicly disclosed.
- Transaction identifiers are masked and normalised, preventing direct cross-referencing with raw blockchain data.
- BTC edge weights between transactions are not included, restricting the ability to perform value-weighted network analysis or apply weighted ranking algorithms.

To partially validate the dataset, targeted spot checks were performed using open-source blockchain explorers such as BTCscan, exemplified in Appendix E. These checks confirmed the structural consistency of transaction metadata in a sample of deanonymised addresses, though the original licit/illicit classifications could not be independently verified, as these labels are derived from proprietary sources unavailable for public audit. These checks therefore serve as qualitative validation of dataset integrity rather than confirmation of ground-truth labels.

In summary, Elliptic++ provides a robust empirical foundation for this research. It offers sufficient structural and transactional detail to support the classification, subnetwork construction, and ranking stages of the proposed methodology. While the dataset's imbalance, anonymisation, and absence of value-weighted edges impose certain analytical constraints, these are mitigated by the framework's focus on explainable, investigator-centric subnetworks built from high-confidence illicit nodes. The resulting methodology therefore remains both scientifically rigorous and operationally relevant, bridging the gap between academic AML research and real-world investigative practice.

3.2 Classification Model

The classification stage identifies suspicious Bitcoin transactions that form the foundation for subsequent subnetwork construction and ranking. Two supervised learning models were implemented—logistic regression and random forest—to balance interpretability, a key requirement for regulatory explainability, with predictive performance necessary for detecting complex laundering behaviour. Logistic regression is a statistical model that predicts the probability of a binary outcome using a linear combination of input features passed through a logistic (sigmoid) function. Random forest is an ensemble learning method that builds multiple decision trees and combines their outputs to improve classification accuracy and reduce overfitting.

This stage addresses several key gaps in prior AML research. First, many studies rely on datasets where illicit labels are sourced from open-data repositories or community reporting rather than verified law-enforcement intelligence, limiting the reliability of model evaluation [17, 24, 36]. To mitigate this, this research leverages the Elliptic and Elliptic++ datasets [46, 16, 8], developed by a regulated blockchain analytics firm supplying AML intelligence to financial institutions and regulators, providing the most credible and reproducible foundation available.

Second, while many AML models achieve strong technical performance, few prioritise explainability or auditability—both essential for adoption within compliance frameworks. The models in this study are explicitly designed to balance predictive accuracy with interpretability, ensuring that outputs can be understood and justified to internal auditors and regulators.

Logistic regression serves as a transparent baseline widely used in AML research [26, 31], offering simplicity and auditability but limited capacity to capture non-linear relationships. Random forest, by contrast, models complex feature interactions while remaining interpretable through feature-importance measures [46, 16, 17]. Together, these models provide a comparative framework to assess trade-offs between explainability, accuracy, and operational practicality in AML detection.

3.2.1 Feature Engineering

In addition to the features provided by the Elliptic and Elliptic++ datasets, twelve supplementary features were engineered to capture behavioural and structural transaction characteristics relevant to money laundering detection. These features reflect both transactional irregularities (e.g., unusually high or low values, disproportionate fees) and structural properties (e.g., atypical input–output relationships) commonly observed in laundering typologies.

The engineered features include: in–out BTC ratio, input–output address ratio, BTC output dispersion, high-fee flag, micro-transaction flag, transaction density, fees per byte, fees ratio, fees per input, rounded amount flag, input address percentile, and output address percentile.

For the 965 transactions that could not be deanonymised in the Elliptic++ dataset, missing values were imputed using the median of each feature column. The engineered features were then integrated with the masked local and aggregated features from the original Elliptic dataset and the 17 augmented features introduced in Elliptic++. As the original Elliptic features were already normalised, normalisation was applied only to the Elliptic++ and newly engineered features to ensure consistency across the feature space.

The resulting feature set captured both direct transaction metrics and relational indicators that could enhance model discrimination between licit and illicit behaviour.

3.2.2 Feature Relationships

After feature engineering, a feature–target relationship analysis was conducted to understand how individual variables relate to transaction labels and to inform model selection. Performance in AML classification depends not only on how models handle class imbalance but also on how well they adapt to the structural characteristics of features. Following prior work on

feature relevance in financial crime detection [26, 46, 16], the dataset was analysed using four complementary metrics: Pearson correlation, Spearman correlation, linear regression (R^2), and mutual information. Each feature was then categorised as linear, non-linear, or weak according to the following thresholds:

- **Linear:** Features with strong Pearson correlation ($|r| > 0.5$) and strong linear fit ($R^2 > 0.3$).
- **Non-linear:** Features with weak Pearson correlation ($|r| < 0.3$) but strong monotonic or information-based relationships (Spearman $|\rho| > 0.5$ or mutual information > 0.1).
- **Weak:** Features that did not meet either criterion, indicating low correlation and limited explanatory or predictive value.

The analysis shows that most features (153) are weak predictors, while a smaller but meaningful subset (43) exhibit non-linear relationships with the target variable. No features met the threshold for strong linearity. This distribution confirms that linear models are inherently limited in capturing key AML signal patterns, whereas methods that can aggregate weak, non-linear relationships—such as tree-based ensembles—are better suited to this dataset. This insight provided empirical justification for evaluating both linear (logistic regression) and non-linear (random forest) classifiers in the subsequent sections.

3.2.3 Training and Validation Strategy

Model training was conducted using the labelled subset of the dataset, which contains 46,654 transactions (approximately 90% licit and 9.8% illicit). The remaining 157,205 transactions (77%) were unlabelled and were not used during model training or validation. Instead, once the best-performing model was finalised, it was applied to this unlabelled portion to generate predictions of illicit activity, representing the model's operational deployment phase.

To ensure realistic model evaluation, a temporal cross-validation strategy was used, as shown in Figure 3.1 and Figure 3.2. The labelled data were divided into ten folds, each representing a continuous time period of 30,000 observations, aligning with temporal training approaches adopted in prior AML research [26, 31, 46]. In each fold, the earlier half of transactions was used for training, and the later half for testing. This approach mirrors how financial crime detection operates in production environments, where models must learn from past transactions and classify new, unseen ones as they occur.

If the data were split randomly, the model could unintentionally learn patterns from future transactions that would not yet exist at prediction time. This "peeking into the future" problem, known as *data leakage*, gives artificially high results during testing and leads to poor real-world performance. Temporal validation prevents this by ensuring that the model only uses information available at the time of classification, providing a more realistic assessment of predictive ability [10, 9].

Hyperparameter tuning was performed independently within each fold, using only the training portion of that fold. The test portion remained unseen until that fold's evaluation. This

ensured that tuning did not leak information from future data and that results reflected true out-of-sample performance. Running tuning separately for each fold also provided a consistent and unbiased comparison between logistic regression and random forest models while minimising the risk of overfitting to any specific time window.

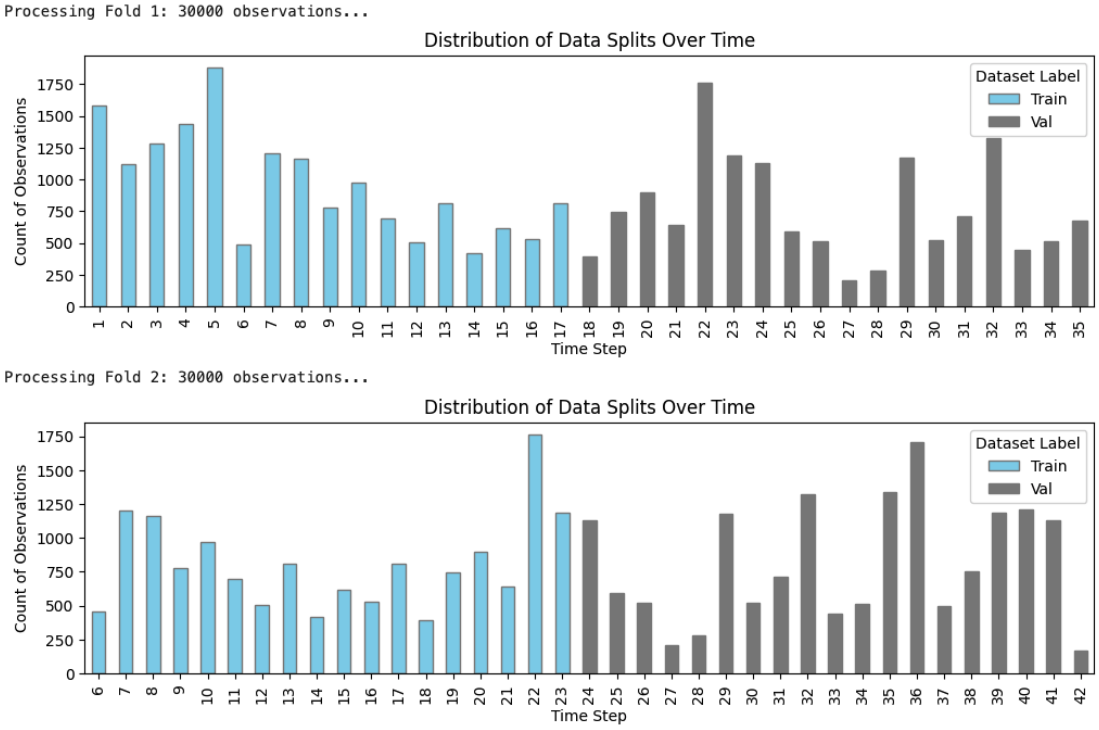


Figure 3.1: Distribution of data splits across time steps for cross-validation folds.

	Fold	Threshold	Total Size	Training Size	Validation Size	Class 0 Size	Class 1 Size	Train Class Distribution	Val Class Distribution	Min Time Step	Max Time Step	Accuracy	Recall	Precision	F1-Score	AUC-ROC	Gini Score	True Positive	True Negative	False Positive	False Negative
0	Fold 1	0.5000	30000	13953	16047	25865	4135	{0.0: 11860, 1.0: 2093}	{0.0: 14005, 1.0: 2042}	8	43	0.9826	0.8958	0.9760	0.9342	0.9921	0.9841	3704	25774	91	431
1	Fold 2	0.5000	30000	11630	18370	26555	3445	{0.0: 9745, 1.0: 1885}	{0.0: 16810, 1.0: 1560}	14	46	0.9759	0.9231	0.8739	0.8978	0.9891	0.9781	3180	26096	459	265
2	Fold 3	0.5000	30000	13788	16212	26133	3867	{0.0: 11990, 1.0: 1798}	{0.0: 14143, 1.0: 2069}	4	40	0.9734	0.8293	0.9590	0.8895	0.9235	0.8471	3207	25996	137	660
3	Fold 4	0.5000	30000	14795	15205	25867	4133	{0.0: 12707, 1.0: 2088}	{0.0: 13160, 1.0: 2045}	7	42	0.9842	0.9083	0.9751	0.9405	0.9942	0.9885	3754	25771	96	379
4	Fold 5	0.5000	30000	15445	14555	26316	3684	{0.0: 13709, 1.0: 1736}	{0.0: 12607, 1.0: 1948}	3	38	0.9729	0.8390	0.9336	0.8838	0.9280	0.8561	3091	26096	220	593
5	Fold 6	0.5000	30000	14561	15439	26318	3682	{0.0: 13085, 1.0: 1476}	{0.0: 13233, 1.0: 2206}	3	37	0.9714	0.8074	0.9526	0.8740	0.9288	0.8577	2973	26170	148	709
6	Fold 7	0.5000	30000	11706	18294	26626	3374	{0.0: 9816, 1.0: 1890}	{0.0: 16810, 1.0: 1484}	15	47	0.9744	0.9268	0.8572	0.8906	0.9886	0.9771	3127	26105	521	247
7	Fold 8	0.5000	30000	12266	17734	26127	3873	{0.0: 10356, 1.0: 1910}	{0.0: 15771, 1.0: 1963}	11	44	0.9807	0.8823	0.9653	0.9219	0.9907	0.9814	3417	26004	123	456
8	Fold 9	0.5000	30000	14915	15085	26345	3655	{0.0: 13434, 1.0: 1481}	{0.0: 12911, 1.0: 2174}	2	37	0.9717	0.8082	0.9520	0.8742	0.9195	0.8390	2954	26196	149	701
9	Fold 10	0.5000	30000	11802	18198	26540	3460	{0.0: 9901, 1.0: 1901}	{0.0: 16639, 1.0: 1559}	14	46	0.9760	0.9275	0.8727	0.8993	0.9897	0.9794	3209	26072	468	251

Figure 3.2: Random Forest cross-validation results across folds.

3.2.4 Logistic Regression

Logistic regression was implemented as the baseline classifier to provide a benchmark for performance comparison. Its advantages include simplicity, interpretability, and extensive prior use in AML-related machine learning studies [26, 46, 16].

Features were standardised prior to training. Hyperparameter tuning was conducted within each fold using only the training portion, with a search over regularisation type (L1, L2, and Elastic Net), penalty strength (C), and class weighting. The "saga" solver was selected to support all penalty types. Recall was used as the primary scoring metric, consistent with AML objectives where failing to identify illicit transactions is costlier than generating false positives. A probability threshold of 0.50 was selected as the initial operating point (see Figure B.1).

The final performance of logistic regression is summarised in Table 3.1. While the model achieved the highest recall, its Gini coefficient was substantially lower than random forest, indicating weaker ability to discriminate between licit and illicit transactions. Precision was also poor, meaning that many flagged transactions were false positives. Combined with lower overall accuracy, this suggests that logistic regression, even after optimisation, trades reliability for marginal recall gains. This outcome does not indicate overfitting, as evaluation occurred on temporally held-out data, but rather reflects structural limitations: the model's linear form cannot capture the non-linear relationships prevalent in blockchain transaction data.

In practical AML environments, this behaviour would overwhelm investigators with false alerts, diluting focus on genuinely suspicious transactions. Logistic regression therefore serves primarily as a benchmark for interpretability and comparison, rather than as an operational detection model.

3.2.5 Random Forest

Following the baseline evaluation of logistic regression, a Random Forest classifier was implemented due to its consistent success in prior AML research using the Elliptic and Elliptic++ datasets [26, 46, 16]. Random Forest models are well suited to the weak and non-linear feature relationships identified earlier, offering strong predictive performance while maintaining interpretability.

The same temporal 10-fold cross-validation procedure was used to ensure direct comparability with logistic regression and to prevent data leakage. Within each fold, hyperparameters were optimised for the number of estimators, maximum tree depth, minimum samples per split, and class weighting. Both models were tuned using recall as the primary scoring metric to prioritise sensitivity to illicit activity.

Threshold optimisation based on confusion matrix analysis resulted in a threshold of 0.40 (see Figures B.1 and B.2). Compared with logistic regression, Random Forest achieved higher accuracy, precision, F1-score, AUC-ROC, and Gini coefficient (see Table 3.1). These improvements indicate stronger overall discriminatory power and a better balance between sensitivity

and reliability. Precision was notably higher, meaning that a greater share of flagged transactions were genuinely illicit.

Importantly, these gains did not arise from overfitting. The nested temporal validation ensured that model tuning and evaluation were fully separated, and the ensemble's design—aggregating many shallow decision trees—allowed it to capture subtle non-linear patterns while maintaining low variance. As a result, the Random Forest demonstrated the best trade-off between operational sensitivity and investigative efficiency, making it both statistically robust and practical for real-world AML deployment.

Table 3.1: Comparison of model performance metrics before and after hyperparameter tuning.

Model	Threshold	Accuracy	Recall	Precision	F1-Score	AUC-ROC	Gini
Logistic Regression (Pre-tuning)	0.50	0.9445	0.6913	0.7270	0.7087	0.8316	0.6632
Logistic Regression (Post-tuning)	0.50	0.8480	0.9164	0.3833	0.5406	0.8785	0.7569
Random Forest (Pre-tuning)	0.40	0.9704	0.6990	0.9972	0.8219	0.8494	0.6988
Random Forest (Post-tuning)	0.40	0.9745	0.8823	0.8601	0.8711	0.9334	0.8668

3.2.6 Model Selection

The comparative performance of Logistic Regression and Random Forest is presented in Table 3.1. Logistic regression achieved the highest recall, meaning it was most sensitive to detecting illicit cases. However, this came at a substantial cost to precision, accuracy, and the Gini coefficient, resulting in a high rate of false positives and reduced discriminatory power. In operational settings, such performance would overwhelm investigators with excessive alerts and diminish the efficiency of AML workflows.

By contrast, the Random Forest model achieved a more balanced profile. Although its recall was slightly lower, it consistently outperformed logistic regression across all other metrics, including precision, F1-score, AUC-ROC, and Gini. This balance demonstrates its ability to detect illicit transactions more reliably while maintaining manageable alert volumes. The ensemble's capacity to aggregate weak and non-linear signals allowed it to generalise effectively without overfitting, producing outputs that are both stable and interpretable.

Random Forest was therefore selected as the final classifier. It provided the optimal trade-off between sensitivity and precision, aligning with the practical and regulatory demands of AML operations. The model's output—flagged illicit transactions—served as the input for the next stage of the methodology, where subnetworks of linked suspicious activity were constructed for deeper network-based analysis.

3.3 Subgraph Construction

Following transaction classification, this stage reconstructs networks of illicit activity by linking together suspicious transactions into coherent, traceable subnetworks. Prior research in AML has often analysed entire blockchain graphs to detect clusters of illicit actors. However, this assumes full visibility of the transaction network—a condition rarely available in operational contexts. In practice, waiting for sufficient data to construct a complete background

Table 3.2: Summary of subgraph construction steps

Step	Description
1. Inputs	Extract a directed transaction-to-transaction table over a defined time window, with binary labels from the classification model: illicit or licit.
2. Choose a seed	Select a single illicit transaction (an AML "alert") as the starting point.
3. Forward BFS	Explore outward along transaction edges. Stop expansion when a licit node is reached. Only illicit nodes and their edges are added. Track visited nodes to avoid loops.
4. One subgraph per seed	The BFS produces an illicit-only subgraph. If no illicit neighbours are found, the seed itself forms a single-node subgraph.
5. Deduplicate and consolidate	After repeating for all seeds: (i) merge exact duplicates, and (ii) merge subgraphs that share at least one illicit node.
6. Output	The result is a set of consolidated illicit-only subnetworks. These become inputs for ranking and visualisation.

graph also introduces significant risk, as it allows criminal activity to continue unchecked during the delay. Investigative teams must therefore act on partial and recent information, following the flow of funds as it occurs rather than retrospectively analysing the entire network.

The goal of this stage is to identify the underlying criminal activity within a broader organised crime structure by tracing the flow of illicit funds from low-level money mules toward more senior coordinators. This mirrors the investigative workflow observed in financial crime investigations, where understanding the hierarchy and movement of value is critical to identifying key decision-makers and network enablers.

This seed-based design addresses three major gaps identified in the literature:

1. It replaces the unrealistic assumption of complete graph visibility with a localised, incremental reconstruction process.

2. It transforms classification outputs into actionable investigative units that can be individually prioritised, ranked, and visualised.

3. It supports scalability and explainability, key requirements for AML systems operating under regulatory oversight.

Each subnetwork represents a self-contained segment of potentially illicit flow, suitable for separate triage or ranking analysis.

Algorithm 1: Illicit-Only Subgraph Construction

Input: Directed transaction graph $G = (V, E)$ with binary labels $\{illicit, licit\}$, set of illicit seed transactions $S \subseteq V$

Output: Set of consolidated illicit-only subnetworks $\mathcal{G}$

$\mathcal{G} \leftarrow \emptyset$; `// Initialise output set of subnetworks`

foreach $s \in S$ **do**

- Initialise subgraph $G_s \leftarrow \{s\}$;
- Initialise queue $Q \leftarrow [s]$;
- Initialise visited set $V_s \leftarrow \{s\}$;
- **while** Q *not empty* **do**
  - $u \leftarrow$ dequeue(Q);
  - **foreach** $(u, v) \in E$ **do**
    - **if** *label*$(v) = illicit$ ***and*** $v \notin V_s$ **then**
      - Add v and edge (u, v) to G_s;
      - Add v to V_s;
      - Enqueue v into Q;
    - **if** *label*$(v) = licit$ **then**
      - **continue** ; `// Do not add licit nodes or expand past them`
- Add G_s to $\mathcal{G}$;

`// --- Deduplication stage ---`

Step 1: Exact duplicates;

Remove duplicate subnetworks from $\mathcal{G}$ where node and edge sets are identical;

Step 2: Intersection merge;

while *there exist* $G_i, G_j \in \mathcal{G}$ *with illicit node overlap* **do**

- Merge G_i and G_j into G_{ij};
- Replace G_i, G_j with G_{ij} in $\mathcal{G}$;

return $\mathcal{G}$;

3.3.1 Efficiency and Computational Scalability

Anchoring expansion on illicit seeds and halting at licit nodes prevents exponential traversal across the predominantly licit blockchain graph. This targeted breadth-first search scales with the visited portion only, $O(V_s + E_s)$, rather than the entire transaction space. Consequently, subnetworks are computationally lightweight to generate and can be refreshed in near-real-time or frequent batch intervals, aligning with the operational cadence of AML monitoring.

3.3.2 Practicality under Class Imbalance

Because illicit activity represents less than 1% of total blockchain volume [12], full-graph network analysis is dominated by licit behaviour, obscuring meaningful illicit relationships. Restricting traversal to illicit nodes keeps subnetworks manageable while retaining high investigative value. It also prevents dilution of risk signals that occurs when licit edges are overrepresented.

3.3.3 Investigative Focus

By constructing networks solely from illicitly labelled transactions, the approach concentrates investigative attention on high-risk flows. Each illicit transaction—analogous to a compliance alert—serves as the root of a subnetwork that captures only those subsequent transactions also classified as illicit. This mirrors investigative practice, ensuring that no alert is ignored and that analysts are presented with compact, interpretable visual structures rather than the overwhelming sprawl of global transaction graphs.

3.3.4 Auditability and Reproducibility

The forward-directed, stop-at-licit procedure is deterministic, auditable, and easy to communicate to regulators. The resulting subnetworks are reproducible, satisfying compliance expectations for transparency and model governance. This contrasts with probabilistic or learning-based graph construction methods, which may produce variable structures between runs.

3.4 Ranking Algorithm

The ranking algorithm establishes investigative priority—a triage mechanism that directs analysts toward the most influential transactions within each illicit subnetwork. In practical AML operations, investigators face large volumes of suspicious transactions and limited resources. Without prioritisation, time is often spent on low-impact cases while central laundering pathways remain undiscovered.

This component addresses a significant gap in prior AML and blockchain research, which has focused primarily on detection and classification performance rather than on post-detection prioritisation. Few studies propose an interpretable, regulator-auditable ranking framework to determine which transactions within a detected network should be examined first. The algorithm developed in this thesis fills this gap by providing an operationally practical, transparent, and reproducible triage system designed to help investigators follow the flow of illicit funds efficiently and systematically.

The method also supports the overarching investigative goal of tracing illicit transactions upwards through the network hierarchy—from low-level money mules toward more senior or coordinating actors. By prioritising structurally central and high-value transactions, the ranking process helps reveal those entities most likely to control or direct laundering operations.

3.4.1 Rationale and Method Design

PageRank, a graph-based centrality measure originally developed to rank web pages [35], forms the structural foundation of the ranking method. Its recursive formulation captures how influence propagates through connected nodes—making it well suited to blockchain transaction graphs where value flows directionally from one transaction to another.

Prior AML and financial network research has applied centrality metrics such as PageRank, degree, and betweenness to identify key actors in criminal or financial systems [30, 18, 52, 14, 37]. However, these studies generally analyse entire static graphs or aggregate criminal clusters, which limits their operational relevance. In real investigative environments, analysts rarely have full visibility of all transaction data and must instead work with partial illicit subnetworks derived from recent or ongoing alerts.

While PageRank effectively measures structural importance, money laundering is also driven by value flow—how and where illicit funds converge or disperse. To capture both dimensions, this algorithm integrates PageRank with financial and degree-based indicators: inbound Bitcoin value, in-degree, and inverse out-degree. This composite approach ensures that transactions are prioritised not only by network position but also by their role in consolidating or redistributing value.

All measures are expressed as percentiles within each subnetwork to mitigate the extreme skew typical of blockchain data, where high-volume service nodes (such as exchanges and mixers) dominate standardised scales. Percentiles indicate each transaction's relative standing rather than its raw value, preserving meaningful variation within the subnetwork. Unlike min–max or z-score scaling, percentile transformation is robust to outliers, preventing very large transactions from compressing the scale and allowing smaller yet structurally significant transactions to remain visible in the ranking.

Together, these design choices ensure that the algorithm remains computationally efficient, interpretable, and regulator-aligned—balancing network-theoretic rigour with the practical requirements of financial crime investigation.

3.4.2 Composite Ranking Calculation

The ranking algorithm integrates structural and financial indicators to quantify the relative influence of each transaction within an illicit subnetwork. Unlike traditional network metrics that capture only one aspect of importance, this composite approach combines connectivity, value accumulation, and flow behaviour into a single interpretable score. The aim is to support investigative triage by ranking transactions that are both central to the network and significant in terms of illicit value flow.

Four complementary indicators are used to represent distinct aspects of transaction behaviour:

1. **Unweighted PageRank** — measures structural centrality and connectivity within the subnetwork, treating all edges equally. This captures how transactions relate to one another independent of transaction volume, reflecting structural influence rather than absolute value.
2. **Inbound Bitcoin value** — the total BTC received by a transaction, identifying aggregation points where funds converge and potentially accumulate illicit value.
3. **In-degree** — counts the number of distinct preceding transactions feeding into a node, indicating consolidation from multiple sources.

4. **Out-degree (inverse)** — penalises transactions that disperse funds widely, highlighting those that retain value rather than distributing it across multiple outputs.

Each metric is converted into a normalised percentile (0–100) within its respective subnetwork to maintain comparability across networks of varying size and density. The final composite score is then computed as a weighted sum of these percentile components:

$$\text{CompositeScore}_i = w_{pr}\, PR_{\text{pct},i} + w_{val}\, InBTC_{\text{pct},i} + w_{in}\, InDeg_{\text{pct},i} + w_{out}\, \big(100 - OutDeg_{\text{pct},i}\big),$$

where $w_{pr} = 0.60$, $w_{val} = 0.30$, $w_{in} = 0.07$, and $w_{out} = 0.03$. These weights were determined through iterative testing and sensitivity analysis to balance the relative importance of structural connectivity and financial magnitude in the ranking outcome.

Each input component captures a specific behavioural characteristic:

- $PR_{\text{pct},i}$ — PageRank percentile representing *structural influence* by recursively weighting inbound links from other influential transactions.
- $InBTC_{\text{pct},i}$ — inbound Bitcoin value percentile measuring the *financial magnitude* or total value received by a transaction.
- $InDeg_{\text{pct},i}$ — inbound degree percentile representing *connectivity strength* and the diversity of incoming sources.
- $100 - OutDeg_{\text{pct},i}$ — inverse outbound degree percentile providing an *accumulation bias*, favouring transactions that retain rather than disperse value.

After calculating the composite score, two additional operations are applied to enable comparison across subnetworks and to produce a ranked ordering of transactions:

$$\text{CompositePct}_i = 100 \times \frac{\text{rank}\big(\text{CompositeScore}_i\big)}{N}, \qquad \text{CompositeRank}_i = \underset{\downarrow}{\text{argsort}}\big(\text{CompositeScore}\big),$$

where N is the number of transactions in the subnetwork. CompositePct_i gives the percentile position of each transaction within its subnetwork (0–100), while CompositeRank_i provides a complete ordered ranking from Rank 1 (most influential) to Rank N (least influential).

This percentile-based formulation ensures interpretability and prevents extreme transaction values from dominating results, making the method robust across subnetworks of varying size and structure. By combining structural and financial perspectives into a single interpretable measure, the ranking algorithm produces a consistent and auditable hierarchy that allows AML investigators to prioritise transactions most likely to represent central or high-value points in illicit fund flows.

3.4.3 Investigative Outcomes

The percentile-weighted ranking produces an interpretable and reproducible prioritisation system that directly supports investigative workflows. Its benefits include:

- **Prioritisation:** Highlights transactions most central to the flow of illicit funds.
- **Efficiency:** Directs investigative effort to high-impact nodes under resource constraints.
- **Comparability:** Percentiles standardise ranking across subnetworks of different sizes.
- **Transparency:** Each component is explicitly defined and weighted, ensuring auditability.
- **Investigative depth:** Once prioritised, detailed transaction attributes (size, direction, counterparties) are analysed by investigators.

In essence, the ranking algorithm transforms complex transaction graphs into practical investigative roadmaps. By combining structural analysis with financial significance, it bridges the gap between machine learning classification and investigative decision-making. The result is a clear, data-driven triage mechanism that helps AML professionals trace value flows efficiently, focus on the most influential nodes, and identify key actors within laundering networks.

3.5 Visualisation

The final stage of the methodology builds upon the preceding components—classification, subgraph construction, and ranking—by converting their analytical outputs into visual and tabular formats that investigators can interpret and act upon. Classified illicit transactions form the nodes of each subnetwork, which are subsequently ranked by their structural and financial importance and visualised to provide an investigator-focused representation of suspicious activity. This ensures that analytical results are translated into practical, decision-support tools for real-world AML investigations. While previous research has applied visualisation primarily as a means of demonstrating analytical performance rather than as an operational aid, studies such as [46, 16, 14] typically depict full background graphs or sample subgraphs to illustrate model behaviour. These large-scale representations, though useful for demonstrating methodology, are unsuitable for day-to-day investigative work due to their scale and lack of focus. This section addresses that gap by generating targeted, reproducible subnetworks designed for direct use by investigators.

Two complementary perspectives are produced: a transaction-to-transaction (txn–txn) view and an address-to-address (addr–addr) view. The txn–txn view captures direct Bitcoin flows between individual transactions, while the addr–addr view aggregates these flows into wallet-level relationships, revealing how value consolidates and redistributes across addresses. Together, these views balance granular transactional detail with higher-level behavioural context.

In both visualisations, node size represents unweighted PageRank, highlighting structurally central nodes within each illicit subnetwork, while node rank indicates their investigative

priority derived from the composite ranking procedure. Representative subnetworks are shown in Figure 3.4, illustrating both transaction and address perspectives of the same network and enabling investigators to trace how illicit value flows through transactions and between wallets.

Visualisation serves as the bridge between automated detection and human analysis. It converts complex analytical outputs into clear, interpretable artefacts that direct investigative attention toward key actors and relationships. In doing so, it enhances explainability, supports regulatory auditability, and ensures that advanced analytical methods remain transparent and operationally useful for financial crime investigations.

3.5.1 Investigative summary table

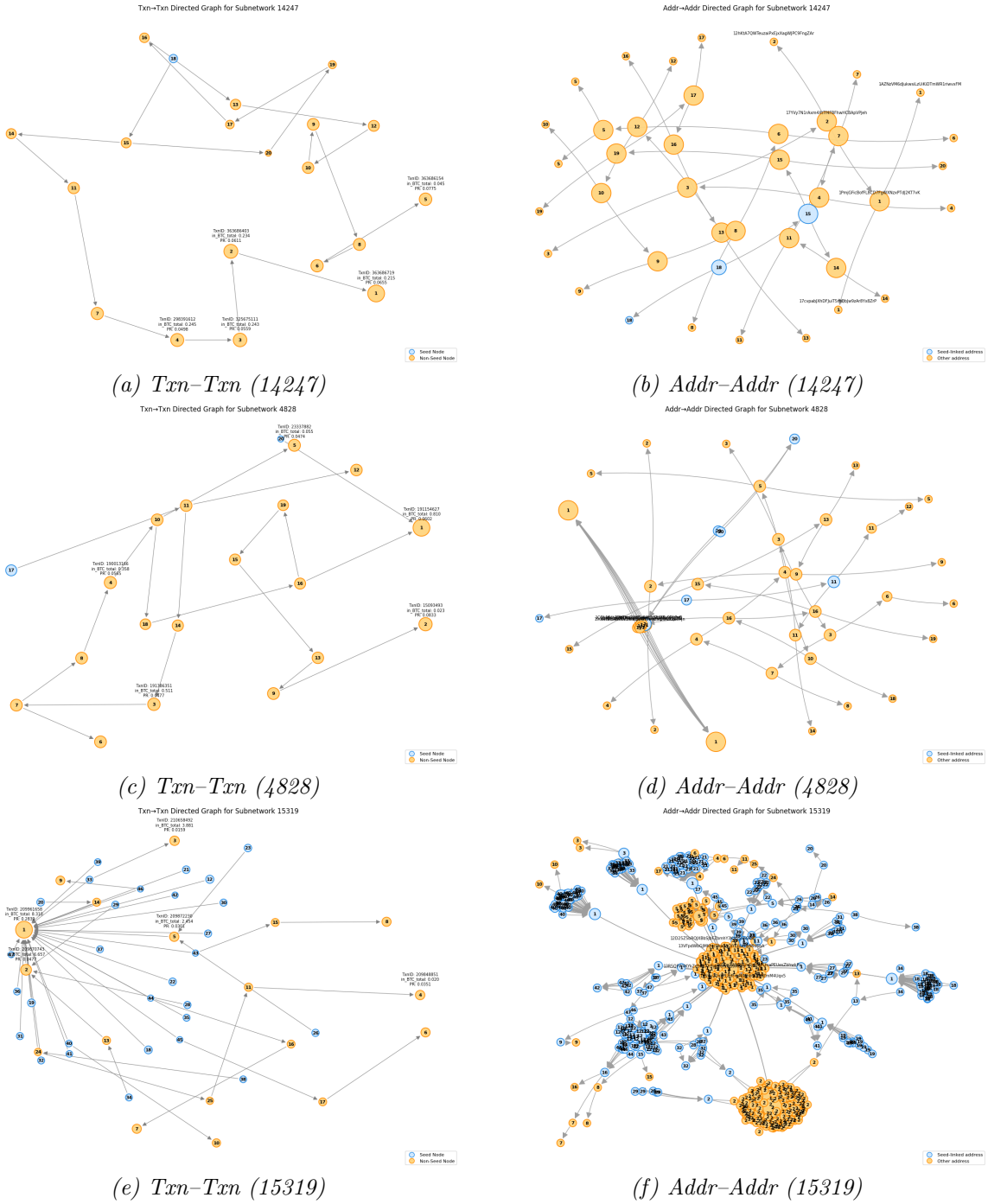
Each illicit subnetwork is also accompanied by a structured summary table (Table 3.3), which provides key transaction attributes and metadata in an auditable format. Investigators use the visualisations to identify priority nodes and the table to obtain detailed information for deep-dive investigations into specific transactions and wallets.

	subnetwork_id	txn_id	illicit_flag	seed_flag	investigation_order	in_BTC_total	composite_raw	input_addresses	output_addresses	n_inputs	n_outputs
0	0	230658142	True	False	1	3.5091	0.9869	125AS1eUZkjPNayhHEt2KLiHVzqZ5jExT1; 12ATFXLSx7...	1XFR1USgYp7gDHPhuNfqVjGF2hcKyLxo2	248	1
1	0	27405707	True	False	2	3.2574	0.9534	115Zkr1WmRWTqUoMvvMnyFVmiBcuo7b9QD; 115c96NaBJ...	3J1MmSusQRs7b4XA6hxK37t6PYK5gwf3VH	382	1
2	0	43560505	True	False	3	6.1129	0.9469	112hDfLPx3gGzUWVBmzfhNeeEtet4NW56G; 113X4nQpfK...	1KXepBmv9L2LsSxscaoi1Cr8XHazzK44T9; 36fFx8qprB...	380	2
3	0	230659438	True	False	4	0.1830	0.8423	1K8eEFjhEeh36iW7fMVZRY8Mpv1GPwkwSU	14hfBvNgqU8zBVmEUsiCj5td5QnU8eLaFA; 1FzMYbNq4D...	1	2
4	0	232377194	True	False	5	0.0416	0.7639	14MhC6hhUGePGZgNUhuk3YLtwoGZTAPRp4	1FrVYLBRPrmZ3bpeWX6SXYaY73fZpzYeM9; 1NDe8tkEPE...	1	2

Figure 3.3: Summary table excerpt from Subnetwork ID 0.

3.5.2 Outcomes

The combination of ranked network visualisations and structured tables delivers both interpretability and practical investigative value. Visualisations enable rapid triage of suspicious networks, while tabular summaries provide the detail required for subsequent casework and compliance reporting. Together, these outputs transform analytical findings into operational insights that investigators and regulators can readily understand and verify.



(a) Txn–Txn (14247) *(b) Addr–Addr (14247)*

(c) Txn–Txn (4828) *(d) Addr–Addr (4828)*

(e) Txn–Txn (15319) *(f) Addr–Addr (15319)*

Figure 3.4: Txn–Txn and Addr–Addr subnetwork visualisations for three subnetworks. Each row shows the two perspectives of the same subnetwork.

Chapter 4

Results and Analysis

This chapter presents the empirical results of the proposed anti-money-laundering (AML) framework and evaluates its effectiveness in addressing the research goals. It begins with the classification of Bitcoin transactions as illicit or licit, proceeds to the construction and ranking of illicit subnetworks, and concludes with the visualisation of those subnetworks for investigative interpretation. Together, these stages demonstrate how analytical outputs can be transformed into clear, investigator-oriented intelligence that supports decision-making in real-world financial crime investigations.

The analysis is structured to bridge a key gap identified in earlier chapters: while existing studies typically report high model performance on benchmark datasets, they rarely translate those results into practical or explainable tools suitable for operational AML use. The results presented here address this gap by demonstrating how classification probabilities, structural network measures, and visual summaries can collectively guide investigators in identifying and prioritising suspicious activity. This progression ensures that each analytical step contributes not only to improved detection accuracy but also to transparency, traceability, and explainability — qualities that are essential for compliance and auditability in regulated environments.

Each subsection therefore not only reports quantitative metrics but also interprets their significance in investigative and regulatory contexts. The evaluation considers both technical performance and practical implications, highlighting trade-offs, limitations, and sources of uncertainty. This ensures that the analysis reflects the dual objectives of analytical robustness and interpretability, aligning with the principles of responsible and regulator-compatible machine learning in financial forensics.

4.1 Classification Results

4.1.1 Classification Performance

The first analytical stage of the results evaluates the performance of the Random Forest (RF) classifier, which provides the foundation for the proposed anti-money-laundering framework.

This stage directly addresses Research Question 1 — identifying which classification approach is most effective for detecting suspicious Bitcoin transactions in an imbalanced dataset while maintaining transparency and auditability for both investigators and regulators. The model's outputs also underpin the construction of illicit-only subnetworks (RQ2), ranking of key actors (RQ3), and network visualisation for interpretability (RQ4).

A Random Forest classifier was trained and validated on the labelled portion of the Elliptic++ dataset which contains both licit and illicit transactions (Table 4.1). The labelled subset comprises 46,564 transactions, of which only 9.8% are illicit and 90.2% are licit. This strong imbalance reflects the real-world challenge of identifying illicit activity within large-scale financial networks, where legitimate transactions overwhelmingly dominate. Prior AML research using the Elliptic and Elliptic++ datasets [46, 16, 22, 8] has focused primarily on improving model accuracy or experimenting with advanced architectures such as Graph Convolutional Networks. However, these studies rarely consider whether the resulting models are explainable, auditable, or operationally useful to investigators and regulators. The present analysis addresses this gap by assessing the classifier's effectiveness not only in statistical terms but also in its capacity to generate trustworthy, interpretable, and regulator-aligned outputs that can support real-world AML decision-making.

The RF model achieved strong predictive performance on the labelled data, with a recall of 88% and a precision of 86%, as shown in Table 4.2 and Table 4.3. These results indicate that the model captures most illicit cases while keeping false positives within a manageable range — a critical trade-off in compliance settings where the cost of missing illicit activity far outweighs the burden of additional alerts. Recall, in particular, measures the proportion of truly illicit transactions that are correctly identified, directly determining the model's ability to expose genuine criminal activity and mitigate risk. A lower recall would mean that a larger share of high-risk transactions remain undetected, posing both compliance and reputational risk to reporting entities. In contrast, a moderate number of false positives — while adding investigative workload — is generally acceptable when the system succeeds in surfacing the majority of illicit flows. In practical terms, for every 100 illicit transactions, about 88 were correctly identified and 12 were missed. This recall-oriented tuning aligns with AML model evaluation recommendations [26, 31], which emphasise sensitivity to illicit cases as the key metric for regulatory effectiveness and financial-crime prevention.

Feature-importance analysis of the trained Random Forest model (see Appendix C) reveals that both *local* and *aggregate* transaction features were most influential in predicting illicit activity. As described by Weber *et al.* [46], local features capture attributes intrinsic to a transaction — such as the number of inputs and outputs, transaction value, and fees — while aggregate features summarise the behaviour of neighbouring transactions, including maxima, minima, and correlation statistics. The dominance of both feature groups suggests that illicit detection benefits from combining direct transactional characteristics with contextual network information. Although specific feature names in Elliptic++ are masked, the consistency of this pattern with previous analyses [46, 14] supports the interpretability of the model and demonstrates that important predictive cues arise from both individual and relational transaction properties.

To support downstream network analysis, the classifier was then applied to the previously unlabelled "Unknown" transactions while retaining existing licit and illicit ground-truth labels.

As shown in Table 4.4, of the 157,205 unlabelled transactions (77% of the dataset), approximately 28.7% were classified as illicit and 71.3% as licit. This reclassification expanded the total number of illicit transactions to 49,707, a tenfold increase relative to the original labelled set, while removing the Unknown category entirely. Table 4.5 shows that over 90% of the final illicit group originated from previously unlabelled data. This outcome underscores both the reach and the responsibility of the classifier: it effectively surfaces new potentially suspicious activity, yet its predictions remain probabilistic and should be validated through subsequent network-level analysis.

Overall, the classification stage transforms a sparse and highly imbalanced dataset into a richer, investigator-ready foundation for network analysis. By selecting a recall-optimised, interpretable model and transparently reporting its decision trade-offs, this stage directly addresses a key gap identified in the AML literature — the lack of methods that translate machine-learning outputs into auditable, explainable, and operationally meaningful intelligence for investigators and regulators. The next sections build on this foundation by examining how these classified transactions form subnetworks of related activity and how those networks can be ranked and visualised to guide AML investigation and prioritisation.

Table 4.1: Dataset composition by class label

Class Label	Count	% of Total	% of Labelled
Illicit	4,545	2.23%	9.76%
Licit	42,019	20.62%	90.24%
Unknown	157,205	77.15%	–
Total	**203,769**	**100%**	**100%**

Table 4.2: Random Forest performance (threshold = 0.4) on labelled data only.

Metric	Value
Total Size	46,564
Accuracy	0.9745
Recall	0.8823
Precision	0.8601
F1-Score	0.8711
AUC-ROC	0.9334
Gini	0.8668
True Positive	4,010
True Negative	41,367
False Positive	652
False Negative	535

Table 4.3: Predictions vs. Actual labels (counts and within-group percentages).

(a) Counts

Actual	Pred. Illicit	Pred. Licit
Illicit	4,144	401
Licit	1,948	40,071
Unknown	45,162	112,043
Total	51,254	152,515

(b) Percentages within Actual group

Actual	Pred. Illicit (%)	Pred. Licit (%)
Illicit	91.18	8.82
Licit	4.64	95.36
Unknown	28.73	71.27
Total	25.16	74.84

Table 4.4: Final labels after applying predictions only to Unknown transactions.

(a) Counts

Actual	Final Illicit	Final Licit
Illicit	4,545	–
Licit	–	42,019
Unknown	45,162	112,043
Total	49,707	154,062

(b) Percentages within Actual group

Actual	Final Illicit (%)	Final Licit (%)
Illicit	100.00	–
Licit	–	100.00
Unknown	28.73	71.27

Table 4.5: Contribution of Unknown predictions to final label totals.

Final Label	From Known Labels	From Unknown Predictions	% from Unknown
Illicit	4,545	45,162	90.9%
Licit	42,019	112,043	72.7%

4.1.2 Model Comparison

The decision to adopt Random Forest (RF) as the primary classifier was informed by comparative model performance across prior studies and by the structure of the feature space. This subsection addresses Research Question 1 by evaluating how the tuned RF model in this study performs relative to the best reported implementations of comparable algorithms in the Elliptic and Elliptic++ literature, while considering both predictive effectiveness and model transparency.

As shown in Table 4.6, the tuned RF developed in this study achieved a recall of 88.2% and an F1-score of 0.87, outperforming all comparable classifiers tested on the same dataset family. The models included in the comparison table represent the best-performing configurations reported for each model type. In Weber *et al.* [46], each model—Logistic Regression (LR), Multi-Layer Perceptron (MLP), Graph Convolutional Network (GCN), and Random Forest (RF)—was optimised using the full Elliptic feature set, which combines local features, aggregate features, and node embeddings. In Elmougy and Liu [16], the Elliptic++ dataset was used to evaluate both individual classifiers (RF, XGBoost, and MLP) and ensemble configurations that combined two or three classifiers (e.g., RF+XGB and RF+MLP+XGB), all trained after feature selection. The results shown here therefore compare this study's tuned single RF model against the strongest single and ensemble classifiers from the literature. Despite the advantage typically gained from ensemble aggregation, the tuned RF in this study achieved a recall of 88.2%, improving on Elmougy and Liu's best ensemble (72.9%) by +15.3 percentage points and Weber *et al.*'s RF (67.5%) by +20.7 percentage points, while maintaining balanced precision.

All three studies employ temporal evaluation rather than random splits, but the validation design differs in its robustness. Weber and Elmougy each used a single chronological split between training and test data, providing a snapshot of performance over time. In contrast, this study applied a ten-fold temporal cross-validation framework with hyperparameter tuning within each fold. This procedure yields a more reliable estimate of generalisable performance,

reducing variance associated with a single test window and ensuring that hyperparameters are optimised for temporal consistency. The improved results therefore reflect both model optimisation and methodological rigour rather than dataset differences alone.

Recall remains the most important evaluation criterion for financial-crime detection, as it measures the proportion of truly illicit transactions correctly identified. In compliance terms, high recall directly reduces residual risk by minimising the likelihood that suspicious activity goes undetected. Prioritising recall over metrics such as accuracy or AUC is essential in highly imbalanced datasets, where a trivial licit-only model could exceed 90% accuracy while missing all illicit activity. The tuned RF's recall of 88% ensures that the majority of high-risk transactions are surfaced for review, aligning with regulatory guidance and model governance expectations [2, 5, 26, 31].

In summary, the Random Forest classifier offers the best balance between sensitivity, interpretability, and computational efficiency. It outperforms both single and ensemble models from the Elliptic and Elliptic++ studies, improving recall by 15–21 percentage points while maintaining transparency and auditability for investigators and regulators. These properties justify its selection as the core classification model used to generate labelled data for subsequent network construction, ranking, and visualisation stages.

Table 4.6: Comparison of classification model performance across studies, ranked by Recall. Note: Elmougy's RF, RF+MLP+XGB, and RF+XGB results are reported after feature selection.

Rank	Model	Precision	Recall	F1
1	This Study (RF)	0.8601	**0.8823**	**0.8711**
2	Elmougy [16] (RF+MLP+XGB)	0.9680	0.7290	0.8340
3	Elmougy [16] (RF)	**0.9860**	0.7270	0.8360
4	Elmougy [16] (XGB)	0.7930	0.7180	0.7540
5	Elmougy [16] (RF+XGB)	0.9870	0.7170	0.8260
6	Weber [46] (RF)	0.9710	0.6750	0.7960
7	Weber [46] (MLP)	0.7800	0.6170	0.6890
8	Elmougy [16] (MLP)	0.6110	0.6130	0.6120
9	Weber [46] (LR)	0.5370	0.5280	0.5330
10	Weber [46] (GCN)	0.8120	0.5120	0.6280
11	Elmougy [16] (LSTM)	0.7090	0.2230	0.3390

4.1.3 Threshold Analysis

The choice of decision threshold has a critical influence on model performance, particularly under strong class imbalance where only a small fraction of transactions are illicit. Threshold optimisation in this study was conducted using validation results from the ten-fold temporal cross-validation process, ensuring that the chosen operating point reflects consistent behaviour across time periods rather than a single test window. Figure 4.1 visualises the confusion matrix at the selected threshold.

As shown in Figure B.2 in Appendix B, performance was evaluated across a range of thresholds. At very low thresholds (e.g., 0.1), the Random Forest classifier achieved near-perfect recall (99%) but extremely poor precision (24%), generating more than 14,000 false positives—an unmanageable workload for investigators. At high thresholds (e.g., 0.8–0.9), precision approached 100%, but recall fell below 55%, excluding nearly half of the illicit cases and creating a significant compliance risk. A threshold of 0.4 provided the most practical trade-off, yielding 88% recall and 86% precision across the labelled subset. This configuration produced 652 false positives and 535 false negatives, meaning that roughly 12% of illicit transactions were missed and approximately 16 false alerts were generated for every 100 true illicit cases detected.

This threshold achieves a balanced compromise between investigative workload and compliance assurance. In regulatory terms, recall reflects the proportion of truly illicit transactions successfully identified—the key measure of residual risk within a transaction monitoring system. Missing illicit cases (false negatives) directly increases compliance exposure and undermines reporting obligations under frameworks such as the *AML/CTF Act 2006* and AUSTRAC's *Transaction Monitoring* guidance [2, 5]. Conversely, excessive false positives primarily increase investigative effort but pose lower regulatory risk. Prioritising recall over precision is therefore a deliberate compliance-aligned choice, ensuring that potential suspicious activity is surfaced even at the cost of additional review effort.

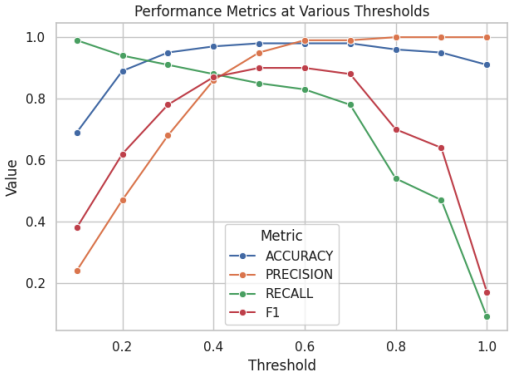


Figure 4.1: RF confusion matrix at $t = 0.4$ illustrating the recall–precision balance.

Accuracy was not suitable for threshold selection, as fewer than 10% of transactions were illicit—meaning a model that classifies all transactions as licit would still exceed 90% accuracy while failing to detect any high-risk activity. Instead, recall and precision provide more meaningful indicators of investigative value and system reliability. The selected threshold also preserves model explainability, as the decision boundary remains interpretable to investigators and auditors—a requirement often lacking in black-box AML models.

While a threshold of 0.4 produced the most balanced outcome in this dataset, it should be regarded as a tunable parameter that can be adjusted according to organisational risk appetite or jurisdictional expectations. Lower thresholds can be used for proactive or intelligence-led monitoring, while higher thresholds may be appropriate for operational production systems seeking to reduce alert volume.

Finally, extending predictions at the chosen threshold to the unlabelled portion of the dataset demonstrates its downstream impact: most of the newly identified illicit activity emerged from the previously "Unknown" transactions. This reinforces that robust threshold selection not only improves model reliability but also directly enhances the quality of data feeding subsequent subnetworks and ranking analyses, supporting Research Question 2.

4.1.4 Classification Limitations

Despite the strong performance of the Random Forest classifier, several limitations constrain the interpretability and generalisability of its results. These arise from both the structure of the dataset and the design assumptions of the model, and they are important to acknowledge to maintain transparency and auditability as required by Research Question 1.

First, the licit and illicit ground-truth labels in the Elliptic dataset, developed by IBM researchers in collaboration with the analytics firm Elliptic, cannot be independently verified. Any bias or misclassification in these annotations propagates through model training and evaluation, introducing epistemic uncertainty that limits validation against external benchmarks.

Second, the dataset is dominated by "Unknown" transactions, and while the model expands this category into predicted licit and illicit labels, these remain probabilistic and unverified. As shown in the confusion matrix results in Appendix B (Figure B.2), even at the optimised 0.4 threshold, 535 illicit cases were missed and 652 licit cases were incorrectly flagged, reflecting the inherent trade-off between recall and precision that all AML systems must balance.

Third, the dataset lacks wallet-type or service-level metadata that would distinguish exchanges, mixers, payment processors, or other intermediaries. Without this contextual information, large legitimate service nodes can dominate network connectivity, introducing structural noise that obscures laundering patterns and reduces interpretability in downstream network analysis.

Fourth, inspection of wallet timestamps indicates that the Elliptic dataset captures activity primarily between 2016 and 2017, as seen in Appendix E, reflecting laundering typologies of that period. Since then, blockchain-based laundering has evolved to include decentralised exchanges, cross-chain mixers, and privacy coins. These changes introduce domain drift, meaning that relationships learned from historical data may weaken as typologies evolve.

Finally, while recall was prioritised to minimise undetected illicit cases in line with regulatory expectations, the optimal balance between recall and precision is context-dependent. Thresholds should be adapted to the investigative capacity, risk appetite, and jurisdictional requirements of each implementing institution.

Taken together, these factors highlight that the classifier, while effective, represents only one layer of the investigative process. Subsequent subnetwork extraction, ranking, and visualisation steps are designed to mitigate these limitations by providing structural context, improving interpretability, and ensuring that analytical results remain auditable, explainable, and regulator-aligned.

4.2 Network Development and Visualisation Results

Building on the classification results, this section examines how identified illicit transactions cluster into subnetworks and how these are visualised to reveal laundering structures. This analysis directly addresses Research Question 2, which explores how illicit-only subnetworks

can be efficiently constructed to mirror investigative workflows and reduce computational overhead.

Previous studies such as Weber et al. [46], Elmougy et al. [16], and Ouyang et al. [34] typically analysed or visualised the entire transaction graph—often containing millions of licit nodes—and focused on improving predictive accuracy through complex models such as graph convolutional networks. While these studies advanced classification performance, they offered limited interpretability and scalability for investigative use. In particular, they did not develop methods for isolating or visualising distinct illicit-only subnetworks that investigators could directly explore or prioritise. This lack of operationally oriented, explainable visual analytics represents a key research gap that this work addresses.

Each subnetwork in this study was constructed using a forward breadth-first search (BFS) beginning from a known illicit transaction and expanding only through other illicit transactions until a licit boundary was reached. When multiple seeds converged on the same illicit node, their paths were merged into a single connected component. This method reflects how analysts trace suspicious flows in real-world AML investigations—progressively following confirmed illicit activity while excluding irrelevant licit background.

From 49,707 illicit transactions, 32,507 de-duplicated subnetworks were initially extracted, later reduced to 26,012 distinct illicit subnetworks after merging networks which had at least a common node. The majority of networks were small, with 76.46% containing a single illicit transaction (Table 4.7). This fragmentation is consistent with observations by Weber et al. [46] and Elmougy et al. [16], who found that illicit activity typically appears at the periphery of the transaction graph, where small clusters and low-degree nodes dominate. However, by constructing and analysing illicit-only subnetworks, this study extends prior work beyond academic performance metrics toward explainable, investigator-focused network representations that can be directly used in operational triage and analysis.

Subsequent analysis focuses on subnetworks with two or more transactions (n = 6,123), particularly those extending beyond two (n = 2,571). These represent the more structurally complex networks, where layering, consolidation, or obfuscation behaviour becomes visible through visual analysis.

Table 4.7: Distribution of subnetworks by transaction count

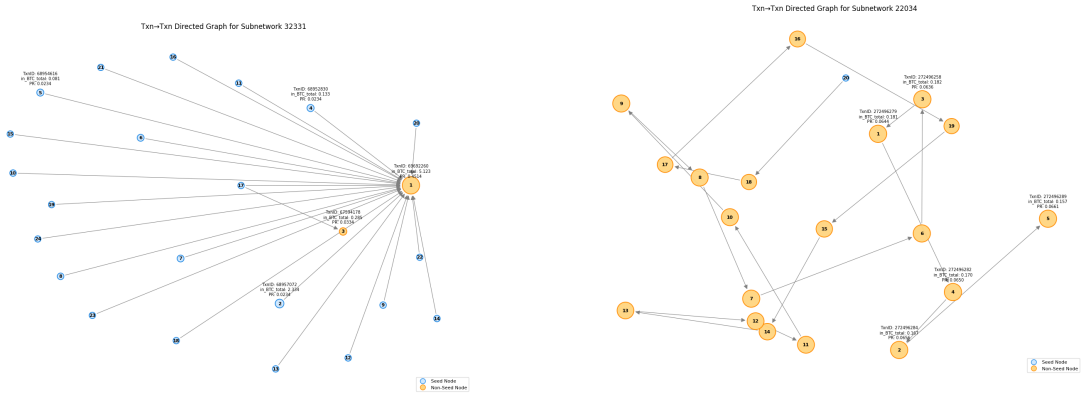
Transactions in network	**Count of NW**	**% of total (26,012)**	**% of networks ≥2 (6,123)**
1	19,889	76.46%	–
2	3,552	13.66%	58.00%
3	993	3.82%	16.22%
4	467	1.80%	7.63%
5	276	1.06%	4.51%
5+	835	3.21%	13.64%
Total	**26,012**	**100%**	**100%**

4.2.1 Analysis of Extended Subnetworks ($\geq$ 2 Transactions)

Network structure and visual patterns. The structural and visual analysis in this section addresses Research Question 4, which focuses on developing interpretable network visualisations that minimise the misidentification of service nodes such as exchanges or mixers.

Figure 4.2 illustrates two dominant topological patterns among extended subnetworks. Some form star-like structures, where many illicit transactions converge on a single node—typical of exchanges, mixers, or service hubs that consolidate high transaction volumes. Others form chain-like structures, where transactions are connected sequentially, consistent with the layering typology described in AML literature [18, 20, 4].

The joint distribution of transaction count and depth (Table 4.8) shows that both forms coexist. Shallow, star-like networks dominate, while longer chain-like subnetworks emerge in a smaller fraction, potentially representing coordinated layering or value movement across services. These patterns, clearly visible in the transaction-to-transaction visualisations, demonstrate how focused network extraction enables interpretable differentiation between laundering strategies—something largely absent from prior full-graph studies.



(a) Star-like txn–txn subnetwork (ID 32331)

(b) Chain-like txn–txn subnetwork (ID 22034)

Figure 4.2: Examples of star-like and chain-like transaction-to-transaction subnetworks.

Size and depth distribution. Among subnetworks with at least two transactions, 58% were simple two-transaction chains, while 42% extended further (Table 4.7). Depth analysis (Table 4.9) shows that 50.77% extended to depth three or more and 19% reached depth five or greater. These deeper subnetworks correspond to multi-layered laundering structures, where funds move through multiple intermediaries to obscure origin and ownership.

These findings reveal that illicit Bitcoin activity is fragmented but not random. Most flows terminate after one or two hops, but a smaller proportion form larger and deeper subnetworks that exhibit layering and consolidation typical of professional laundering. This outcome supports the research goal of creating interpretable, regulator-aligned representations of illicit behaviour that go beyond predictive modelling alone.

Table 4.8: Count of subnetworks by depth and transaction count (networks with ≥2 transactions, $n = 6{,}123$)

Txns in network	Depth = 1	Depth = 2	Depth = 3	Depth = 4	Depth = 5	Depth = 5+	Row Total
1–5	3,949	937	302	100	0	0	5,288
6–10	17	86	105	69	78	99	454
11–15	5	16	24	22	22	57	146
16–20	2	5	8	6	9	41	71
21+	3	13	20	26	18	84	164
Total	**3,976**	**1,057**	**459**	**223**	**127**	**281**	**6,123**

4.2.2 Visual Comparison of Transaction and Address Networks

The comparison between transaction-to-transaction (txn–txn) and address-to-address (addr–addr) networks further answers Research Question 4 by demonstrating how dual visual perspectives enhance interpretability and reduce the misidentification of service nodes.

As shown in Figure 4.3, subnetworks that appear as deep sequential chains in the txn–txn view often collapse into star-shaped structures in the addr–addr view due to address reuse, internal routing, or wallet consolidation within services. These dual perspectives reveal different but complementary aspects of the same activity: the txn–txn view captures temporal flow and sequence, while the addr–addr view highlights entity-level aggregation.

This combined visual framework directly addresses the research gap in explainable AML visualisation. Whereas prior studies have presented static or aggregate visualisations to demonstrate model performance, this work produces investigator-oriented, reproducible views that differentiate between genuine laundering flows and benign service behaviour—enhancing both interpretability and compliance value.

4.2.3 Limitations

Several limitations affect the interpretation of these results.

First, the BFS traversal expanded only through transactions classified as illicit, ensuring computational efficiency but making results dependent on classifier accuracy. Misclassified nodes can fragment paths, breaking continuity and creating artificially small subnetworks.

Second, differences in representational level can lead to misinterpretation. At the txn–txn level, sequential spending patterns resemble chains, while at the addr–addr level, these same flows appear as dense stars due to address reuse by exchanges or mixers (Figure 4.3). These hubs may not correspond to central criminal actors but to legitimate service infrastructure that aggregates flows.

Finally, the dataset lacks wallet-type metadata, providing no visibility into whether nodes represent service providers or end-user wallets. This limitation introduces structural noise and underscores the need for integrating off-chain or entity-level data—an open research

challenge noted by Gruber [21], Deprez et al. [14], and Samadi et al. [37], who emphasise the importance of linking on-chain and off-chain intelligence to improve the reliability of AML analytics.

Despite these limitations, the network development and visualisation framework effectively addresses Research Questions 2 and 4, closing a key gap in the literature. It demonstrates that illicit-only subnetworks can be efficiently constructed and visualised in a way that mirrors investigative practice, enhances interpretability, and supports explainable reasoning for both investigators and regulators.

Table 4.9: Distribution of subnetworks by depth

Depth group	**Count of NW (all)**	**% of total (26,012)**	**Count of NW (≥2)**	**% of networks ≥2 (2,147)**
1	23,865	91.75%	–	–
2	1,057	4.06%	1,057	49.24%
3	459	1.76%	459	21.38%
4	223	0.86%	223	10.39%
5+	408	1.57%	408	19.00%
Total	**26,012**	**100%**	**2,147**	**100%**

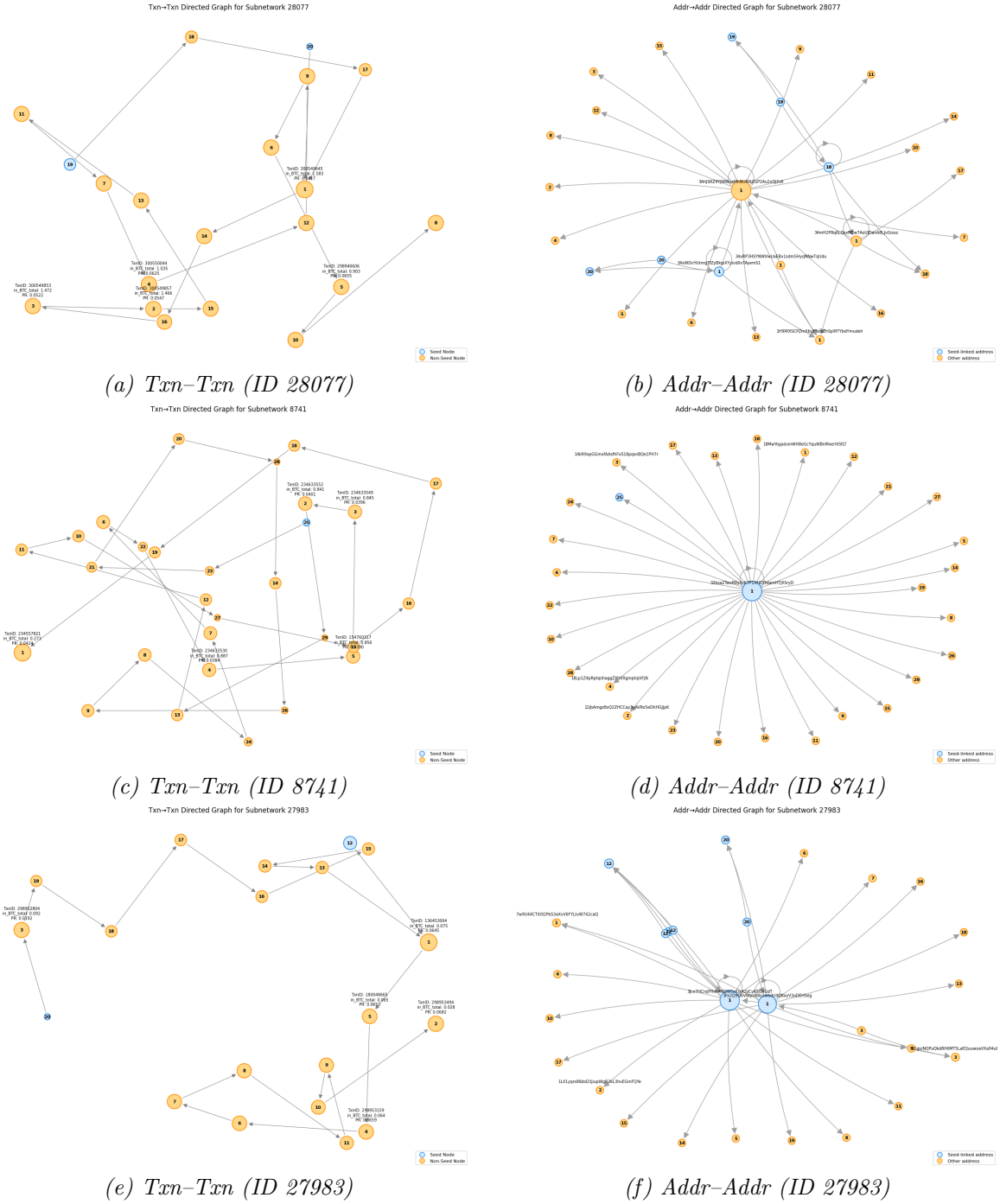
4.3 Ranking Results

The goal of ranking in this study is not merely to quantify influence but to support investigative triage. In anti-money laundering (AML) investigations, analysts must determine which transactions or wallets to review first among thousands of suspicious cases. Traditional network ranking algorithms, while effective at identifying structural influence, were not designed to prioritise investigative order in financial crime contexts. This section develops and evaluates a ranking system tailored to that operational goal, directly addressing Research Question 3: *How can a ranking method that integrates structural and financial indicators be designed to prioritise key actors within illicit subnetworks?*

4.3.1 Rationale for Composite Ranking

Previous AML network studies, such as Weber [46], Elmougy [16], Ouyang [34], and Samadi [37], have applied individual or pattern-based ranking measures—most commonly PageRank, degree-based scores, or connectivity motifs—to evaluate model performance or characterise node influence. However, these approaches do not explicitly integrate financial magnitude or provide interpretable outputs suitable for investigative triage.

Each traditional centrality measure captures a different aspect of network importance but has clear limitations when applied to Bitcoin transaction graphs (see Appendix D for metric definitions):



(a) Txn–Txn (ID 28077) *(b) Addr–Addr (ID 28077)*

(c) Txn–Txn (ID 8741) *(d) Addr–Addr (ID 8741)*

(e) Txn–Txn (ID 27983) *(f) Addr–Addr (ID 27983)*

Figure 4.3: Txn–Txn vs Addr–Addr views for three subnetworks. Each row shows the same subnetwork in two perspectives placed side by side. In several cases, a deep chain at the transaction level appears as a shallow, star-shaped structure at the address level due to address reuse and consolidation.

- *Unweighted PageRank* estimates structural influence based on incoming links but ignores transaction value, allowing nodes with many small transfers to appear more important than those controlling large flows.
- *HITS (Hub and Authority)* differentiates between senders and receivers but tends to overvalue dense, low-value clusters and is unstable in small directed graphs.
- *Degree centrality* measures transaction activity volume but not influence or control, meaning high-degree mixers or services may appear important despite limited risk.
- *Betweenness centrality* highlights nodes acting as bridges between subnetworks but may prioritise transient intermediaries rather than value controllers.
- *Eigenvector, harmonic, and Katz centralities* measure influence propagation but are sensitive to sparse graph structure and may overemphasise nodes in dense clusters.
- *Coreness* identifies dense subgraph membership but ignores transaction direction and value, limiting investigative interpretation.

These weaknesses underscore the research gap: existing network measures alone cannot provide a reliable, interpretable ordering of investigative priority. To address this, the composite ranking calculation introduced in Section 3.4.2 was developed to integrate both structural influence and financial significance into a single, explainable metric. This approach was specifically designed to align with AML investigative workflows, producing ranked outputs that can support analyst triage and prioritisation across large, complex illicit subnetworks.

4.3.2 Weight Sensitivity and Threshold Selection

The weighting configuration determines how strongly structural versus transactional properties influence ranking outcomes. Given that each component reflects a different investigative priority, the chosen weights directly affect which transactions are surfaced as high-risk. To identify a practical and interpretable balance, multiple weighting schemes were tested.

For each configuration, the top 20% of ranked nodes—drawn from subnetworks containing five or more transactions—were examined to measure how much of the total Bitcoin value and PageRank influence they captured (Table 4.10). This approach quantifies how effectively each weighting concentrates financial and structural importance within the top-ranked nodes, serving as a proxy for investigative focus.

Results showed consistent trade-offs across configurations.

- *PR-heavy* weightings concentrated structural influence (median PageRank percentile $\approx$ 92–93) but captured smaller Bitcoin shares, emphasising connectivity over monetary impact.
- *Value-heavy* weightings captured greater proportions of Bitcoin flow but reduced structural coherence, often identifying isolated high-value transactions with limited network centrality.

Intermediate configurations achieved the best balance. The *current* weighting (0.60 PageRank, 0.30 Value, 0.07 In-degree, 0.03 Out-degree) maintained high structural alignment (median PR percentile ≈ 91.7) while capturing meaningful financial concentration (median BTC share $\approx 26\%$ of the subnetwork BTC). This weighting captures nodes that are both structurally central and financially dominant, aligning with AML priorities that favour identifying key actors who control the flow of illicit value rather than simply those most connected.

Overall, this sensitivity analysis confirms that the selected weighting provides a stable, interpretable compromise between structure and value, supporting the objective of RQ3: to design a composite ranking method that prioritises key actors within illicit subnetworks for investigative triage.

*Table 4.10: Weight configurations and performance (top 20% of nodes, sorted by median PageRank percentile). The **current** row shows the selected weighting.*

Weight Config.	w_{pr}	w_{val}	w_{in}	w_{out}	**BTC Share**	**PR Pct**	n	**Top Frac**
pr_extreme	0.80	0.15	0.03	0.02	0.23	**92.86**	1111	0.20
pr_heavy	0.70	0.20	0.07	0.03	0.24	**92.31**	1111	0.20
pr_mid	0.55	0.35	0.07	0.03	0.28	**91.67**	1111	0.20
current	**0.60**	**0.30**	**0.07**	**0.03**	**0.26**	**91.67**	**1111**	**0.20**
deg_heavy	0.40	0.30	0.20	0.10	0.28	**91.38**	1111	0.20
balanced	0.50	0.40	0.07	0.03	0.28	**91.18**	1111	0.20
equal	0.25	0.25	0.25	0.25	0.27	**90.00**	1111	0.20
deg_equal	0.33	0.33	0.17	0.17	0.29	**90.00**	1111	0.20
val_mid	0.40	0.50	0.06	0.04	0.37	**75.00**	1111	0.20
val_heavy	0.20	0.70	0.06	0.04	0.47	**46.43**	1111	0.20
val_extreme	0.10	0.80	0.05	0.05	0.48	**35.71**	1111	0.20

4.3.3 Correlation Across Ranking Metrics

After establishing the composite rank, correlations were analysed against established network metrics to evaluate how closely the new measure aligns with existing concepts of structural influence (see Appendix D). This analysis assesses whether the composite ranking preserves the relative ordering of important nodes while introducing financial context.

Spearman's ρ was used to measure pairwise correlation across subnetworks containing five or more transactions. The composite rank correlated most strongly with PageRank ($\rho = 0.91$), harmonic centrality ($\rho = 0.87$), and Katz centrality ($\rho = 0.86$), confirming that it inherits the structural backbone of influence-based metrics while extending them through transaction-value weighting.

Lower or negative correlations with degree, betweenness, and coreness measures indicate that these capture different behavioural patterns—such as activity volume, bridging roles, or local clustering—rather than overall influence or control. This pattern is consistent with Elmougy [16], who found that structural measures such as PageRank and degree centrality alone contribute limited value to identifying illicit nodes unless combined with transactional indicators.

Overall, the correlation analysis confirms that the composite ranking integrates the most relevant aspects of structural influence and financial aggregation. It therefore meets the objective of RQ3: providing a balanced, interpretable ranking method that supports AML investigators in prioritising key actors with both network and monetary significance.

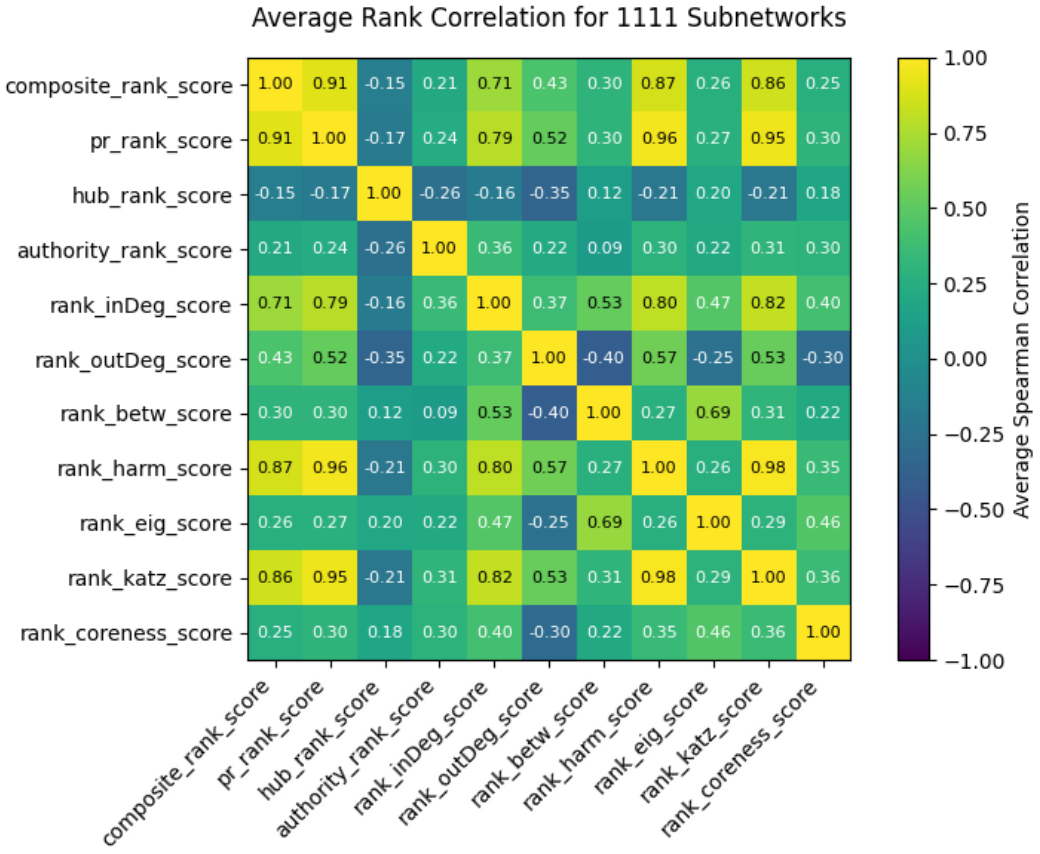


Figure 4.4: Average Spearman correlation between ranking methods across 1,111 subnetworks (minimum 5 transactions per subnetwork).

4.4 Constraints, Limitations, and Assumptions

This study was conducted under several constraints that shape the interpretation and generalisability of its findings. These constraints arise from the nature of the available data and from deliberate methodological trade-offs designed to balance practicality, scalability, and interpretability. While these factors limit absolute precision, they do not compromise validity. Rather, they reflect the operational realities of anti–money laundering (AML) research on pseudonymous blockchains and the need for transparent, auditable, and reproducible analytical frameworks.

4.4.1 Data Constraints

The Elliptic and Elliptic++ datasets, though the most widely used public benchmarks for Bitcoin AML research, present significant limitations. Only about 2% of transactions are labelled as illicit, resulting in severe class imbalance (RQ1). The majority (77%) are unlabelled, and the proprietary process used by IBM and Elliptic to assign labels has not been disclosed, preventing independent verification of ground truth. Consequently, all model predictions must be treated as probabilistic. In addition, transaction identifiers are masked, restricting external validation, and the absence of edge-level Bitcoin value weights limits precise modelling of transaction flows. These data constraints exemplify the broader research gap: a lack of transparent, verifiable AML datasets for benchmarking and validation [16, 14].

4.4.2 Methodological Constraints

Methodological design emphasised scalability and interpretability, supporting RQ1 and RQ3's focus on explainability and investigator usability. This required limiting algorithmic complexity—deep learning or graph neural networks were not implemented—to maintain transparency and auditability under regulatory expectations. The subnetwork construction process (RQ2) used a forward breadth-first search restricted to illicit nodes, improving efficiency but introducing dependency on classifier accuracy. Misclassifications, particularly illicit transactions labelled as licit, fragment flows and may underestimate network connectivity.

Although the Elliptic++ dataset includes wallet-level identifiers, this research intentionally operated at the transaction level. Address clustering was avoided to reduce additional assumptions and uncertainty about ownership relationships, which can distort network boundaries. Instead, addresses were summarised in tabular outputs (see Table 3.3) for reference rather than used for clustering or network aggregation. This design choice ensured that subnetwork structures remained transparent, interpretable, and reproducible, though it limited visibility into higher-level entity interactions (RQ4).

4.4.3 Computational and Practical Trade-offs

The framework was developed under limited computing resources and without live blockchain integration. As a result, processing was performed in batch mode rather than real-time, preventing dynamic modelling of temporal behaviour or transaction velocity. Percentile-based ranking was adopted for cross-network comparability, favouring interpretability over granularity. While this reduces sensitivity to fine value differences, it enhances usability for compliance analysts—a deliberate trade-off that supports auditability and regulatory transparency.

4.4.4 Assumptions

The study assumes that the labelled illicit transactions reflect genuine criminal and laundering activity. It also assumes that illicit funds flow forward through transaction links in accordance with Bitcoin's UTXO model, and that observed structures approximate historical laundering typologies. Finally, the ranking method assumes that nodes with higher composite scores (as

defined in Section 3.4.2) represent central actors or choke points in illicit networks, consistent with AML investigative priorities.

Chapter 5

Conclusion and Future Work

5.1 Summary and Contributions

The purpose of this research was to develop a practical and transparent framework for detecting, ranking, and visualising money laundering activity on the Bitcoin blockchain. The goal was not only to improve analytical accuracy but to design a system that investigators, regulators, and technical specialists can all understand, trust, and use. In anti-money laundering contexts, where accountability is critical, every analytical step must be transparent, traceable, auditable, and explainable. This research demonstrates that these principles can be achieved through a straightforward and interpretable design.

This study set out to answer four interconnected research questions that span classification, network construction, ranking, and visualisation. Transactions were first classified as illicit or licit (RQ1) using a transparent machine learning model. From these results, illicit-only subnetworks were constructed (RQ2) to isolate flows of suspicious activity. Each transaction within these networks was then ranked using a composite metric (RQ3) that integrates structural influence and financial significance, supporting investigative triage. Finally, results were presented through interpretable network visualisations (RQ4) that communicate complex transactional behaviour in a regulator-friendly format.

In addressing these research questions, this study also responds to a critical gap identified in prior work. Existing studies such as Weber [46], Elmougy [16], and Deprez [14] highlight the lack of AML detection systems that are both technically robust and operationally explainable. By integrating machine learning, network analysis, and interpretability into a unified pipeline, this research demonstrates that transparency and analytical sophistication can coexist in AML investigation tools.

Together, these components form a coherent and operationally realistic framework that bridges the gap between research and practice. The approach remains simple enough for non-technical investigators to understand, yet rigorous enough to meet regulatory standards of transparency, reproducibility, and auditability. Every output—whether a label, a ranking, or a network view—can be explained and defended, making the framework both credible and practical in compliance and enforcement environments.

Beyond its technical contribution, this thesis also advances an interpretive framework for designing AML tools that are human-centred and regulator-aligned. It argues that simplicity and interpretability are not limitations but essential qualities of effective financial crime detection. An AML framework that can be clearly understood by both data scientists and compliance officers has greater regulatory and investigative value than one that is opaque or overly complex. By focusing on clarity, usability, and accountability, this work provides a foundation for future AML systems that are transparent, trustworthy, and directly actionable.

5.2 Future Work

While the framework developed here establishes a transparent foundation for blockchain-based AML detection, several directions remain open for further exploration and development.

1. **Analytical expansion:** Linking transaction-level subnetworks to wallet or entity-level clusters would enable clearer identification of exchanges, mixers, and service providers, building on recent efforts to integrate off-chain context [37, 14]. Temporal analysis could further reveal evolving or recurring laundering typologies over time.
2. **Operational integration:** Incorporating live blockchain data and near real-time analytics would allow for continuous monitoring, while an interactive dashboard could unify classification, ranking, and visualisation into a single investigative interface.
3. **Generalisability testing:** Applying this framework to other blockchains—such as Ethereum or cross-chain ecosystems—would assess how well the design principles of transparency, explainability, and ranking generalise beyond Bitcoin.

Through these developments, and as demonstrated in this work, the framework presented in this thesis provides both a methodological and operational contribution to the field of anti-money laundering research. Methodologically, it demonstrates how machine learning, network analysis, and ranking can be combined into a single, explainable framework that balances analytical rigour with interpretability. Operationally, it delivers a transparent, regulator-aligned approach that supports investigative triage through traceable, auditable, and defensible outputs.

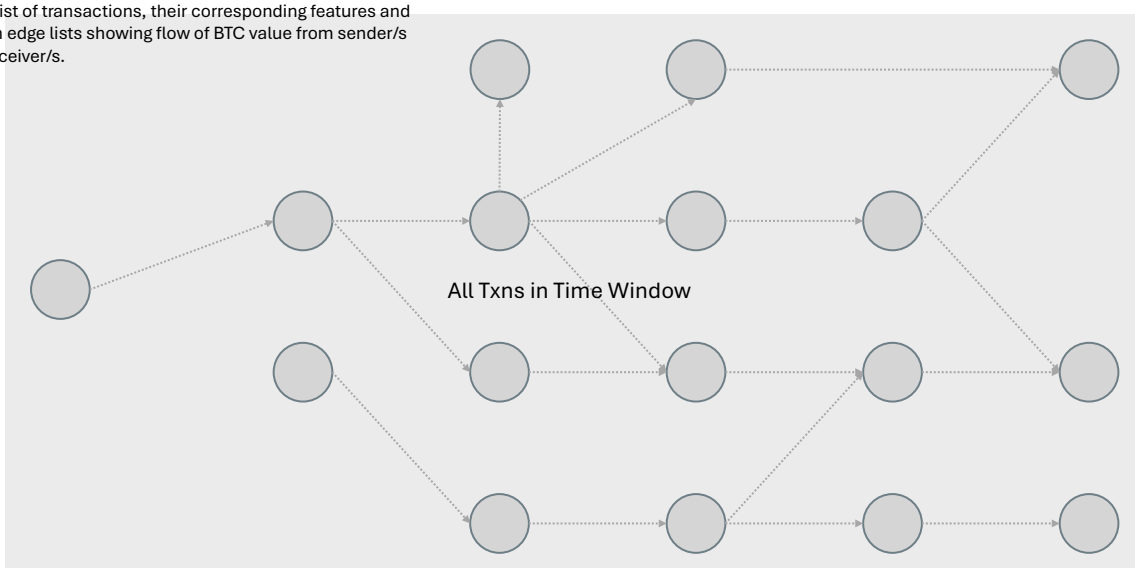
By unifying classification, network extraction, ranking, and visualisation in a coherent and explainable process, this research lays the groundwork for AML systems that are not only technically effective but also aligned with the real-world needs of investigators, compliance officers, and policymakers.

Appendix A

Methodology Overview

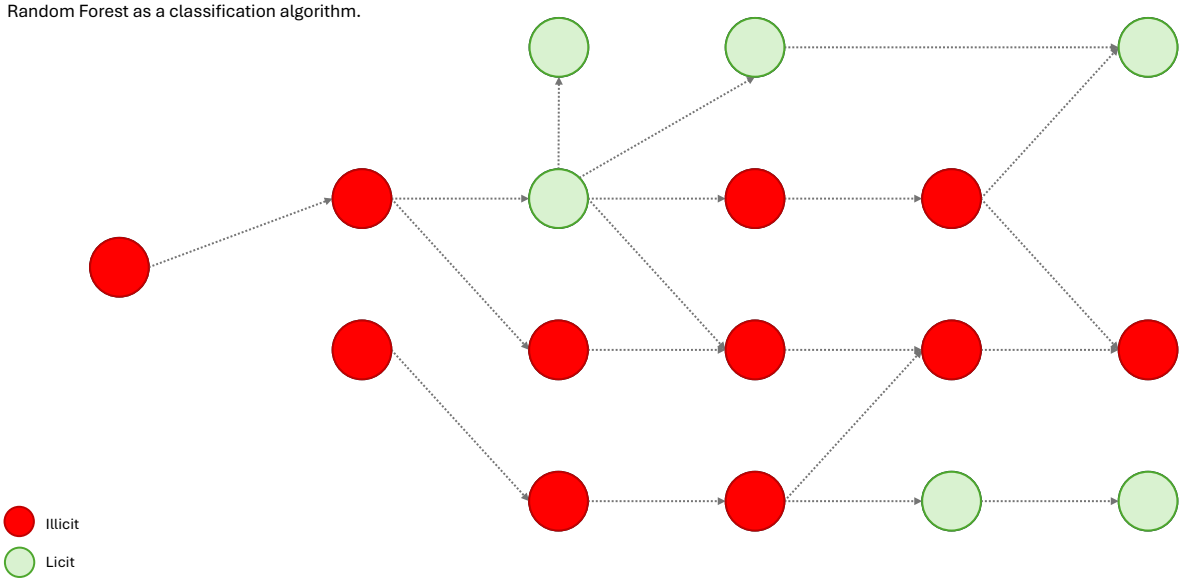
Methodology Diagram

Step 1: Get transactions (txn) for a specified window of time.
Get a list of transactions, their corresponding features and txn-txn edge lists showing flow of BTC value from sender/s and receiver/s.



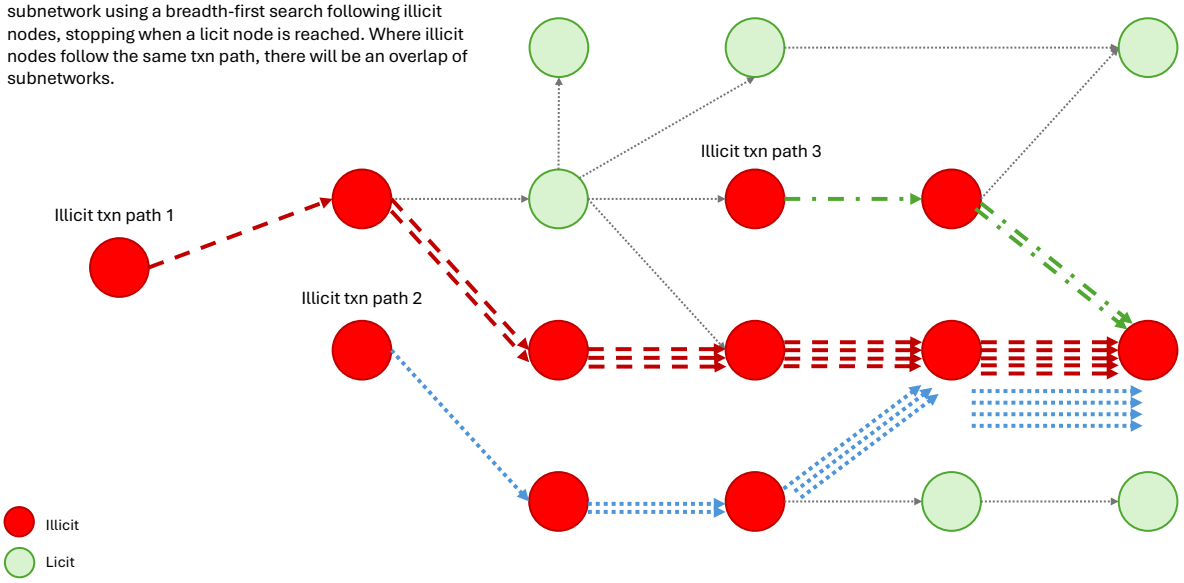
Methodology Diagram

Step 2: Classify txn as suspected illicit or licit.
Classify nodes as illicit or licit using a Random Forest as a classification algorithm.



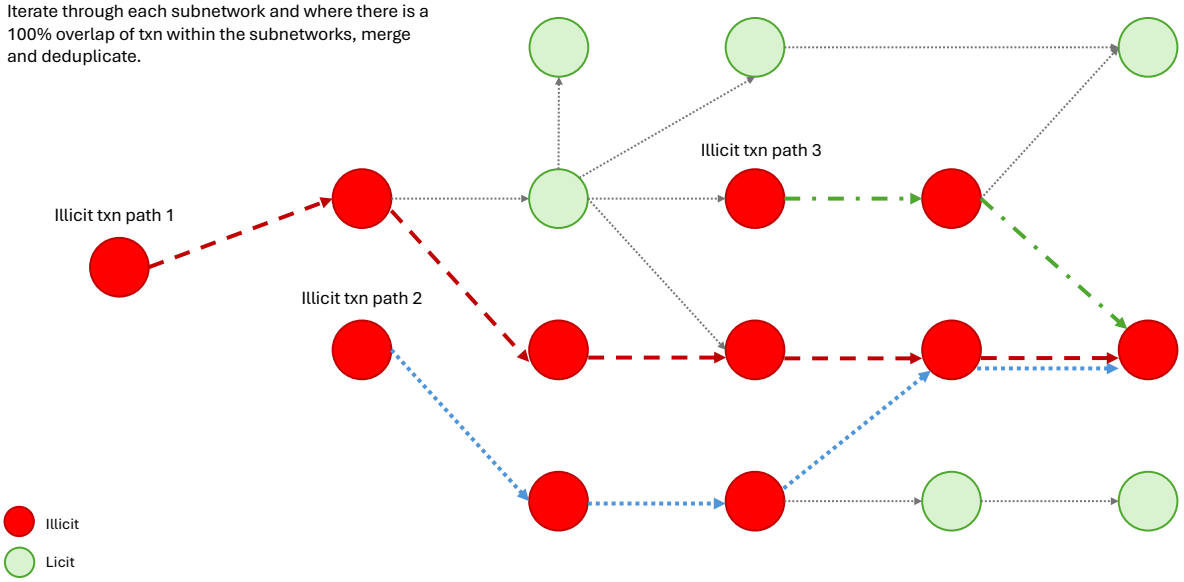
Methodology Diagram

Step 3: For each illicit txn, build a subnetwork.
Starting from each illicit node, build a forward-directed subnetwork using a breadth-first search following illicit nodes, stopping when a licit node is reached. Where illicit nodes follow the same txn path, there will be an overlap of subnetworks.



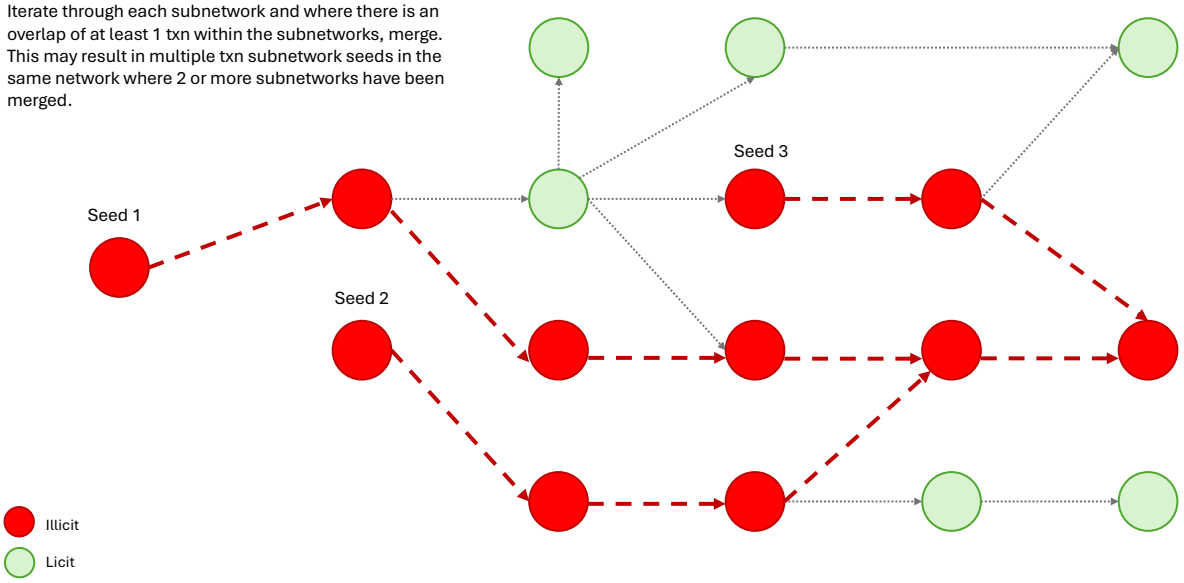
Methodology Diagram

Step 4: Deduplicate subnetworks where there is a 100% overlap
Iterate through each subnetwork and where there is a 100% overlap of txn within the subnetworks, merge and deduplicate.



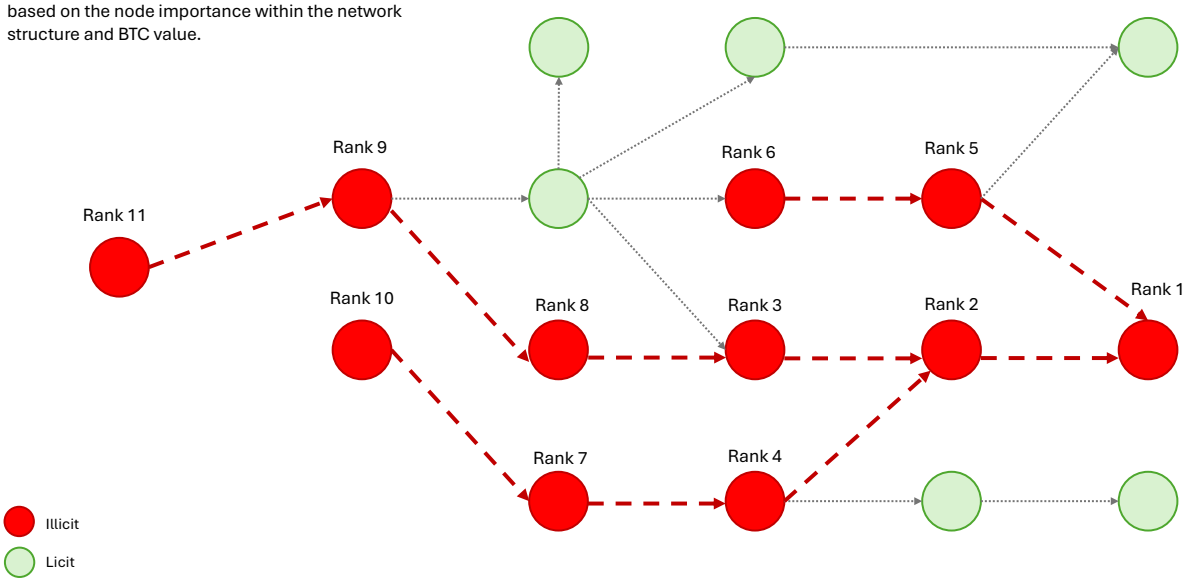
Methodology Diagram

Step 5: Merge subnetworks where there is at least 1 overlapping node
Iterate through each subnetwork and where there is an overlap of at least 1 txn within the subnetworks, merge. This may result in multiple txn subnetwork seeds in the same network where 2 or more subnetworks have been merged.



Methodology Diagram

Step 6: Rank txns within each network
For each txn in the subnetwork, assign a rank which is based on the node importance within the network structure and BTC value.



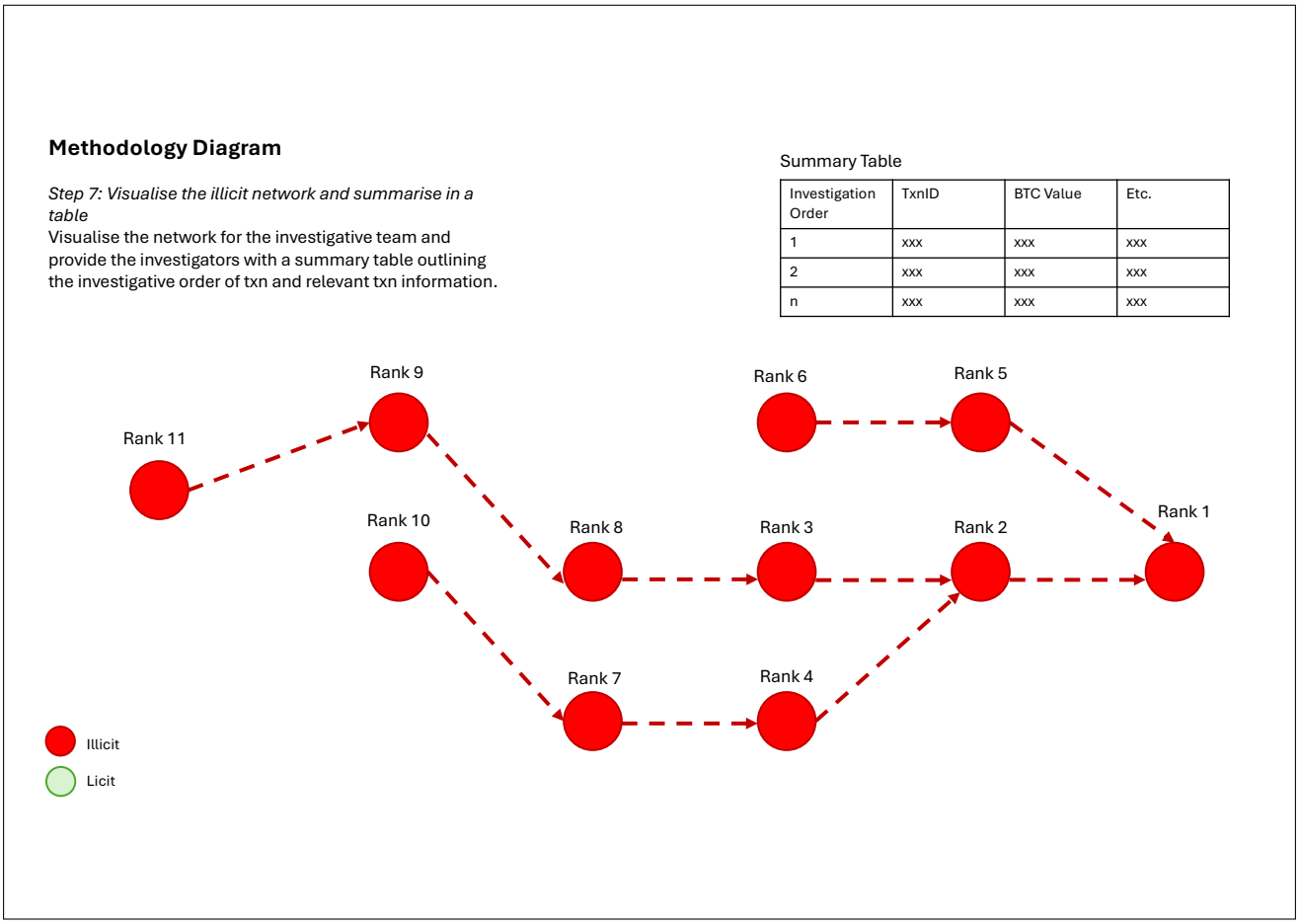
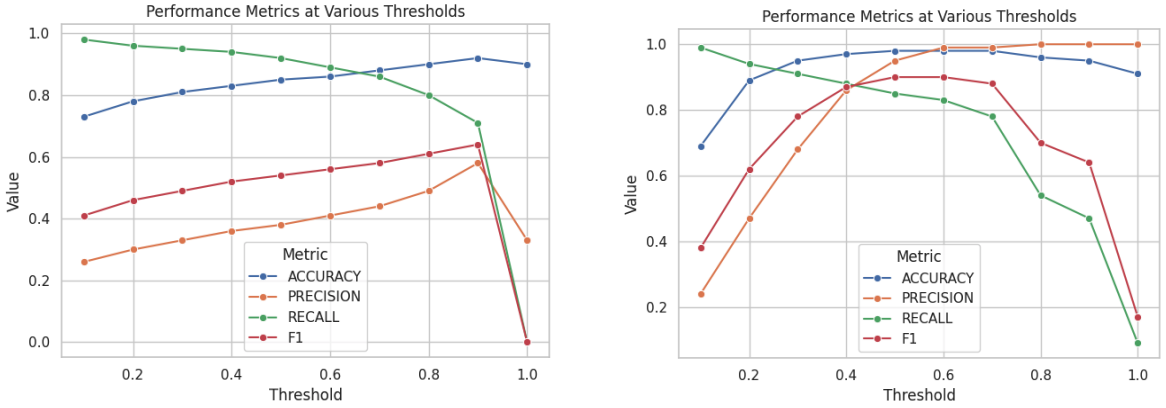


Figure A.1: Seven-step methodology overview.

Appendix B

Confusion Matrices



(a) Logistic Regression *(b) Random Forest*

Figure B.1: Confusion matrices of the classifiers at the selected decision threshold. (a) Logistic Regression. (b) Random Forest. The matrices show the distribution of true positives, false positives, true negatives, and false negatives.

THRESHOLD	ACCURACY	RECALL	PRECISION	F1	TRUE_POSITIVE	TRUE_NEGATIVE	FALSE_POSITIVE	FALSE_NEGATIVE	TOTAL
0.1000	0.6900	0.9900	0.2400	0.3800	4496	27455	14564	49	46564
0.2000	0.8900	0.9400	0.4700	0.6200	4267	37127	4892	278	46564
0.3000	0.9500	0.9100	0.6800	0.7800	4144	40071	1948	401	46564
0.4000	0.9700	0.8800	0.8600	0.8700	4010	41367	652	535	46564
0.5000	0.9800	0.8500	0.9500	0.9000	3876	41827	192	669	46564
0.6000	0.9800	0.8300	0.9900	0.9000	3793	41967	52	752	46564
0.7000	0.9800	0.7800	0.9900	0.8800	3560	41994	25	985	46564
0.8000	0.9600	0.5400	1.0000	0.7000	2474	42007	12	2071	46564
0.9000	0.9500	0.4700	1.0000	0.6400	2128	42017	2	2417	46564
1.0000	0.9100	0.0900	1.0000	0.1700	429	42019	0	4116	46564

Figure B.2: Confusion matrix table of the Random Forest classifier showing performance metrics at different thresholds.

Appendix C

Feature Importance

	feature	perm_importance_mean	perm_importance_std	gini_importance	rank_perm	rank_gini
0	Local_feature_53	0.0031	0.0015	0.0656	1.0000	2.0000
1	Local_feature_55	0.0030	0.0015	0.0589	2.0000	3.0000
2	Aggregate_feature_70	0.0026	0.0012	0.0069	3.0000	33.0000
3	Aggregate_feature_66	0.0020	0.0007	0.0027	4.0000	65.0000
4	Aggregate_feature_65	0.0019	0.0008	0.0031	5.0000	60.0000
5	size	0.0018	0.0010	0.0294	6.0000	11.0000
6	Local_feature_81	0.0015	0.0008	0.0034	7.0000	54.0000
7	Local_feature_2	0.0012	0.0009	0.0146	8.0000	18.0000
8	Local_feature_90	0.0011	0.0007	0.0146	9.0000	19.0000
9	Local_feature_80	0.0009	0.0006	0.0043	10.0000	45.0000
10	Aggregate_feature_51	0.0008	0.0004	0.0044	11.0000	44.0000
11	Local_feature_59	0.0006	0.0005	0.0069	12.0000	32.0000
12	Local_feature_3	0.0006	0.0007	0.0018	13.0000	87.0000
13	Local_feature_61	0.0005	0.0002	0.0050	14.0000	39.0000
14	Local_feature_18	0.0005	0.0006	0.0118	15.0000	22.0000
15	Local_feature_65	0.0005	0.0004	0.0074	16.0000	29.0000
16	Local_feature_76	0.0004	0.0006	0.0036	17.0000	50.0000
17	Local_feature_47	0.0004	0.0009	0.0663	18.0000	1.0000
18	output_address_percentile	0.0004	0.0010	0.0184	19.0000	15.0000
19	Aggregate_feature_61	0.0004	0.0003	0.0011	20.0000	100.0000
20	fees	0.0004	0.0009	0.0146	21.0000	17.0000
21	Aggregate_feature_16	0.0004	0.0003	0.0011	22.0000	102.0000
22	Aggregate_feature_20	0.0004	0.0002	0.0004	23.0000	144.0000
23	Local_feature_60	0.0003	0.0002	0.0108	24.0000	24.0000
24	Aggregate_feature_59	0.0003	0.0002	0.0005	25.0000	130.0000

Figure C.1: Top 25 features ranked by permutation and Gini importance from the Random Forest classifier. Both local and aggregate features exhibit the highest importance scores, consistent with findings by Weber et al. [46].

Appendix D

Ranking Metrics and Correlation Definitions

This appendix supports Section 4.3 and defines the network ranking metrics and correlation methods used in this study. All definitions are sorted alphabetically for clarity and written in accessible terms to assist both technical and non-technical readers.

Authority Score (HITS Algorithm): Derived from the Hyperlink-Induced Topic Search (HITS) algorithm, this measures how much a node receives from major senders. High authority scores indicate nodes that collect funds from influential sources, such as exchanges, collectors, or laundering endpoints.

Betweenness Centrality: Measures how often a node lies on the shortest paths between others. High betweenness nodes act as bridges linking subnetworks and may represent intermediaries that connect otherwise separate laundering flows.

Composite Rank (Thesis Contribution): A key contribution of this thesis. The composite rank is a weighted combination of PageRank, inbound Bitcoin value, in-degree, and inverse out-degree. It ranks nodes by both network influence and illicit value accumulation, supporting investigative triage and regulator-facing explainability.

Coreness: Determines how deeply a node is embedded within a dense transaction cluster. Core nodes often correspond to services or exchanges that facilitate large volumes of overlapping flows.

Eigenvector Centrality: Scores nodes based on their connections to other highly connected nodes, representing prestige or structural influence. High eigenvector scores may highlight central exchanges or mixers.

Harmonic Centrality: Measures a node's average closeness to all others, giving higher scores to nodes that can reach many others through short paths. It reflects accessibility and influence in transaction propagation.

Hub Score (HITS Algorithm): Also derived from the HITS algorithm, the hub score measures how actively a node sends Bitcoin to other important nodes. High hub scores indicate dispersers of funds, such as mixers, tumblers, or payment processors.

Inbound Degree (InDeg): Counts the number of unique incoming transactions to a node. Nodes with high inbound degree consolidate funds from multiple sources and may represent aggregation points in laundering chains.

Katz Centrality: Extends eigenvector centrality by considering both direct and indirect connections, applying a decay factor for longer paths. It captures influence propagation and indirect control in transaction networks.

Outbound Degree (OutDeg): Counts the number of unique outgoing transactions from a node. High outbound degree nodes distribute funds broadly, characteristic of dispersion or layering activity.

PageRank (PR): Evaluates structural influence by recursively weighting links from other important nodes. In Bitcoin networks, high PageRank values identify nodes that sit at the centre of major flows of value.

Spearman Rank Correlation (ρ): A non-parametric measure of how similarly two ranking methods order the same set of nodes. It is used to assess the consistency between the composite rank and standard centrality metrics.

Summary: Together, these metrics describe complementary aspects of network behaviour: influence (PageRank, eigenvector, Katz), connectivity (degree measures), bridging (betweenness, harmonic), and clustering (coreness). Correlating them with the composite rank demonstrates that the proposed approach captures both structural and financial significance, bridging the gap between theoretical network metrics and practical AML triage.

Appendix E

Data Validation Spot Check

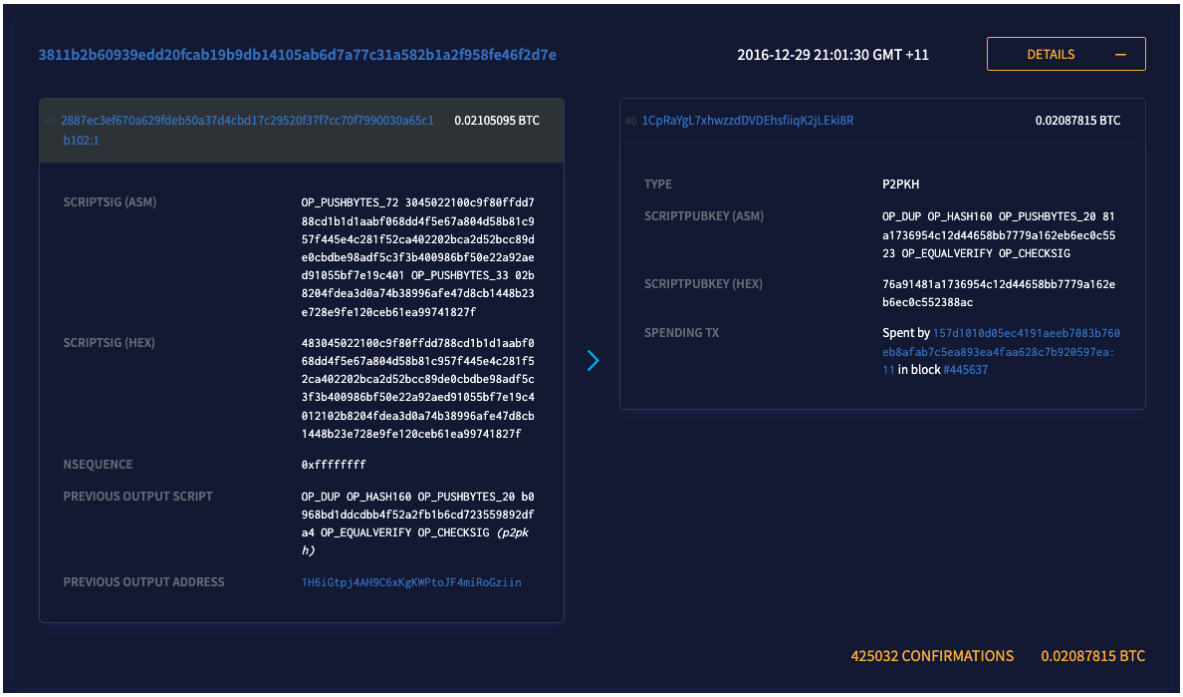


Figure E.1: Data validation spot check for Address 1H6iGtpj4AH9C6xKgKWPtoJF4miRoGziin using BTC Scan.

	index	263829
0	input_address	1H6iGtpj4AH9C6xKgKWPtoJF4miRoGziin
1	txId	139232043

(a) Address to masked transaction lookup.

	index	134009
0	txId	139232043
1	Time step	28
2	in_txs_degree	0.0000
3	out_txs_degree	1.0000
4	total_BTC	0.0209
5	fees	0.0002
6	size	192.0000
7	num_input_addresses	1.0000
8	num_output_addresses	1.0000
9	in_BTC_min	0.0211
10	in_BTC_max	0.0211
11	in_BTC_mean	0.0211
12	in_BTC_median	0.0211
13	in_BTC_total	0.0211
14	out_BTC_min	0.0209
15	out_BTC_max	0.0209
16	out_BTC_mean	0.0209
17	out_BTC_median	0.0209
18	out_BTC_total	0.0209

(b) Transaction feature summary.

Figure E.2: Data validation spot check for Address 1H6iGtpj4AH9C6xKgKWPtoJF4miRoGziin in the Elliptic++ dataset.

Appendix F

Supplementary Resources

All code, trained models, and visualisation modules developed for this research are publicly available at: https://github.com/majorpayne-2021/rmit_master_thesis

The repository contains:

- Python scripts for data preprocessing, classification, subnetwork construction, and ranking;
- BigQuery and Google Cloud Platform (GCP) notebooks for automated data ingestion and model execution;
- Output visualisations for transaction-to-transaction (txn–txn) and address-to-address (addr–addr) subnetworks; and
- Metadata and configuration files required to reproduce all experiments and figures in this thesis.

Dataset Reference: The analytical framework in this research was developed using the *Elliptic++* dataset [16], an extension of the original *Elliptic* dataset released by the MIT–IBM Watson AI Lab [46]. The Elliptic dataset provides over 200,000 Bitcoin transactions labelled as *licit*, *illicit*, or *unknown*, together with transaction features and directional edges that capture the flow of Bitcoin across the network. Elliptic++ builds upon this foundation by adding temporal transaction ordering, enriched node and edge attributes, and an updated set of labels, enabling dynamic and graph-based analysis of illicit activity over time.

The Elliptic++ dataset is publicly available at: https://github.com/git-disl/EllipticPlusPlus

Researchers seeking to reproduce or extend this work are encouraged to clone both repositories. The `rmit_master_thesis` repository contains all analytical and modelling code, while the `EllipticPlusPlus` repository provides the Bitcoin transaction tables and associated labels used for model training and evaluation.

Bibliography

[1] Chad Albrecht et al. "The use of cryptocurrencies in the money laundering process". In: *Journal of Money Laundering Control* 22.2 (2019), pp. 210–216. DOI: 10.1108/JMLC-12-2017-0074. URL: https://doi.org/10.1108/JMLC-12-2017-0074.

[2] AUSTRAC. *AML/CTF Programs*. Accessed: 2024-05-30. 2024. URL: https://www.austrac.gov.au/business/core-guidance/amlctf-programs.

[3] AUSTRAC. *AUSTRAC Takes Action to Stamp Out Financial Crime Through Cryptocurrency ATMs*. Accessed: 2025-10-09. 2024. URL: https://www.austrac.gov.au/news-and-media/media-release/austrac-takes-action-stamp-out-financial-crime-through-cryptocurrency-atms.

[4] AUSTRAC. *Money Laundering in Australia 2011*. Accessed: 2024-05-30. 2011. URL: https://www.austrac.gov.au/business/how-comply-guidance-and-resources/guidance-resources/money-laundering-australia-2011.

[5] AUSTRAC. *Transaction Monitoring*. Accessed: 2024-05-30. 2024. URL: https://www.austrac.gov.au/business/core-guidance/amlctf-programs/transaction-monitoring.

[6] AUSTRAC. *Typologies Paper: AUSTRAC Money Laundering and Terrorism Financing Indicators*. Accessed: 2025-10-08. 2024. URL: https://www.austrac.gov.au/business/how-comply-guidance-and-resources/guidance-resources/typologies-paper-austrac-money-laundering-and-terrorism-financing-indicators.

[7] Australian Securities and Investments Commission. *ASIC Wins Case Against Kraken Crypto Exchange Operator for Design and Distribution Failure*. Accessed: 2025-10-09. 2024. URL: https://www.asic.gov.au/about-asic/news-centre/find-a-media-release/2024-releases/24-186mr-asic-wins-case-against-kraken-crypto-exchange-operator-for-design-and-distribution-failure/.

[8] Claudio Bellei et al. *The Shape of Money Laundering: Subgraph Representation Learning on the Blockchain with the Elliptic2 Dataset*. arXiv preprint arXiv:2404.19109, Presented at NeurIPS 2024 Datasets and Benchmarks Track (under review). 2024. URL: https://arxiv.org/abs/2404.19109.

[9] Gianluca Bontempi, Souhaib Ben Taieb, and Yann-Aël Le Borgne. "Machine learning strategies for time series forecasting". In: *European Business Intelligence Summer School*. Springer, 2012, pp. 62–77. DOI: 10.1007/978-3-642-36318-4_3. URL: https://doi.org/10.1007/978-3-642-36318-4_3.

[10] Vitor Cerqueira, Luís Torgo, and Igor Mozetič. "Evaluating Time Series Forecasting Models: An Empirical Study on Performance Estimation Methods". In: *Machine Learning* 109.11 (2020). Code available at GitHub repository `experiments-performance_estimation`, pp. 1997–2028. DOI: 10.1007/s10994-020-05910-7. URL: https://link.springer.com/article/10.1007/s10994-020-05910-7.

[11] Chainalysis. *Money Laundering and Cryptocurrency: Trends and New Techniques for Detection and Investigation*. Tech. rep. Accessed: 2024-07-03. Chainalysis Inc., July 2024. URL: https://go.chainalysis.com/cryptocurrency-money-laundering-report.html.

[12] Chainalysis. *The 2024 Crypto Crime Report*. Tech. rep. Accessed: 2024-07-03. Chainalysis Inc., 2024. URL: https://go.chainalysis.com/crypto-crime-2024.html.

[13] Tetsuya Deguchi et al. "Hubs and Authorities in the World Trade Network Using the HITS Algorithm". In: *PLOS ONE* 9.7 (2014), e100338. DOI: 10.1371/journal.pone.0100338. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0100338.

[14] Bruno Deprez et al. "Network Analytics for Anti-Money Laundering – A Systematic Literature Review and Experimental Evaluation". In: *arXiv preprint arXiv:2405.19383* (2024). Preprint; see also HTML version at https://arxiv.org/html/2405.19383v1. URL: https://arxiv.org/abs/2405.19383.

[15] David Easley and Jon Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge, UK: Cambridge University Press, 2010. ISBN: 978-0521195331. URL: https://www.cs.cornell.edu/home/kleinber/networks-book/.

[16] Youssef Elmougy and Ling Liu. "Demystifying Fraudulent Transactions and Illicit Nodes in the Bitcoin Network for Financial Forensics". In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*. ACM, 2023, pp. 1–16. DOI: 10.1145/3580305.3599803. URL: https://doi.org/10.1145/3580305.3599803.

[17] Steven Farrugia, Joshua Ellul, and George Azzopardi. "Detection of illicit accounts over the Ethereum blockchain". In: *Expert Systems with Applications* 150 (2020), p. 113318. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2020.113318. URL: https://www.sciencedirect.com/science/article/pii/S0957417420301433.

[18] Emilio Ferrara et al. "Detecting Criminal Organizations in Mobile Phone Networks". In: *Expert Systems with Applications* 41.13 (2014), pp. 5733–5750. DOI: 10.1016/j.eswa.2014.03.024.

[19] Financial Action Task Force. *Updated Guidance for a Risk-Based Approach to Virtual Assets and Virtual Asset Service Providers*. Accessed: 2025-10-09. 2021. URL: https://www.fatf-gafi.org/en/publications/Fatfrecommendations/Guidance-rba-virtual-assets-2021.html.

[20] Andrea Fronzetti Colladon and Eugenio Remondi. "Using social network analysis to prevent money laundering". In: *Expert Systems with Applications* 67 (2017), pp. 49–58. DOI: 10.1016/j.eswa.2016.09.029.

[21] Sarah Gruber. "Trust, Identity and Disclosure: Are Bitcoin Exchanges the Next Virtual Havens for Money Laundering and Tax Evasion?" In: *Quinnipiac Law Review* 32.1 (2013). Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2312110, pp. 135–221. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2312110.

[22] Jingguang Han et al. "Artificial intelligence for anti-money laundering: a review and extension". In: *Digital Finance* 2 (2020), pp. 211–239. DOI: 10.1007/s42521-020-00023-1.

[23] Yining Hu et al. *Characterizing and Detecting Money Laundering Activities on the Bitcoin Network*. arXiv preprint arXiv:1912.12060. 2019. URL: https://arxiv.org/abs/1912.12060.

[24] Phuong Duy Huynh et al. "From Programming Bugs to Multimillion-Dollar Scams: An Analysis of Trapdoor Tokens on Uniswap". In: *Blockchain: Research and Applications* (2025). Final journal version; earlier preprint: arXiv:2309.04700. URL: https://www.sciencedirect.com/science/article/pii/S2096720925000971.

[25] Woochang Hyun, Jaehong Lee, and Bongwon Suh. "Anti-Money Laundering in Cryptocurrency via Multi-Relational Graph Neural Network". In: *Advances in Knowledge Discovery and Data Mining. PAKDD 2023. Part II*. Ed. by Hiroshi Mamitsuka et al. Vol. 13938. Lecture Notes in Computer Science. Springer, 2023, pp. 118–130. DOI: 10.1007/978-3-031-30157-4_10. URL: https://www.researchgate.net/publication/371083167_Anti-Money_Laundering_in_Cryptocurrency_via_Multi-Relational_Graph_Neural_Network.

[26] Martin Jullum et al. "Detecting money laundering transactions with machine learning". In: *Journal of Money Laundering Control* 23.1 (2020). Open Access under CC BY 4.0, pp. 173–186. DOI: 10.1108/JMLC-07-2019-0055. URL: https://doi.org/10.1108/JMLC-07-2019-0055.

[27] Md. Rezaul Karim et al. "Scalable Semi-Supervised Graph Learning Techniques for Anti Money Laundering". In: *IEEE Access* 12 (2024), pp. 50012–50029. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2024.3383784. URL: https://ieeexplore.ieee.org/document/10486886.

[28] Xiangfeng Li et al. "FlowScope: Spotting Money Laundering Based on Graphs". In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-20)*. 2020, pp. 4731–4738. DOI: 10.1609/aaai.v34i04.5906. URL: https://ojs.aaai.org/index.php/AAAI/article/view/5906.

[29] Dan Lin et al. *RiskProp: Account Risk Rating on Ethereum via De-anonymous Score and Network Propagation*. 2023. arXiv: 2301.00354 [cs.SI]. URL: https://arxiv.org/abs/2301.00354.

[30] JiaQi Liu, XueRong Li, and JiChang Dong. "A survey on network node ranking algorithms: Representative methods, extensions, and applications". In: *Science China Technological Sciences* 64.3 (2021), pp. 451–461. DOI: 10.1007/s11431-020-1683-2.

[31] Joana Lorenz et al. “Machine Learning Methods to Detect Money Laundering in the Bitcoin Blockchain in the Presence of Label Scarcity”. In: *Proceedings of the ACM International Conference on AI in Finance (ICAIF '20)*. Also available as arXiv preprint https://arxiv.org/abs/2005.14635. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–8. DOI: 10.1145/3383455.3422549. URL: https://arxiv.org/abs/2005.14635.

[32] Malte Möser, Rainer Böhme, and Dominic Breuker. “Towards Risk Scoring of Bitcoin Transactions”. In: *Financial Cryptography and Data Security – FC 2014 Workshops*. Vol. 8438. Lecture Notes in Computer Science. Presented at FC 2014 Workshops, Barbados. Springer, Berlin, Heidelberg, 2014, pp. 16–32. DOI: 10.1007/978-3-662-44774-1_2. URL: https://www.researchgate.net/publication/285624825_Towards_Risk_Scoring_of_Bitcoin_Transactions.

[33] Satoshi Nakamoto. *Bitcoin: A Peer-to-Peer Electronic Cash System*. White paper. 2008. URL: https://bitcoin.org/bitcoin.pdf.

[34] Shiyu Ouyang et al. “Bitcoin Money Laundering Detection via Subgraph Contrastive Learning”. In: *Entropy* 26.3 (2024), p. 211. DOI: 10.3390/e26030211. URL: https://www.mdpi.com/1099-4300/26/3/211.

[35] Lawrence Page et al. *The PageRank Citation Ranking: Bringing Order to the Web*. Tech. rep. Technical Report. Stanford InfoLab, 1999. URL: http://ilpubs.stanford.edu:8090/422/.

[36] Silivanxay Phetsouvanh, Frédérique Oggier, and Anwitaman Datta. “EGRET: Extortion Graph Exploration Techniques in the Bitcoin Network”. In: *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2018, pp. 244–251. DOI: 10.1109/ICDMW.2018.00043. URL: https://dr.ntu.edu.sg/entities/publication/840aa161-eea9-4670-b670-54e3e4766e4e.

[37] Yasaman Samadi, Hai Dong, and Xiaoyu Xia. “MPOCryptoML: Multi-Pattern Based Off-Chain Crypto Money Laundering Detection”. In: *arXiv preprint arXiv:2508.12641* (2025). Manuscript submitted to IEEE Transactions on Information Forensics and Security. arXiv: 2508.12641 [cs.CR]. URL: https://arxiv.org/abs/2508.12641.

[38] Alex Sangers et al. “Secure Multiparty PageRank Algorithm for Collaborative Fraud Detection”. In: *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2018, pp. 244–251. DOI: 10.1109/ICDMW.2018.00043. URL: https://eprint.iacr.org/2018/917.pdf.

[39] Abdul Khalique Shaikh and Amril Nazir. “A novel dynamic approach to identifying suspicious customers in money transactions”. In: *International Journal of Business Intelligence and Data Mining* 17.2 (2020), pp. 143–158. DOI: 10.1504/IJBIDM.2020.108762.

[40] Abdul Khalique Shaikh, Malik Al-Shamli, and Amril Nazir. “Designing a Relational Model to Identify Relationships Between Suspicious Customers in Anti-Money Laundering (AML) Using Social Network Analysis (SNA)”. In: *Journal of Big Data* 8.1 (2021), p. 20. DOI: 10.1186/s40537-021-00411-3. URL: https://doi.org/10.1186/s40537-021-00411-3.

[41] Ítalo Della Garza Silva, Luiz Henrique Andrade Correia, and Erick Galani Maziero. "Graph Neural Networks Applied to Money Laundering Detection in Intelligent Information Systems". In: *Proceedings of the XIX Brazilian Symposium on Information Systems (SBSI '23)*. Maceió, Brazil: ACM, 2023, pp. 252–259. ISBN: 979-8-4007-0759-9. DOI: 10.1145/3592813.3592912. URL: https://www.researchgate.net/publication/371880273_Graph_Neural_Networks_Applied_to_Money_Laundering_Detection_in_Intelligent_Information_Systems.

[42] Malcolm K. Sparrow. "The Application of Network Analysis to Criminal Intelligence: An Assessment of the Prospects". In: *Social Networks* 13.3 (1991), pp. 251–274. DOI: 10.1016/0378-8733(91)90008-H.

[43] The Australian. *Government Reveals New Laws to Govern Crypto Sector Applying Existing AFSL Rules*. Accessed: 2025-10-09. 2024. URL: https://www.theaustralian.com.au/business/government-reveals-new-laws-to-govern-crypto-sector-applying-existing-afsl-rules/news-story/24b7b122a4bec27f6bc2e9e8ae006ff2.

[44] U.S. Department of Justice. *Binance and CEO Plead Guilty to Federal Charges in $4B Resolution*. Accessed: 2025-10-09. 2024. URL: https://www.justice.gov/archives/opa/pr/binance-and-ceo-plead-guilty-federal-charges-4b-resolution.

[45] United Nations Office on Drugs and Crime. *Hierarchical Model of Organised Crime*. Accessed: 2025-10-10. 2020. URL: https://sherloc.unodc.org/cld/en/education/tertiary/organized-crime/module-7/key-issues/hierarchical-model.html.

[46] Mark Weber et al. *Anti-Money Laundering in Bitcoin: Experimenting with Graph Convolutional Networks for Financial Forensics*. arXiv preprint. 2019. DOI: 10.48550/arXiv.1908.02591. arXiv: 1908.02591 [cs.SI]. URL: http://arxiv.org/abs/1908.02591.

[47] Mark Weber et al. *Scalable Graph Learning for Anti-Money Laundering: A First Look*. arXiv preprint arXiv:1812.00076. NeurIPS 2018 Workshop on Challenges and Opportunities for AI in Financial Services. 2018. URL: https://arxiv.org/abs/1812.00076.

[48] Jiajing Wu et al. "Analysis of Cryptocurrency Transactions from a Network Perspective: An Overview". In: *Journal of Network and Computer Applications* 190 (2021), p. 103139. DOI: 10.1016/j.jnca.2021.103139.

[49] Mike Wu et al. *Tutela: An Open-Source Tool for Assessing User-Privacy on Ethereum and Tornado Cash*. 2022. arXiv: 2201.06811 [cs.CR]. URL: https://arxiv.org/abs/2201.06811.

[50] Zhiying Wu et al. "TRacer: Scalable Graph-Based Transaction Tracing for Account-Based Blockchain Trading Systems". In: *IEEE Transactions on Information Forensics and Security* 18 (2023), pp. 2609–2621. DOI: 10.1109/TIFS.2023.3266162. URL: https://ieeexplore.ieee.org/document/10098630.

[51] Yuexin Xiang et al. *BABD: A Bitcoin Address Behavior Dataset for Pattern Analysis*. arXiv preprint. 2022. arXiv: 2204.05746 [cs.CR]. URL: https://arxiv.org/abs/2204.05746.

[52] Jianying Xiong and Wen Xiao. "Identification of Key Nodes in Abnormal Fund Trading Network Based on Improved PageRank Algorithm". In: *Proceedings of the International Conference on Data Processing and Applications in Management (ICDPAM 2020)*. Vol. 1774. IOP Publishing, 2021, p. 012001. DOI: 10.1088/1742-6596/1774/1/012001.