Capítulo 4

Análise Forense aplicada ao Bitcoin

Ivan da Silva Sendin (UFU), Rodrigo Sanches Miani (UFU), Pedro Henrique Resende Ribeiro (UFU)

Abstract

This chapter presents comprehensive concepts on forensic analysis applied to Bitcoin through machine learning techniques. The study begins with a theoretical overview of Bitcoin and its ecosystem, detailing the decentralized nature of its blockchain and the challenges associated with its pseudonymous transactions. It then explores methods for acquiring and processing blockchain data, highlighting fundamental statistical analyses that reveal transaction patterns and anomalies. A section is also dedicated to the application of heuristics H1 and H2, which are essential for tracing mixed transactions. Additionally, we examine the concept of OSINT to enrich blockchain data with external intelligence, providing deeper insights into suspicious activities. Finally, the chapter explores the application of supervised and unsupervised machine learning models in Bitcoin forensics. These models are evaluated for their effectiveness in detecting illicit activities, identifying suspicious entities, and improving the accuracy of forensic investigations. The findings underscore the potential of combining machine learning with traditional forensic methods to enhance the overall robustness of Bitcoin investigations.

Resumo

Este capítulo apresenta conceitos sobre análise forense de forma abrangente aplicada ao Bitcoin, através de técnicas de aprendizado de máquina. O estudo começa com uma visão teórica sobre o Bitcoin e seu ecossistema, detalhando a natureza descentralizada de sua blockchain e os desafios associados às suas transações pseudônimas. Em seguida, explora métodos para a obtenção e processamento de dados da blockchain, destacando análises estatísticas fundamentais que revelam padrões e anomalias nas transações. Também há um trecho dedicado à aplicação das heurísticas H1 e H2, essenciais para rastrear transações em mixers. Além disso, examinamos o conceito de OSINT para enriquecer os dados da blockchain com inteligência externa, proporcionando uma visão mais profunda sobre atividades suspeitas. Finalmente, o artigo explora a aplicação de modelos de

aprendizado de máquina supervisionados e não supervisionados na forense de Bitcoin. Esses modelos são avaliados quanto à sua eficácia na detecção de atividades ilícitas, identificação de entidades suspeitas e melhoria da precisão das investigações forenses. As descobertas destacam o potencial de combinar aprendizado de máquina com métodos forenses tradicionais para aumentar a robustez geral das investigações sobre Bitcoin.

4.1. Introdução

A análise forense, tradicionalmente associada à investigação de crimes e à identificação de evidências em diversos contextos, ganhou destaque no campo da criptografia e das moedas digitais, especialmente com o crescimento do uso de Bitcoin e outras criptomoedas. A análise forense aplicada ao ambiente digital tem como objetivo a identificação, preservação e interpretação de dados digitais, com o propósito de revelar atividades ilícitas, como a lavagem de dinheiro, compras ilegais e o financiamento de terrorismo. No contexto das criptomoedas, essa análise se torna desafiadora devido à natureza pseudônima e descentralizada das transações realizadas na blockchain.

A segurança e a integridade das transações de Bitcoin dependem fortemente de técnicas criptográficas avançadas. As funções de *hashing* criptográficas, por exemplo, são utilizadas para garantir que os dados armazenados na blockchain sejam íntegros. As propriedades de unidirecionalidade e resistências a colisões e segunda pré-imagem desempenham um papel crucial na construção de estruturas como a Árvore de Merkle *(Merkle Tree*), que permite a verificação eficiente e segura da integridade de grandes conjuntos de dados na blockchain. Essa estrutura é essencial para a organização dos blocos de transações e para assegurar que cada transação seja válida e imutável.

Além disso, as assinaturas digitais, como o ECDSA (*Elliptic Curve Digital Signature Algorithm*) [Johnson et al. 2001], são fundamentais para garantir a autenticidade das transações, permitindo que apenas o proprietário de uma chave privada específica possa autorizar a movimentação dos fundos associados a um endereço na blockchain. Recentemente, o algoritmo Schnorr tem sido discutido como uma alternativa mais eficiente e segura ao ECDSA, oferecendo vantagens como a agregação de assinaturas, que pode melhorar a escalabilidade e a privacidade das transações[1].

No entanto, a utilização dessas tecnologias criptográficas também tem facilitado o surgimento de atividades ilegais, como lavagem de dinheiro e transações em *Darknet Markets* (DNM), onde produtos e serviços ilícitos são comercializados. Além disso, a fuga do protocolo ou a utilização de *mixers* comprometem a rastreabilidade das transações na blockchain, criando obstáculos para as investigações forenses. Isso se torna ainda mais crítico quando essas transações são utilizadas para financiar atividades terroristas[2], representando uma ameaça à segurança global.

Diante desses desafios, é importante que os analistas forenses compreendam as tecnologias subjacentes ao Bitcoin e suas implicações, tanto para o fortalecimento das investigações quanto para a elaboração de políticas que possam diminuir o uso ilícito das criptomoedas. Além disso, é necessário observar os avisos legais associados à coleta e

[1] Schnorr Signatures for secp256k1

[2] Terrorist Financing: Hamas and Cryptocurrency Fundraising.

análise de dados na blockchain, garantindo que as investigações sejam conduzidas em conformidade com as leis e regulamentos existentes.

Neste minicurso, são discutidos os principais métodos utilizados na análise forense utilizados da criptomoeda Bitcoin. Na Seção 4.2 são apresentados os principais conceitos relacionados ao Bitcoin, suas características técnicas principais e o seu ecossistema. Esse conceitos são fundamentais para iniciar as análises. Já na Seção 4.3 são mostradas as principais formas de obtenção dos dados da Blockchain e suas vantagens e desvantagens são discutidas. Nas Seções 4.4 e 4.5 são mostradas como as primeiras análises forenses podem ser feitas, primeiramente com ferramentas estatísticas e depois com técnicas de análises específicas para Blockchain. Na Seção 4.6, mostramos como as técnicas de OSINT podem ser aplicadas na busca de atividades ilegais e na Seção 4.7 é apresentada aplicação de Aprendizado de Máquina. Por fim, a Seção 4.8 apresenta as conclusões deste minicurso e alguns dos desafios que os estudantes e profissionais da área terão que enfrentar.

Esse texto apresenta vários exemplos práticos e sugestões de atividades a serem desenvolvidas. O critério de escolha dos exemplos foi didático e operacional: acreditamos que em cada um deles uma lição seja apresentada e que eles sejam reproduzíveis em ambientes computacionais modestos. Por fim, o material complementar deste minicurso está disponível no repositório de Forense de Bitcoin do NUSEC/FACOM - Núcleo de Segurança da Faculdade de Computação da Universidade Federal de Uberlândia.

4.2. Bitcoin

O Bitcoin [Nakamoto 2009] é primeira e a principal criptomoeda existente. A sua segurança e funcionalidade dependem da existência da Blockchain: uma forma segura e descentralizada de armazenar as transações financeiras executadas pelos seus participantes. A Blockchain é mantida e atualizadas pelos mineradores, que utilizam de uma mecanismo chamado Prova de Trabalho para tal. Sem autoridade central, a corretude da Blockchain é obtida pela Teoria dos Jogos: o comportamento honesto deve gerar mais lucro que o comportamento ilícito. Para uma melhor compreensão do Bitcoin e de criptomoedas recomenda-se a leitura atenta de livros-texto como [Antonopoulos 2014] e [Narayanan et al. 2016]; o trabalho [Narayanan and Clark 2017] também é um excelente ponto de partida para o conhecimento dos diversos recursos criptográficos e tecnologicos usados nas criptomoedas e o principal site da comunidade Bitcoin `https://bitcoin.org/` é um fonte importante de detalhes e especificações técnicas.

Conhecido pela oportunidade de ganhos financeiros e especulativos, o Bitcoin atrai novos investidores a cada dia, estima-se que no mundo todo sejam mais de 500 milhões de usuários de alguma criptomoeda[3]. Outra característica que ajudou na popularização do Bitcoin é o seu "anonimato", que o torna o meio "ideal" para a prática de atividades ilegais, ou pelo menos questionáveis, como meio de pagamentos de *ramsomware,* esquemas de pirâmides, apostas e de mercado de produtos ilegais. Como veremos adiante, esta percepção de anonimato do Bitcoin é falsa e existem abordagens que podem levar a identificação dos seus usuários.

[3]Statista: Base global de usuários de criptomoedas.

Mas o Bitcoin também é adotado em algumas situações tradicionais e lícitas. Por exemplo, o Governo de El Salvador adotou a criptomoeda dentro do seu sistema financeiro [els]; existem cartões de créditos operacionalizados por meio de criptomoedas e a inserção de criptomoedas no sistema financeiros tradicional[4] ocorre em diversos países. A sua facilidade de uso e acesso permitem que o Bitcoin seja uma solução onde o sistema financeiro tradicional está em colapso[5] ou mesmo para pessoas cujo acesso aos bancos é inviável[6][7].

Dado esse contexto complexo, faz-se necessário a possibilidade de investigações com a finalidade de identificar pessoas, pagamentos e atividades ilegais praticadas com Bitcoin. Essa atividade, chamada de Análise ou Investigação Forense, já existe no meio financeiro tradicional e vem se aprimorando por décadas.

Apesar de recente, a **análise forense de Bitcoin** é bem intensa no meio acadêmico com alguns eventos e periódicos abordando o assunto [Dudani et al. 2023, Salisu et al. 2023]. Também existe uma demanda significativa para este tipo de trabalho fora da área acadêmica, algumas empresas que efetuam análises de Blockchain e desenvolvem produtos com essa finalidade também se destacam[8].

4.2.1. Funcionamento

O ponto central do Bitcoin é a sua Blockchain[9]. Como o nome sugere, blockchain é uma cadeia linear de blocos. Cada bloco agrega as informações relacionadas a um conjunto de transações. Uma transação estabelece a movimentação das moedas, bitcoins (₿), de um determinado usuário para outro. Na Figura 4.1 são mostrados três blocos da Blockchain do Bitcoin. Um ponto fundamental da blockchain é o código *hash* que faz uma referência para o bloco anterior. Este campo autentica e traz a integridade para os blocos anteriores, impedindo a sua modificação. As informações de negócio são armazenadas de forma eficiente na raiz de uma Árvore de Merkle.

Os mineradores desempenham um papel fundamental na manutenção da Blockchain. A sua função principal é validar e propagar as transações, isto é, só permitir que transações corretas propaguem pela rede e sejam escritas na blockchain.

Uma grande inovação do Bitcoin foi justamente a implementação de um sistema em que participantes com interesses conflitantes - é razoável considerar que cada participante tem interesse em aumentar a sua quantidade de moedas - pudessem chegar a um consenso em um sistema descentralizado, este processo é chamado de **Consenso de Nakamoto**. Para obter este consenso, cada minerador busca uma **prova de trabalho**: um sequencia de bits (o campo `Nonce` na Figura 4.1) que faz com que o código *hash* do bloco tenha uma certa propriedade, por exemplo ser menor que um determinado valor. Ao encontrar o `nonce` correto, tem-se um bloco válido e o minerador deve propagá-lo

[4]Bancos com Itaú e Banco do Brasil permitem o investimento em criptomoedas do forma simples, a partir dos seus aplicativos.

[5]Banking the unbanked: why emerging markets dominate crypto adoption

[6]The new wave of crypto users: migrant workers

[7]Julian Assange

[8]Chainalysis e Elliptic, por exemplo.

[9]Neste texto usaremos a convenção: blockchain para um estrutura de dados e Blockchain para e estrutura de dados do Bitcoin.

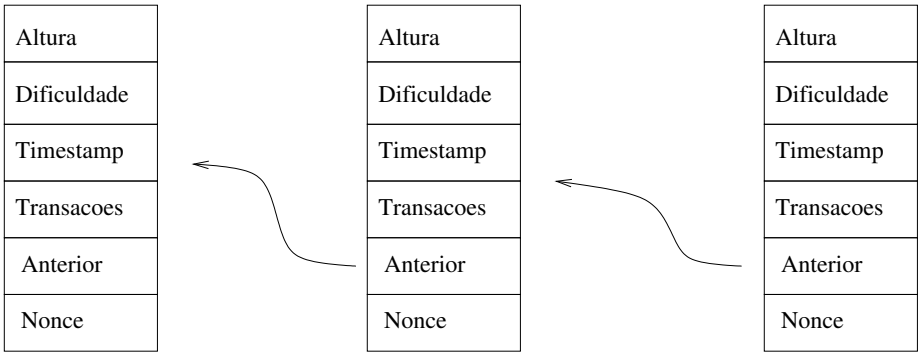

Figura 4.1: Uma blockchain simplificada. A Altura é um valor sequencial crescente. O campo `Transações` contém informações sobre as transações contidas naquele bloco. O campo `Anterior` é o código *hash* do bloco anterior. O código *hash* de cada bloco deve obedecer a uma determinada propriedade, essa característica é obtida por uma busca no campo `Nonce`.

para a rede P2P. O bloco recém minerado é propagado pela rede P2P seguindo um protocolo `gossip`: cada minerador que recebe um bloco deve validá-lo e enviar a novidade para os seus vizinhos. Ao contrário da mineração, a validação do bloco é um processo computacionalmente eficiente: com apenas uma operação de *hashing* o `nonce` é verificado. As transações contidas na raiz da Árvore de Merkle também são validadas, e agora são chamadas de **transações mineradas**. Após isso, o processo se reinicia e os mineradores começam a busca o `nonce` para o próximo bloco que será formado com novas transações que estão circulando na rede P2P. A dificuldade do processo de mineração é ajustada dinamicamente para que em média um bloco seja encontrado a cada 10 minutos, independente do número de mineradores participantes do processo.

A busca feita pelos mineradores é um processo estocástico e funciona como um sorteio: a chance de ganhar o sorteio é proporcional ao poder computacional empregado na busca pelo `nonce`. Ganhar este sorteio poder ser visto como uma oportunidade de publicar um novo bloco que será aceito pelos demais mineradores se estiver correto. Espera-se que o minerador publique apenas blocos corretos: caso ele encontre o `nonce` correto mas use transações ilícitas na construção do bloco, esse fato será percebido pelos demais mineradores e os bloco será descartado e o investimento computacional na busca do `nonce` será perdido.

Os usuários do Bitcoin possuem **endereços** que são basicamente chaves públicas. É comum, e desejável, que uma mesma entidade possua vários endereços diferentes. Para transferir um determinado valor em bitcoin, o usuário deve criar uma transação e assinar com a chave privada correspondente ao endereço de origem dos bitcoins. A maioria dos endereços são codificados usando `Base58`: 58 símbolos [10] são utilizados para codificar uma informação binária. O gerenciamento destes endereços é feito por programas chamados de **carteiras**. Usando a biblioteca Python bitcoinlib é possível implementar uma carteira simples com poucas linhas de código. Na Listagem 4.1 é mostrado um script para

[10] Os 58 símbolos que compõem a Base58 são os dígitos de 1 a 9 - sem o digito 0; as letras maiúsculas, a exceção de I e O; e as letras minúsculas, a exceção do l. Essa escolha peculiar foi feita para evitar dubiedades na leitura e digitação dos endereços.

a criação de um endereço Bitcoin. Na linha 5, a entropia do `/dev/urandom` é usada para gerar os bits da chave privada, que é apresentada na forma de uma *passphrase*. Na linha 8, as informações geradas são armazenadas na base SQL local mantida pela biblioteca utilizando o identificador `SBSEG2024`. Após isso o endereço bitcoin é mostrado na tela, esse endereço é usado para "receber" os valores e, por último, a chave privada é mostrada no formato padronizado para importação por outras carteiras (WIF, *Wallet Inport Format*). A *passphrase* foi pensada para cenários não digitais, ela deve ser anotada em um pedaço de papel, que deve ser mantido de forma segura, ou até mesmo memorizada. Ela tem a propriedade permitir correções, sem comprometer a sua segurança, por exemplo, se ao digitar a *passphrase* alguem digitar `squirel`, esse erro será detectado. A utilização dos bitcoins de um endereço depende do conhecimento da chave privada, Neste exemplo, caso o usuário não anote a *passphrase* e o disco seja corrompido, os bitcoins do endereço serão inutilizados[11]. O endereço produzido pela execução do código foi o 12*yPhuv*7*yJCfepq*7*kZQm*5*ej*93*Fes*5*do*64*n*, muitas vezes, em um texto, o endereço é apresentado de forma compacta: 12*yP...o*64*n*.

Existem outros tipos de endereços usados no Bitcoin. Por exemplo alguns endereços usam `scripts` para validar transações, e outros que implementam variações que permitem o melhor aproveitamento da blockchain. Essas tipificações são percebidas no prefixo do endereço e são mostradas na Tabela 4.1.

```
1  from bitcoinlib.wallets import Wallet
2  from bitcoinlib.mnemonic import Mnemonic
3
4  passphrase = Mnemonic().generate()
5  print("A sua passphrase e:", passphrase)
6  #A sua passphrase e: glimpse flavor enroll peanut enough under final
       reason squirrel twenty sport slender
7  w = Wallet.create('SBSEG2024', keys=passphrase, network='bitcoin')
8
9  print(w.get_key().address)
10 #12yPhuv7yJCfepq7kZQm5ej93Fes5do64n
11
12 print(w.get_key().wif)
13 # xprvA3n4M5SJLFo41eHbv6mMud26vNRKewz6jVFizSHiyKoHyScixABJhBYceudbk1K
14 # imGnLw7YcRwG8qLx3BW6tbmvtjXmMLTwAT8wpM819pH9
```

Listagem 4.1: Exemplo de uso da biblioteca bitcoinlib para criação de endereço.

De um modo simplificado, a **transação** no Bitcoin pode ser vista como uma estrutura formada por dois vetores: o vetor de entrada da transação e o vetor de saída. A entrada de uma transação é composta por referências a saídas não gastas de transações anteriores (`UTxO`, do inglês *Unspent Transaction Output*). A saída da transação é um lista ordenada de tuplas endereços e valores. As transações formam um grafo de movimentação de bitcoins. Cada transação é identificada de forma única pelo *hash* calculado sobre o seu conteúdo.

A primeira transação de cada bloco é chamada de `Coinbase` e é especial: ela não possui nenhum item na entrada, apenas itens na saída. Essa transação é criada pelo minerador, que coloca endereços sob seu controle na saída da transação, emitindo novas

[11] É famoso o caso em que meio bilhão de dólares foram jogados no lixo.

Prefixo	Tipo	Descrição
1	Pay-to-Public-Key-Hash (P2PKH)	Utiliza asssinaturas digitais para validar as transações
3	Pay-to-Script-Hash (P2SH)	O resultado da execução do script valida ou não uma transação
bc1	Segregated Witness (SegWit)	Modifica a forma como os dados são armazenados na blockchain.
bc1p	Taproot (P2TR)	Utiliza assinaturas de Schnorr amplia o uso smart contract no Bitcoin

Tabela 4.1: Tipos de endereços Bitcoin

moedas. Este processo existe justamente para remunerar os mineradores pelo trabalho de manutenção da Blockchain. Uma outra forma de remuneração é paga por cada emissor de transação. Para uma determinada transação, o valor da **taxa** paga é a diferença entre a soma dos valores do `UTxOs` de entrada e a soma dos `UTxOs` de saída. Na Figura 4.2 é apresentado um exemplo onde o usuário que controla os endereços *bcd* e *cde* decide pagar ₿22 para o endereço *xyz*. Ao consultar a lista de `UTxOs` da Blockchain, ele identifica que a saída de posição 1 da transação identificada por 0x1234 e a saída de posição 0 da transação 0x234 não são gastas - `UTxOs`. O `UTxO` $(0x1234, 1)$ é de ₿20, que junto com $(0x2345, 0)$ contabilizam ₿30, suficiente para o pagamento desejado. Uma transação é criada, com referências aos `UTxOs` selecionados na entrada. Como o valor da entrada excede o valor a ser pago um novo endereço é criado (*trc*, no exemplo) para receber o troco. Ainda, podemos observar que ₿1 é destinado como taxa ao minerador. Antes da transação ser enviada para a rede P2P ela precisa ser assinada por todos os endereços que estão "gastando", isto é, estão na entrada. Uma vez que a transação 0x3456 for enviada à rede P2P, suas entradas serão removidas da lista de `UTxOs`. A criação do endereço de troco não é obrigatória: ao criar a transação, o usuário poderia escolher algum endereço já usada para receber o troco. A taxa paga também é de escolha do criador da transação. Os valores registrados na Blockchain são em Satoshi: uma unidade de Satoshi vale ₿10^{-8}.

Uma consequencia direta do modelo `UTxO` criado pelo Bitcoin é que o conceito de saldo é bem diferente do conceito de saldo de uma conta bancária, ou até mesmo de algumas outras criptmoedas. O saldo de uma determinada entidade esta distribuído em diversos endereços; e o saldo de cada endereço está distribuído em diversas `UTxOs` e precisa ser calculado para ser usado.

Ainda, a identificação dos usuários por meio de chaves públicas dá ao Bitcoin uma de suas principais - e mais mal compreendida - características: o pseudo-anonimato. O Bitcoin - e a maioria das criptomoedas[12] - é anônimo pois não existe nome (nome próprio, RG, CPF ou outra forma de indentificação forte) na Blockchain, mas todas as transações e seus dados estão explícitos na Blockchain e passíveis de serem analisadas, por isso o termo pseudo.

As análises feitas sobre o Bitcoin são análises onde o fluxo de dinheiro está dispo-

[12] Algumas moedas são chamadas de *privacy coins* e utilizam mecanismos criptograficos para esconder de fato as informações transitadas na Blockchain.

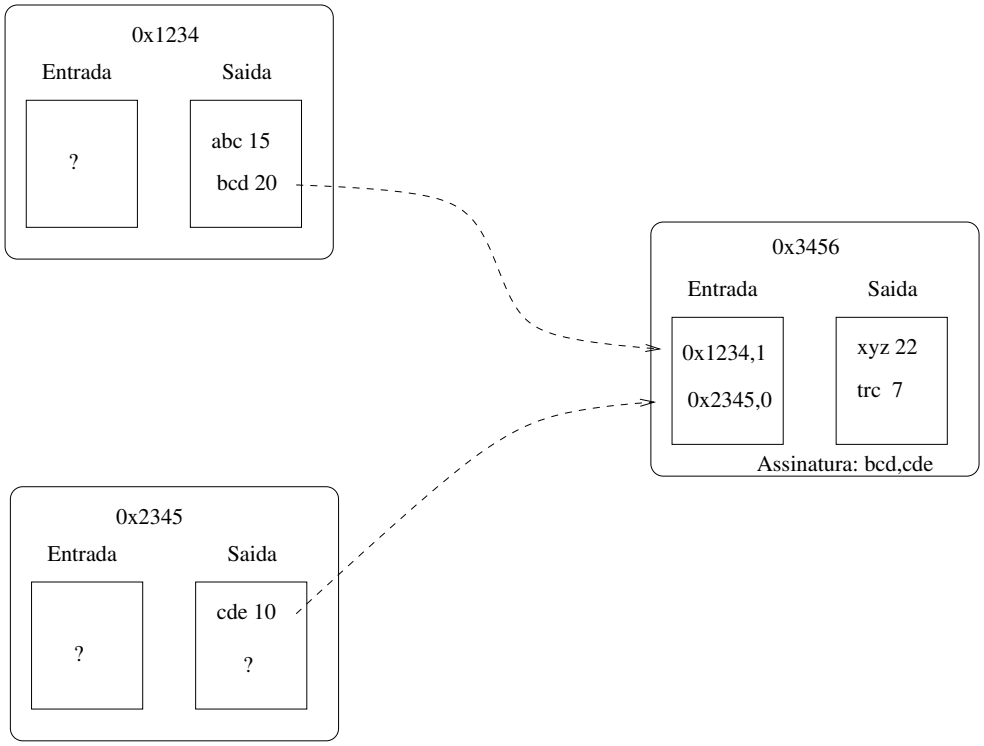

Figura 4.2: Exemplo de Transação. O endereço xyz recebe 22 Bitcoins. Como a soma dos valores das entradas é maior que o valor a ser pago, um endereço (trc) é criado para receber o troco. A diferença entre a entrada e a saída é a taxa, recolhida pelo minerador.

nível de forma explícita na blockchain, mas as entidades participantes destas transações muitas vezes estão anônimas e, justamente, grande parte do esforço nas análises é identificar - ou ao menos produzir inteligência - sobre as entidades participantes.

4.2.2. Ecossistema

Somente o entendimento do protocolo Bitcoin não é suficiente para a compreensão do seu funcionamento no mundo real. Nesta seção apresentamos alguns dos componentes do chamado *Ecossistema do Bitcoin* especialmente relevantes para análise forense.

Conforme dito anteriormente, **os mineradores** desempenham um papel fundamental na manutenção da Blockchain e são remunerados por esse serviço. Considerando o protocolo do Bitcoin, pode-se imaginar que cada um dos participantes poderia gastar algumas horas de sua CPU de uso pessoal para fazer o papel de minerador, concretizando a frase "uma CPU um voto". Atualmente este cenário é impensável; a mineração é executada por grandes empresas que abandonaram o uso de CPUs e agora usam equipamentos especificamente projetados[13] para executar mineração de criptomoedas. Os mineradores ser organizam em *pools*, onde um coordenador executa as tarefas de manutenção da blockchain propriamente dita e se comunica com os demais participantes que executam a tarefa de busca pelo `nonce`. Por isso, os mineradores não são contabilizados pela sua quantidade, mas sim pela fração do poder computacional que eles possuem.

[13] Veja, por exemplo, a empresa Bitman.

O consenso obtido pelo mineradores não é obrigatoriamente a verdade, mas um acordo entre a maioria dos participantes. Caso a maioria dos mineradores decida ignorar o protocolo e, por exemplo, desconsiderar uma determinada transação, essa prática será aceita e a corretude da Blockchain ficará comprometida, esse ataque hipotético é conhecido como **Ataque do 51%** [Aponte-Novoa et al. 2021]. Desta forma a centralização que a organização em *pools* de mineradores provoca é contraditória com a descentralização inicialmente proposta e pode comprometer a segurança do Bitcoin.

Outro ponto crítico no processo de mineração é que o desvio do protocolo pode trazer benefícios aos mineradores. Em [Eyal and Sirer 2014] é proposto uma modificação no processo de mineração, onde o minerador ao minerar o bloco de altura n não envia este bloco para a rede P2P e fica trabalhando sozinho na busca do bloco de altura $n+1$, obtendo vantagem sobre os demais mineradores. Esta abordagem é chamada de **Mineração Egoísta**. Alguns métodos de identificação desta prática já foram propostos e aplicados na blockchain mas ainda sem comprovação de que ela já foi usada na prática [Li et al. 2020a, Li et al. 2020b].

As ***exchanges*** podem ser vistas como casas de câmbio para criptomoedas: os seus clientes conseguem fazer depósitos e saques usando o sistema financeiro tradicional utilizando moeda corrente do seu país; e utilizar o sistema das *exchanges* para compra e venda de criptomoedas. Usualmente as *exchanges* possuem várias opções de criptomoedas e os seus usuários fazem operações de compra e venda especulativas. Um dos objetivos das *exchanges* é prover facilidades aos seus clientes, que não precisam saber o conceito de chave pública e privada para executar operações na blockchain, assim, um cliente pode usar uma interface *web* para enviar bitcoins do seu endereço para outro. De acordo com [Team 2024] diversos países exigem as práticas de *Anti Money Laundering* (AML) e *Know Your Customer* (KYC) que obrigam as *exchanges* identificar devidamente os seus clientes com apresentação de documentos oficiais, fotos e comprovantes de endereço. Essas práticas permitem que as autoridades legais de um país possam investigar pessoas suspeitas de praticarem atividades ilegais[14].

Uma característica importante é que os clientes das *exchanges* não possuem bitcoins de fato: as *exchanges* mantém a custódia dos bitcoins em endereços controlados por elas. Para alguém que observa os endereços dos clientes na blockchain, constatará que eles possuem saldo zero a maior parte do tempo. Quando um usuário de *exchange* deseja enviar bitcoins para algum endereço, a *exchange* faz a transferência para o endereço do cliente e posteriormente faz uma segunda transferência para o endereço determinado pelo cliente. Esse controle feito pela maioria das *exchanges* sobre os bitcoins dos clientes causa duas consequências importantes:

1. Pegadas na Blockchain: esse padrão de transferência dupla e endereços sem saldo deixa pegadas na blockchain que podem ajudar na identificação de endereços de *exchanges* [Sendin 2018];

2. Risco de perda de bitcoins: como o controle fica centralizado nas *exchanges*, problemas técnicos ou má fé podem causar a perda de valores de milhares de clientes

[14]Em 2019, a Interpol, utilizando de informações fornecidas por diversas *exchanges* prendeu 337 pessoas em 38 países por participarem de uma rede de vídeos pornográficos ilegais

de forma rápida e irreversível. Existem inúmeros casos de perdas de bitcoins relacionados a *exchanges*. Um evento de grande repercussão foi o da *exchange* MtGox, que alega ter perdido o equivalente a 480 milhões de dólares com bitcoins roubados, o que teria causado a sua falência [Ishikawa 2017].

Outro ponto de contato entre o Bitcoin e o sistema financeiro tradicional são as **ATMs**: algumas empresas disponibilizam máquinas automáticas para sacar dinheiro em moeda corrente. O processo operacional destas máquinas é bem simples: o usuário utiliza um dispositivo qualquer - eg. *smart phone* - para enviar bitcoins para um endereço informado via QR Code pela ATM, que após alguns instantes, disponibiliza as cédulas. Estas facilidades impedem os controles KYC e AML que as *exchanges* são obrigadas a implementar, tornando as ATMs um mecanismo alternativo para lavagem de dinheiro [Noll 2023, Hyman 2015].

Os **mercados** são importantes participantes no ecossistema das criptomoedas: um dos atrativos das criptomoedas é justamente a sua facilidade de uso. A primeira compra feita com bitcoin foi de duas pizzas e ocorreu no mês de Maio de 2010. O compra foi negociada no fórum bitcointalk e ₿10.000 foram transferidos como pagamento de US$41, o valor das pizzas e ₿0.99 foram pagos de taxa.

É normal esperar que os mercados sejam menos regulamentados que as *exchanges*, devendo atrair o dinheiro obtido de forma ilegal. Uma presença relevante nesta categoria são os mercado de produtos ilegais (**DNM**, do inglês *DarkNet Marketplace*), que oferecem uma gama variada de produtos ilegais. O acesso às plataformas DNM é feito de forma segura, geralmente usando a rede Tor[15]. Um exemplo notório de DNM é o Silk Road [Christin 2012]. Este mercado iniciou as suas atividades em 2011 e movimentou milhares de dólares em criptomoedas até que em 2013 foi fechado e seu responsável foi condenado à prisão perpétua. Além das questões óbvias de privacidade, a operação de um DNM exige certos cuidados que um mercado tradicional não precisa ter: os impasses não podem ser resolvidos no sistema judiciário. Existem duas abordagens principais para o funcionamento dos DNMs:

Custódia O cliente faz o pagamento para o DNM, para uma carteira de custódia. Após receber o produto o DNM recolhe a sua comissão e o restante do valor é transferido para o vendedor. Caso o comprador indique que não recebeu o produto, o valor é devolvido para o cliente;

Reputação É fácil ver que o sistema de custódia não é infalível, por exemplo, o cliente pode receber o produto comprado e mesmo assim pedir o dinheiro de volta. Alguns DNMs operam por um sistema de reputação, onde os vendedores são avaliados pelos compradores.

O DNMs são foco de de atenção na análise de atividades ilegais e objeto de intensa pesquisa, pela sua própria natureza ilegal, mas também pelos produtos vendidos, que podem ser usados em outras atividades ilegais mais danosas, como senhas, *malwares* ou até mesmo armas restritas.

[15] https://www.torproject.org

A privacidade oferecida pelo pseudo-anonimato do Bitcoin é bem limitada, para aumentar o anonimato dos participantes, ainda no início do uso da criptomoeda, foi desenvolvido o conceito de `CoinJoin` [16]: *n* pessoas interessadas em aumentar a sua privacidade na blockchain se unem para fazer uma transação com *n* `UTxOs` e *n* itens na saída da transação. Com os valores de entrada e saídas iguais e com a ordem de saída sendo uma permutação aleatória da ordem de entrada, quem observar uma transação `CoinJoin` na Blockchain não vai saber qual `UTxO` - portanto qual usuário - está relacionada com qual saída, apagando as pegadas que as transações normalmente produzem. A Figura 4.3 ilustra um exemplo de transação `CoinJoin`. Os operadores deste tipo de serviço são chamados de ***mixers***, que coletam as informações dos participantes das transações usando a rede Tor ou VPN e fazem o embaralhamento. Dois grandes *mixers* disponíveis são o Wasabi e o Whirlpool, e justamento pelo poder deste tipo de operação, o uso de *mixer* é ilegal em alguns países e alvo de interesse de pesquisas [Stütz et al. 2022, Tironsakkul et al. 2022].

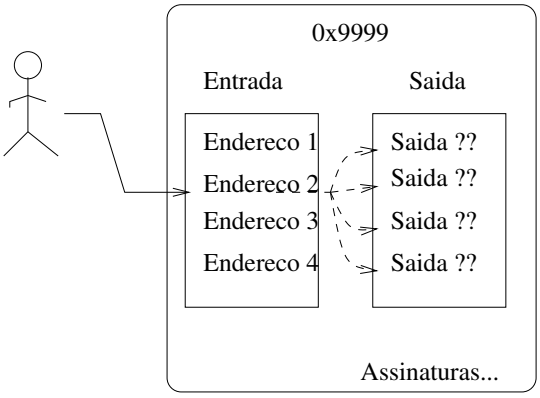

Figura 4.3: Neste exemplo de transação `CoinJoin` destacamos o usuário do *Endereco2* que ao participar de uma transação do tipo `CoinJoin` elimina os passos que normalmente seriam deixados na blockchain. Como a ordem das saídas foi embaralhada e os valores são iguais, a observação da blockchain não permite saber quem pagou(ou recebeu) de quem.

Enquanto apostas são rigorosamente regulamentadas ou proibidas em muitos países, o pseudoanonimato e a inexistência de fronteiras do Bitcoin o tornam um meio atraente para operação de **plataformas de apostas**. É bem sabido que as apostas são utilizadas para lavagem de dinheiro. Estudos recentes [Dalal et al. 2021, Ampel et al. 2023] indicam que as plataformas de apostas são usadas para remover o caminho deixados pelas transações, de forma similar ao efeito obtido pelo uso dos *mixers*.

Apesar de não fazer parte do ecossistema do Bitcoin, os resgates relacionados a *ransomware* [Dalal et al. , Ampel et al. 2023] são relevantes nos estudos de forense pois esse tipo de *malware* possui uma relação de dependência das criptomoedas e os seus endereços fazem movimentações para dificultar a análise.

Além destes participantes, qualquer pessoa com interesse em investigações envolvendo criptomoedas deve acompanhar as empresas de **Análise de Blockchain**. Essas empresas fazem constantes análises da blockchain de inúmeras criptomoedas, produzindo

[16] CoinJoin: Bitcoin privacy for the real world

relatórios; bases de dados que classificam endereços, entidades e transações; software e APIs de análise de Blockchain e consultorias para o sistema de justiça de diversos países. Algumas empresas de destaque nesta área são: Chainalisys, Elliptic, TRM Labs e Arkham Intelligence. Ainda, a plataforma WalletExplorer implementa um buscador capaz de classificar um endereço ou transação em categorias como *exchanges*, apostas, *mixers* entre outros. O WalletExplorer é amplamente utilizado na análise de blockchain e citado em diversos trabalhos acadêmicos. Outra plataforma de interesse é a ChainAbuse que recebe e classifica atividades ilegais praticadas com criptomoedas.

4.2.3. Prática

Utilizando a plataforma Blockchain.com é possível consultar a blockchain da maioria das criptomoedas utilizando uma interface web amigável. Utilize esta plataforma para:

P1 Consultar a transação da primeira compra feita com Bitcoin.

P2 As transações de `CoinJoin` podem ser mais ou menos complexas, de acordo com o serviço de *mixer* utilizado. Consulte duas transações apontadas com transações de `CoinJoin` por [Stütz et al. 2022]: transação 1 e transação 2.

4.3. Obtenção dos Dados

Um passo primordial na análise forense do Bitcoin é a obtenção da blockchain propriamente dita. A rede de comunicação do Bitcoin é uma rede P2P não permissionada, isto é, qualquer pessoa com acesso à Internet pode consultar a rede, criar um nó local do Bitcoin ou até mesmo validar as transações e minerar novos blocos. Consequentemente, os dados da Blockchain são públicos e podem ser obtidos com facilidade. Apesar disso, temos que considerar que a quantidade de dados é grande e a sua organização pode ser complexa. Por exemplo, a entrada de uma transação não contém endereços, mas referências a endereços de transações anteriores, então para interpretar uma transação e obter os endereços de entrada é necessário consultar diversas transações anteriores. Além disso, informações que estão fora da blockchain - como por exemplo o valor em dolar de uma transação - muitas vezes precisam ser obtidas de outras fontes. As Seções 4.3.1, 4.3.2 e 4.3.3 apresentam algumas formas comuns de acessar os dados da Blockchain.

4.3.1. Acesso Direto

Os dados sobre o Bitcoin podem ser obtidos diretamente da sua Blockchain utilizando um dos vários clientes de Bitcoin disponíveis. Uma vez que o cliente Bitcoin esteja instalado, é possível acessar qualquer nó da rede P2P para fazer consultas. Na prática, por questões de desempenho é importante rodar um nó Bitcoin localmente para que as consultas sejam resolvidas localmente de forma mais rápida.

O cliente oficial do Bitcoin está disponível para diversas plataformas e pode ser obtido aqui. Ao executar o binário `bitcoind`, o download da blockchain terá inicio. O processo de download pode demorar alguns dias e consumir algumas centenas de gigabytes: durante o download a consistência dos blocos é verificada gerando um número alto de consulta aos dados recém obtidos, assim é altamente recomendável que o armazenamento seja feito em SSD.

Uma vez que o nó do Bitcoin esteja rodando, a Blockchain poderá ser consultada usando o binário `bitcoin-cli`, distribuído no mesmo pacote, ou por `scripts` utilizando explicitamente RPC. Nas Listagens 4.2 e 4.3 são apresentados exemplos de acesso ao nó local, que para afeito de demonstração foi iniciado para ser acessado pelo usuário `forensics` com a senha `123456`. Em ambas listagens, a tarefa executada é a mesma: o procedimento `getblockhash` é invocado com o parâmetro 5000 ao `bitcoind`, que esta aberto na porta TCP 8332. Este, por sua vez, consulta a base de dados e retorna o código *hash* do bloco solicitado.

```
yoda@dagobah:~$bitcoin-cli -rpcuser=forensics -rpcpassword=123456
   getblockhash 5000
000000004d78d2a8a93a1d20a24d721268690bebd2b51f7e80657d57e226eef9
```

Listagem 4.2: Uso do bitcoin-cli para acessar um nó local do Bitcoin. Neste exemplo o código hash do bloco de altura 5000 é obtido através do método `getblockchash` com o parâmetro 5000

```
import requests
import json

headers = {'content-type': 'text/plain'}
datastr='{"jsonrpc": "1.0", "id": "curltest", "method": "getblockhash",
    "params": [5000]}'
response = requests.post('http://127.0.0.1:8332/', headers=headers,
   data=datastr, auth=('forensics', '123456'))
r = response.json()
print(r['result'])
#000000004d78d2a8a93a1d20a24d721268690bebd2b51f7e80657d57e226eef9
```

Listagem 4.3: Exemplo de script Python para acessar bitcoind

Ainda, é possível acessar a Blockchain armazenada localmente de forma direta, sem a necessidade de consultas ao `bitcoind`. A biblioteca Bitcoin Blockchain Parser, desenvolvida em Python, disponibiliza um `parser` para os dados armazenados. Além de facilitar alguns aspectos na implementação, pois oferece classes que encapsulam algumas complexidades das consultas, esta abordagem é computacionalmente mais eficiente, pois elimina o *overhead* do uso do RPC. Na Listagem 4.4 é mostrado um código para calcular o saldo das carteiras ao final do primeiro ano do Bitcoin. Para determinar o período de um ano é feito uma conta simples de seis blocos por hora, 24 horas por dia e 365 dias por ano, para analises que exijam maior precisão é possível usar a propriedade `timestamp` do bloco.

```
import os
from blockchain_parser.blockchain import Blockchain

PATH = '/media/ssd/FULLNODE/blocks/'
INDEXPATH = '/media/ssd/FULLNODE/blocks/index'

def updateBalance(ad,a,v):
    if a in ad:
        ad[a] = ad[a]+v
        return
    ad[a] = v

```

```
def getOutputData(txo,h,i):
    o = txo[h][i]
    return (o.addresses,o.value)

txo = {}
a = {}

blockchain = Blockchain(os.path.expanduser(PATH))

for block in blockchain.get_ordered_blocks(os.path.expanduser(INDEXPATH
    ), end=6*24*365):
    print("height=%d block=%s" % (block.height, block.hash))
    for tx in block.transactions:
        txo[tx.hash] = tx.outputs

        if not tx.is_coinbase():
            for i in tx.inputs:
                (endereco,valor) =  getOutputData(txo,i.
    transaction_hash,i.transaction_index)
                if len(endereco)==1:
                    updateBalance(a,endereco[0].address,-valor)

        for no, output in enumerate(tx.outputs):
            if (len(output.addresses)==1):
                updateBalance(a,output.addresses[0].address,output.
    value)

print(sorted(a.items(), key=lambda item: item[1],reverse=True)[:10])
```

Listagem 4.4: Exemplo de uso da biblioteca Bitcoin Blockchain Parser

4.3.2. Navegadores de Blockchain

Os navegadores de blockchain ou *block explorers* são serviços oferecidos por algumas empresas que visam facilitar o acesso à Blockchain. O acesso pode ser feito usando o navegador ou por meio de uma API própria. Muitas vezes as informações oferecidas por estes serviços são obtidas de forma mais simples. Por exemplo, para determinar os endereços de entrada de uma transação, algumas consultas devem ser feitas para interpretar os `UTxOs`. Quando os *block explorers* são utilizados, este tipo de informação é obtida diretamente.

Um exemplo de empresa que provê este tipo de informação é a Blockchain.com[17] que oferece uma API para acessar os dados da blockchain. Por exemplo, ao acessar a `url` `https://blockchain.info/latestblock` será obtido informações básicas sobre o último bloco minerado. Já na Listagem 4.5 é apresentado um exemplo que faz duas requisições ao provedor: a primeira requisição obtém o código *hash* do último bloco minerado, com esta informação uma segunda requisição é feita e as transações contidas no bloco são obtidas. Um ponto negativo desta abordagem é que este tipo de serviço é pago, um mesmo endereço de IP pode fazer apenas algumas poucas consultas por minuto sem o cadastro na plataforma.

[17] A mesma empresa mantém os domínios `blockchain.info` e `blockchain.com` e não é um *site* oficial do Bitcoin.

```
import requests
import json

url = 'https://blockchain.info/latestblock'
resp = requests.get(url=url)
data = resp.json()
print (data)
print (data['hash'])

url2 = 'https://blockchain.info/rawblock/'+data['hash']

resp = requests.get(url=url2)
data = resp.json()

for tx in data['tx']:
    print (tx)
```

Listagem 4.5: Acesso às transações do último bloco minerado

4.3.3. Dumps

Quase que como uma extensão dos navegadores de blocos, também existem empresas que oferecem `dumps` dos dados da blockchain, muitas vez organizados por data e por tipo de dado. Por exemplo, a BlockChair disponibiliza os dados organizados em arquivos tipo TSV[18] e acompanhados de informações além das existentes na Blockchain, como o valor em dólares das movimentações. A utilização dos `dumps` é interessante pois obtém-se os dados necessários de forma rápida, mas sem precisar obter a Blockchain inteira.

4.3.4. Prática

Como visto na Seção 4.2 os mineradores desempenham um papel fundamental na manutenção da Blockchain, estabelecendo a "verdade" por um sistema de votação que considera o poder computacional de cada minerador. Um problema desta abordagem é que a grosso modo quando a maioria do poder computacional estabelecer um "fato", ele será aceito pelo sistema. Esse comportamento é conhecido como Ataque do 51%. Por exemplo, na Figura 4.2, se a maioria com mais da metade do poder computacional concordar em excluir uma determinada transação isso pode ocorrer. Por isso é importante que haja um distribuição do poder computacional entre os mineradores. É possível inferir com precisão o poder computacional dos mineradores observando os blocos minerados: para uma determinada janela de tempo se um minerador produziu $p\%$ dos blocos, ele deve ter o poder computacional de $p\%$ do total dos mineradores.

Os blocos podem ser obtidos de Blockchair e o acesso ao arquivo pode ser feito com mostrado na Listagem 4.6. Os dados obtidos da Blockchair vêm com o campo `guessed_miner` que indica o provável minerador do bloco. A análise do poder dos mineradores deverá indicar que nenhum deles tem mais da metade poder computacional,e consequentemente, não poderá executar o Ataque do 51%. Porém um outro cenário para o Ataque do 51% é o conluio entre os mineradores: dois ou mais mineradores se juntam para obter o controle da rede. O termo **Coeficiente de Nakamoto** [Milad et al. 2024]

[18]Tab Separated Values, formato que pode ser lido,por exemplo, pela biblioteca Pandas.

refere-se o número mínimo de mineradores que precisam se unir para obter o poder computacional suficiente para controlar a rede. Importante notar que a escolha de um mês para análise é arbitrária, dentro do mesmo mês é possível analisar os dados usando uma janela deslizante e chegar a números um pouco diferentes. A conclusão final sobres estes dados exige uma análise criteriosa.

```
import pandas as pd
b = pd.read_csv('blockchair_bitcoin_blocks_20240101.tsv.gz', sep='\t')
for m in b['guessed_miner']:
    print(m)
```

Listagem 4.6: Leitura de um bloco.

```
Foundry USA Pool
Foundry USA Pool
AntPool
AntPool
AntPool
Foundry USA Pool
Foundry USA Pool
F2Pool
F2Pool
F2Pool
AntPool
AntPool
MaraPool
MaraPool
```

Listagem 4.7: Lista parcial dos mineradores de 01/01/2024

A Listagem 4.7 apresenta um excerto dos mineradores do dia primeiro de Janeiro de 2024, produzido pela Listagem 4.6. Um olhar criterioso sobre estes dados nos faz questionar o sistema de "sorteio" do processo de mineração: existe muita repetição consecutiva de mineradores, levantando a dúvida se a listagem apresentada é realmente resultado de um processo estocástico ou se existe algum evento que faça a mineração consecutiva ocorrer com maior frequência. De fato a Mineração Egoísta, quando aplicado com sucesso faz com que o minerador produza mais blocos do que o esperado para o seu poder computacional. A principal evidência deixada na Blockchain por este ataque é a existência de blocos consecutivos com uma frequência maior do que o esperado em um processo estocástico.

Desta forma, a busca por evidências de Mineração Egoísta pode ser feita usando um algoritmo simples, baseado em um Teste de Permutação, com o seguinte raciocínio: o primeiro passo é contar quantas minerações consecutivas o minerador fez em uma determinada janela de blocos da Blockchain. O próximo passo consiste em produzir uma permutação aleatória sobre esse mesmos dados e contar as minerações consecutivas ocorridas neste teste. Cada vez que o número de minerações consecutivas dos dados originais é maior que o número de minerações consecutivas dos dados permutados aleatoriamente, temos um indício que os dados originais não foi produzido por um processo estocástico. Este processo está descrito no Algoritmo 1. Esta abordagem tem limitações, pois para termos uma chance de reproduzir a realidade, é necessário dividir a Blockchain em janelas - usualmente de um mês - incorrendo em testes múltiplos [P et al. 2016] que podem

comprometer os resultados obtidos.

Algoritmo 1 Busca por evidências de Mineração Egoísta

Input B: uma subsequencia da Blockchain,k: quantidadde de testes; M: identificação do minerador
$count \leftarrow 0$
$c \leftarrow$ número de blocos em sequencia de M em B
for $i \in 1..k$ **do**
    $p \leftarrow$ permutação aleatório sobre B
    $cp \leftarrow$ número de blocos em sequencia de M in p
    if $c \geq cp$ **then**
        $count \leftarrow count + 1$
    end if
end for
if $\frac{count}{k} > 0.95$ **then**
    M é Suspeito!
end if

P3 É possível usar a URL https://blockchain.info/rawaddr/$ENDERECO para obter o saldo atualizado de $*ENDERECO*. Utilize este serviço para obter o saldo dos endereços do primeiro ano do Bitcoin, obtidos na Listagem 4.4.

P4 Utilizando os dados da BlockChair, determinar o poder computacional dos mineradores em Janeiro de 2024.

P5 Com os mesmos dados da prática anterior, determinar o Coeficiente de Nakamoto em Janeiro de 2024.

P6 Implementar e executar a detecção de Mineração Egoísta (Algoritmo 1) e aplicar nos dados de Janeiro de 2024.

4.4. Análises Estatísticas

Análises estatísticas revelam informações importante sobre os dados da Blockchain [Xi et al. 2020,Molitor et al. 2023] fornecendo *insights* para outras análises e informações que podem alimentar os modelos de Aprendizado de Máquina. Algumas análises comumente feitas sobre os endereços:

Saldo O saldo de um endereço é claramente um dado de interesse. Com já foi mencionado anteriormente, o saldo é uma propriedade que precisa ser calculada contabilizando os `UTxOs`. Como o saldo varia no tempo, muitas vezes é considerado o saldo médio, mínimo e máximo nas análises;

Tempo de vida Os endereços podem ter padrões bem diferentes de uso: alguns podem ser descartáveis enquanto outros podem ser usado por anos. Essas diferenças podem ser observadas analisando o tempo de vida de um endereço, isto é, a sua primeira aparição até a sua última movimentação. A unidade de medida pode ser blocos ou dias;

Número de Transações Um usuário com preocupação de privacidade deve usar endereços descartáveis, com uma transação de depósito e outra saque. Já um cliente de uma *exchange* pode fazer diversas movimentações com o mesmo endereço.

Para as transações, as seguintes análises podem ser feitas:

Número de Entradas/Saídas As transações são compostas por uma lista de entrada e uma lista de saída, os tamanhos destas listas podem indicar o tipo de transação. Por exemplo, na Figura 4.4 é mostrado uma transação de 1 para n e um outra de n para n, este cenário pode ser uma *exchange* que mantém os bitcoins sob sua custódia, e deve criar uma transação para prover fundos aos endereços dos clientes e posteriormente, uma transação com um número grande de entradas e um número igual de saídas, para os clientes efetuarem os seus pagamentos;

Valores Existem várias análises que podem ser feitas nos valores de uma transação. As análises podem ser feitas para cada item da entrada (ou saída) ou pelo total da transação. Na Seção 4.4.1 é mostrado como os valores podem ser usados para detectar um serviço de *mixer*;

Taxa O valor pago como taxa por cada transação também pode ser usado para classificar as transações. É normal que carteiras e *exchanges* possuem valores *default* para a taxa a ser paga, então este item pode ser útil na classificação das transações.

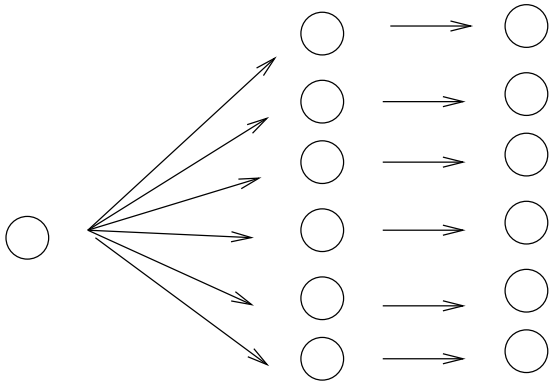

Figura 4.4: Duas transações. A primeira, representada pelo círculo mais à esquerda, representa um único endereço fazendo a transferência para n endereços. Na segunda transação, cada um dos n endereços faz uma transferência para um outro endereço.

Uma métrica bastante usada em análises que envolvem dados financeiros é o Índice Gini, que mede a desigualdade de um determinado recurso em uma população [Juodis et al. 2024]. Tal métrica gera valores entre 0 e 1, com valores maiores indicando maior desigualdade. Para exemplificar o seu uso, os dados de transações da *exchange* brasileira Mercado Bitcoin e do *mixer* Wasabi foram obtidos na plataforma WalletExplorer e os valores das transações foram usados para calcular o Índice Gini, os resultados são mostrados na Figura 4.5. Como o Índice Gini da *exchange* é maior que o do *mixer*, sabemos que esta *exchange* tem mais desigualdade nos valores de suas transações.

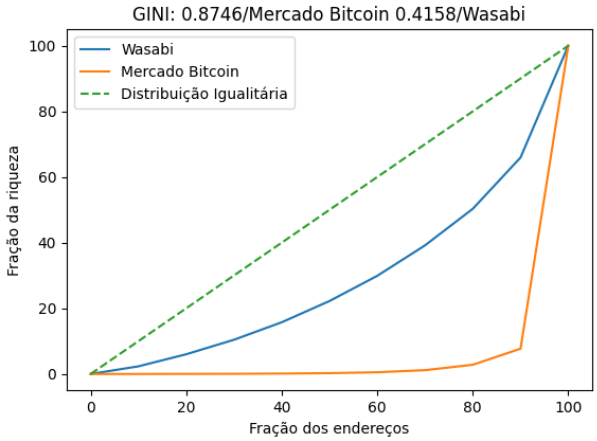

Figura 4.5: Índice Gini das transações da *exchange* MercadoBitcoin e do *mixer* Wasabi.

Outra métrica comumente usada é a Lei de Benford [Vičič and Tošić 2022], que analisa a frequência do dígito mais significativo dos elementos de um conjunto de dados. Amostras de dados que se desviem dos valores esperados pela Lei de Benford podem indicar que os seus valores não foram obtidos por um processo normal [Benford 1938]. Na Listagem 4.8 é mostrado um *script* em Python para calcular os valores de acordo com a Lei de Benford para os últimos cinco blocos da Blockchain. O resultado do `script` é mostrado na Figura 4.6 onde também são mostrados a frequência dos dígitos para as transações do *mixer* Whirlpool.

```
import json
import requests

def firstD(i):
    i = int(i)
    while i>=10:
        i=i/10
    return i

r = requests.get('https://blockchain.info/latestblock')
resp = r.json()
url = 'https://blockchain.info/rawblock/'
lh = resp['hash']

h=[0]*9
for i in range(5):
    r = requests.get(url+lh)
    resp = r.json()
    for i in range(1,resp['n_tx']):
        valortx = sum([o['value'] for o in resp['tx'][i]['out']])
        h[firstD(valortx)-1] =  h[firstD(valortx)-1] +1
    print(h)
    lh = resp['prev_block']
```

Listagem 4.8: Obtenção dos dados dos últimos 5 blocos para o cálculo dos valores da Lei de Benford.

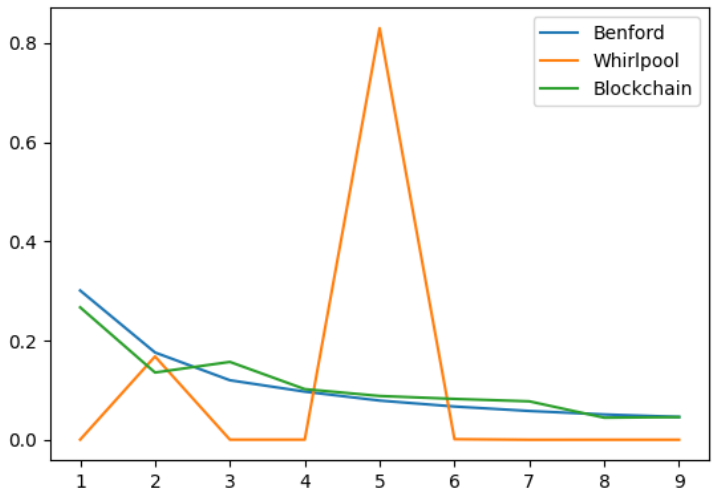

Figura 4.6: Lei de Benford. Comparativo da distribuição dos dígitos mais significativos obtidos das transações dos últimos 5 blocos da Blockchain e de um conjunto de transações do *mixer* Whirlpool.

4.4.1. Whirlpool

Para ilustrar como as análises estatísticas podem ser usadas para obter *insights* sobre os dados mostraremos uma aplicação das análises para o *mixer* Whirlpool [Schnoering and Vazirgiannis 2023]. Identificar transações do *mixer* Whirlpool é uma tarefa fácil: as transações sempre têm 5 saídas com os mesmo valor. Na Figura 4.7 é mostrado o número de transações de Whirlpool por bloco no ano de 2022. Mesmo sem uma análise estatística formal, é possível perceber que existem *bursts* de ocorrências dessas transações, e então, inferir que elas estão de alguma forma relacionadas.

A Figura 4.8 ilustra o resultado da análise de 144 blocos consecutivos, também do ano de 2022. Esta quantidade de blocos corresponde a aproximadamente um dia de transações. A complexidade e variedade dos subgrafos indicam que alguns usuários de *mixer* fazem reuso da ferramenta, provavelmente objetivando aumentar o anonimato das transações.

Na Figura 4.9 é mostrada a análise de um dos endereços pertencentes a uma transação de *mixer* do blocos analisados. Na figura, observa-se que $bc1qjdd...z9th$ dividiu o seu saldo em 12 transações, sendo 10 delas de *mixers*. Assim, o controlador do endereço controla 10 novos endereços de um conjunto de 50 possíveis.

4.4.2. Prática

P7 Obter o saldo de todos os endereços existentes ao final do primeiro ano do Bitcoin. Consultar a plataforma Blockchain.com para obter a situação atual dos endereços.

P8 Dada uma transação, calcular a razão entre o número de entradas e saídas.

P9 Usando o WalletExplorer, obter as transações das de algumas *exchanges*, calcular o

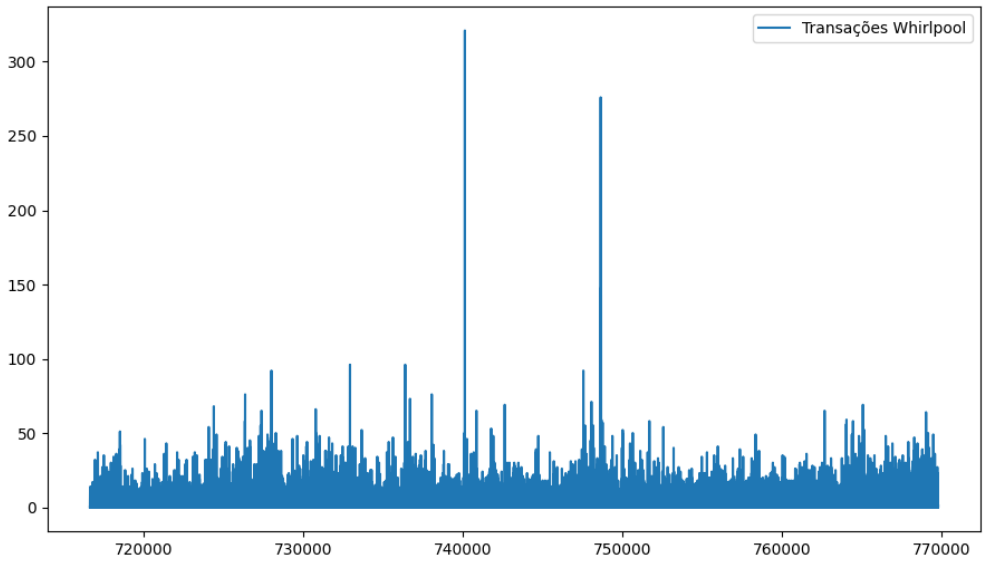

Figura 4.7: Número de transações identificadas com Whirlpool por bloco. Dos blocos 716599 a 769786, correspondente ao ano de 2022.

Índice Gini de cada uma delas e comparar com a Figura 4.5.

4.5. Análise da Blockchain

Existem algumas análises específicas para serem aplicadas na Blockchain. A seguir são apresentadas algumas comumente utilizadas.

4.5.1. Heurística das Entradas Múltiplas

Umas das mais importantes heurísticas utilizadas na análise da Blockchain é chamada justamente "Heurística Um" ou simplesmente H1, como um referência a uma heurística principal. Ela é uma consequência direta do modelo `UTxO` adotado pelo Bitcoin. Como visto na Seção 1 deste trabalho uma transação normalmente é composta por diversos `UTxOs`, comumente de endereços diferentes. Assim, uma transação é composta por diversos endereços diferentes de entrada que precisam ser assinados por diversas chaves privadas diferentes. Considerando que o processo de assinatura deve ser feito em um ambiente seguro, é bem razoável inferir que uma mesma entidade controla todos os endereços de entrada de uma transação. Utilizando esta heurística no exemplo da Figura 4.2 podemos concluir que *bcd* e *cde* são controlados pela mesma entidade. A propriedade de unir transações pela H1 é transitiva: se uma $tx1$ tem ao menos um endereço em comum com uma $tx2$ e $tx2$ tem ao menos um endereço em comum com $tx3$, os endereços de entrada de $tx1$,$tx2$ e $tx3$ pertencem a mesma entidade. Observe que H1 não se aplica sobre os endereços de saída: em uma única transação o usuário pode fazer pagamentos para entidades diferentes.

A aplicação da H1 é comumente chamada de clusterização: um **cluster de ende-**

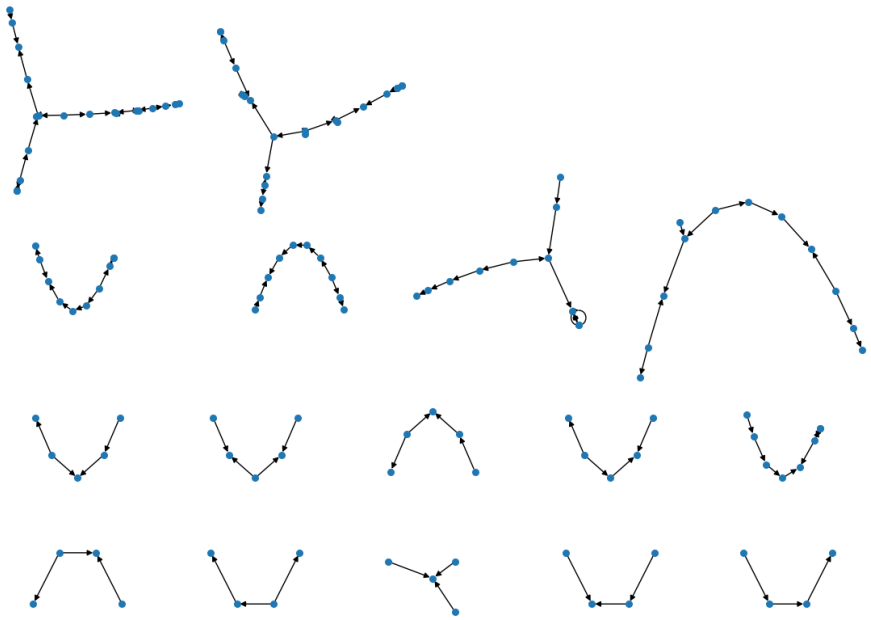

Figura 4.8: Componentes conexas de tamanho maior que três das transações de Whirlpool. Os vértices correspondem a transações no período escolhido. As arestas indicam a existência de pagamento entre os endereços das transações.

reços é um conjunto de endereços controlados pela mesma entidade. Esse é um passo importante da análise forense, por exemplo, se um endereço do cluster está envolvido em atividades ilegais e um outro endereço fez compras em um *marketplace* com entrega, existe a possibilidade desta pessoa ser identificada pelas autoridades.

A clusterização H1 não deve ser confundida com o mesmo termo comumente usado em Aprendizado de Máquina. Uma implementação do H1 é apresentada no Algoritmo 2.

Algoritmo 2 Clusterização usando H1

```
1:  Input TX.inputs: Lista de endereços de entradas de uma lista de transações
2:  Clusters ← ∅
3:  for enderecos ∈ Tx.inputs do
4:      for c ∈ Clusters do
5:          if enderecos ∩ c ≠ ∅ then
6:              enderecos ← enderecos ∪ c
7:              Remove c de Clusters
8:          end if
9:      end for
10:     Insere enderecos em Clusters
11: end for
12: return Clusters
```

A H1 é amplamente usada na análise de Blockchain e é tida como um método

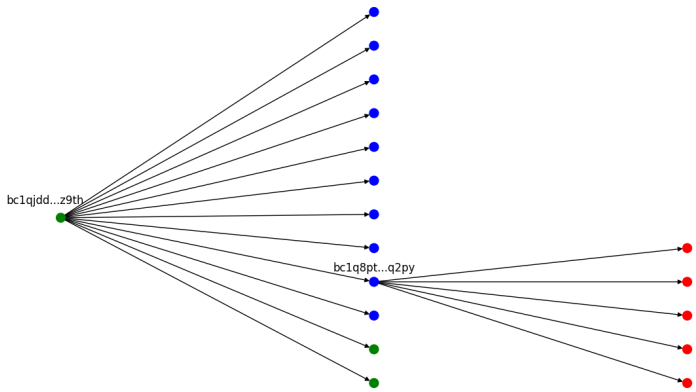

Figura 4.9: Grafo da vizinhança do endereço $bc1q8pt...q2py$. Vértices em azul indicam que os endereços fazem parte da entrada de uma transação de *mixer*, e em vermelho, saída. Arestas e vértices omitidos para melhor visualização.

"a prova de erros" [Xi et al. , Harrigan and Fretter 2017], mas existem pelo menos dois cenários onde o seu uso pode produzir erros:

Transações `CoinJoin` Este tipo de transações permite aos usuários deliberadamente usarem serviços de *mixers* para "juntar" transações, invalidando a H1 [Schnoering and Vazirgiannis 2023];

Exchanges Em geral quando alguém possui criptomoedas em *exchanges* a chave privada dessas criptomoedas está sob o cuidado da *exchange*. É comum que a *exchange* junte diversas transações de vários clientes antes de enviar para a rede P2P. Este é um cenário onde a falha da heurística H1 pode trazer uma informação diferente: o cluster formado pelo H1 pode ser considerado a mesma entidade: a *exchange*, e essa entidade é formada por diversas subentidades, os clientes.

As informações produzidas pelo uso da H1 são muito importantes no processo de análise forense: saber que um determinado indivíduo controla diversos endereços pode ser usado para determinar sua localização geográfica e, determinar que uma carteira é controlada por uma *exchange* pode fornecer informações precisas sobre a pessoa e, eventualmente, até o confisco dos bitcoins.

4.5.2. Heurística do Troco

Em geral, as transações exigem o uso do endereço de troco, que é controlado pela mesma entidade da entrada da transação.

A Heurística do Troco - ou simplesmente H2 - também é usada para clusterizar endereços. A H2 não é tão precisa quanto a H1, carteiras diferentes podem implementar o endereço de troco de formas diferentes, tornando a classificação da H2 um pouco mais complexa e imprecisa. Em geral, para classificar o endereço trc de uma transação Tx como endereço de troco, as seguintes características serão consideradas [Xi et al. , Zhang et al. 2020, Ermilov et al. 2017, Meiklejohn et al. 2013]:

- A transação Tx não deve ser uma transação do tipo Coinbase;
- É a primeira aparição de trc;
- trc é o único endereço de primeira aparição;
- O valor pago para trc é menor que a menor `UTxO`.

4.5.3. Dust

A técnica do **Dust** (ou poeira, em português) é um método ativo do uso do H1 [Loporchio 2023] que visa comprometer o pseudo-anonimato das criptomoedas baseadas em `UTxO`. Esta técnica consiste em transferir pequenas quantidades de bitcoins - daí o nome *dust* - para um quantidade grande de endereços. Espera-se que as carteiras produzam uma transação que "recolha" as pequenas quantidades de bitcoins dos endereços controlados por ela em um único endereço. Este processo remove várias `UTxOs`, simplificando o processo de novas transações. Como a transação criada possui várias entradas quando a H1 é aplicada, será revelado quais carteiras fazem parte do mesmo cluster.

4.5.4. Identificação de DNMs

Os DNMs são preocupações constantes nas análises feitas buscando atividades ilícitas. Vários trabalhos abordam o problema de identificar atividades relacionadas a DNM, muitos deles usando Aprendizado de Máquina [dos Reis et al. 2023,Kanemura et al. 2019,Hiramoto and Tsuchiya 2020]. Aqui, destacamos duas heurísticas simples que podem ser usadas na busca por atividades relacionadas a DNMs:

Transações de Custódia Em [Hiramoto and Tsuchiya 2020], é apontado que alguns DNMs mantém sob sua custódia o pagamento feito pelos compradores, repassando o valor para os vendedores após o comprador informar que recebeu o produto.

Valor Conhecido Segundo [Dolejška et al. 2023] é possível fazer um `web scrapping` nas plataformas de alguns DNMs e detectar o momento e o valor das vendas de certos produtos. Usando esta informação é possível procurar na blockchain por suspeitos.

4.5.5. Prática

P10 Aplicar a clusterização H1 nas transações do primeiro ano do Bitcoin. Após a obtenção dos clusters, calcular o saldo de cada um deles.

P11 Desenvolver uma função para selecionar os endereços de custodia em uma janela de tempo. Você pode considerar que o endereço de custodia tem apenas duas transações e a transação de pagamento possui uma ou duas saídas, o vendedor e, possivelmente, um outro endereço para o DNMs receber a sua comissão. Em um fração destas carteiras, é possível que o valor seja reembolsado para o comprador.

P12 Desenvolver uma função para selecionar as transações de uma janela de tempo que estão dentro de uma faixa de valores.

P13 Desenvolver uma função para identificar candidatos a endereços de troco de acordo com Seção 4.5.2

P14 De acordo com [Schnoering and Vazirgiannis 2023], transações candidatas de `CoinJoin` do *mixer* Whirpool podem ser encontradas com uma heurística bem simples: transações com 5 entradas; e 5 saídas com os mesmos valores. Implemente um algoritmo para encontrar estas transações.

4.6. OSINT

A *Open Source Intelligence* (OSINT) surgiu no início da Segunda Guerra Mundial, marcada pelo estabelecimento do Serviço de Monitoramento da BBC na Grã-Bretanha em 1939 e do Serviço de Monitoramento de Emissões Estrangeiras (FBMS) em 1941 nos Estados Unidos. A origem da expressão "*Open Source Intelligence*" e do acrônimo OSINT na literatura ocorreu em um artigo de Robert Steele escrito em 1990 [Steele 1990, Block 2023]. De maneira proeminente, uma edição especial do *American Intelligence Journal* em 1993 foi inteiramente dedicada à inteligência de fontes abertas e reuniu os artigos apresentados durante a primeira conferência sobre inteligência de fontes abertas realizada em 1992 [Block 2023].

A OSINT representa a abordagem mais elementar na obtenção de informações, caracterizando-se pela coleta de dados a partir de fontes abertas, como internet, transmissões, documentos, entre outros, para processamento posterior. Ao empregar informações acessíveis publicamente, há benefícios como a obtenção em tempo real e a facilidade de acesso aos dados, além de ser um método de coleta de informações de custo mais baixo [Hwang et al. 2022]. Um exemplo da estrutura básica da OSINT é mostrado na Figura 4.10.

No contexto das empresas, o conhecimento representa uma forma de poder. As organizações fazem uso de fontes de OSINT para explorar novos mercados, vigiar as atividades dos concorrentes, planejar estratégias de marketing e antecipar eventos que possam impactar as operações presentes e o crescimento futuro. Anteriormente, o acesso às fontes de OSINT estava restrito a grandes corporações com orçamentos substanciais para inteligência. Contudo, nos dias de hoje, devido à difusão generalizada da Internet, até mesmo pequenas empresas com recursos financeiros limitados podem empregar fontes de OSINT e integrar as informações obtidas em seus planos de negócios [Hwang et al. 2022].

Contrastando com os usos benéficos mencionados, as fontes de OSINT também podem ser exploradas de maneiras prejudiciais, por exemplo organizações terroristas podem empregar essas fontes para planejar ataques. Elas podem coletar informações sobre o alvo, analisar plataformas de mídia social para recrutar mais membros, obter informações militares inadvertidamente divulgadas pelo governo (como métodos de fabricação de explosivos) e utilizar diversos canais de mídia para disseminar propaganda globalmente. Adicionalmente, os dados obtidos por meio de OSINT podem servir como base para a prática de diversos crimes cibernéticos [Hwang et al. 2022].

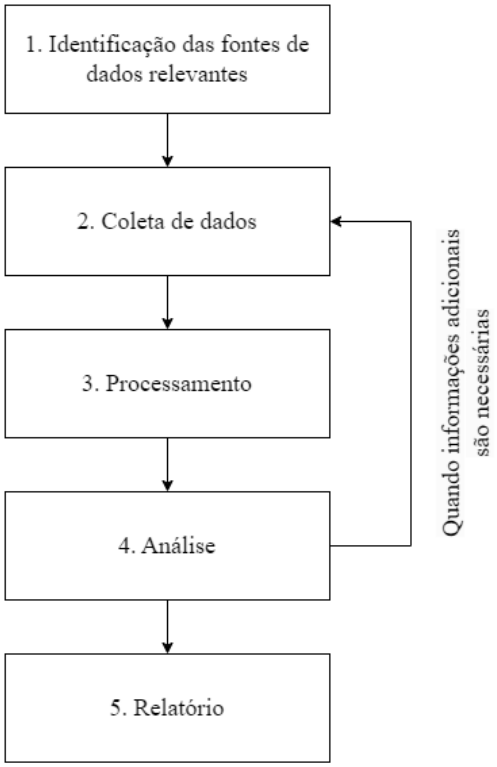

Figura 4.10: Estrutura da OSINT.

4.6.1. Aplicação em investigações digitais

Com o advento da Internet, diversos setores da economia começaram a explorar os recursos e as vantagens proporcionadas por essa tecnologia. Surgiram, por exemplo, diversos sites de comércio eletrônico e aplicativos bancários, o que possibilitou que os usuários realizassem transações *online*. Além disso, as plataformas de mídia social surgiram como ferramentas estratégicas importantes para as organizações divulgarem produtos, serviços e facilitar a comunicação com funcionários e clientes [Yeboah-Ofori and Brimicombe 2018].

Páginas do Facebook e *feeds* do Instagram e Twitter são explorados para fins de inteligência empresarial. Isso se deve pelo fato de que as corporações necessitam de uma capacidade de inteligência crescente para serem competitivas, pois precisam antecipar e reagir às mudanças que ocorrem no contexto em que operam. Por essa razão, as redes sociais se mostram como uma fonte rica de informações produzidas por usuários e empresas de todo o mundo [Costa et al. 2012].

No contexto das criptomoedas, uma das principais razões para o aumento da literatura sobre o tema é a disponibilidade de dados, uma vez que séries históricas extensas de preços são gratuitas e fáceis de baixar de fontes de dados muito diversas, como sites de classificação de moedas e empresas especializadas em moedas digitais [Vidal-Tomás 2022]. Consequentemente, formuladores de políticas, investidores e acadêmicos podem analisar criptomoedas sem quaisquer limitações em relação aos dados. Como exemplo,

pode-se citar diversos sites que apresentam informações sobre transações de criptomoedas como o blockchain.com, etherscan.io, Wallet Explorer, dentre outros.

Portanto, a ampla disponibilidade de dados na internet desempenha um papel fundamental na aplicação da OSINT nas investigações digitais. O fácil acesso a extensas fontes de informações *online*, incluindo transações em *Blockchain*, fóruns especializados e redes sociais, oferece aos investigadores uma base robusta para coleta e análise de dados. Essa riqueza de dados permite a identificação de padrões de comportamento, rastreamento de transações financeiras e a criação de perfis detalhados de indivíduos ou entidades envolvidas em atividades suspeitas com criptomoedas. A OSINT, aproveitando essa abundância de informações acessíveis publicamente, torna-se uma ferramenta valiosa na desmistificação de operações ilícitas envolvendo ativos digitais, contribuindo assim para a eficácia das investigações digitais nesse cenário em constante evolução.

4.6.2. Ética e privacidade em OSINT

A integridade na prática de OSINT é intrinsecamente ligada tanto à legalidade e conformidade quanto ao uso responsável. O respeito pelas leis e regulamentações locais e internacionais é crucial para assegurar que a coleta e análise de dados ocorram dentro de parâmetros éticos e legais. Isso inclui a consideração de normas de privacidade, proteção de dados e quaisquer restrições específicas aplicáveis ao contexto da investigação. Paralelamente, o uso responsável é essencial para garantir que as informações obtidas sejam interpretadas, divulgadas e aplicadas de maneira ética. Envolve a ponderação sobre as possíveis consequências das ações, a minimização de danos potenciais e a busca por uma disseminação equitativa e justa das informações. Juntos, o compromisso com a legalidade e conformidade, aliado ao uso responsável, forma a base ética necessária para uma prática de OSINT que respeita tanto os parâmetros legais quanto os princípios éticos fundamentais [Böhm and Lolagar 2021].

A verificação de fontes é um princípio essencial na prática de OSINT, destacando a importância de confirmar a credibilidade e autenticidade das fontes de informação. Os profissionais de OSINT devem adotar métodos robustos para garantir a confiabilidade dos dados coletados, evitando o uso e a disseminação de informações falsas ou não verificadas [Tabatabaei and Wells 2016]. Paralelamente, a proteção da privacidade emerge como um imperativo ético, com a necessidade de salvaguardar os dados pessoais dos indivíduos investigados. Ao coletar e analisar informações, é crucial adotar práticas que evitem a divulgação indevida de dados sensíveis e respeitem as normas de privacidade estabelecidas. A combinação eficaz desses dois princípios contribui para a credibilidade das investigações de OSINT, assegurando a precisão e a integridade das informações, ao mesmo tempo que protege os direitos fundamentais de privacidade das partes envolvidas [Bean 2011].

4.6.3. Trabalhos desenvolvidos por empresas

Diversas empresas e grupos especializados oferecem serviços de análise forense de criptoativos, como o Bitcoin. Exemplos de empresas desse ramo são a Chainalysis, CipherTrace, Elliptic, Coinfirm, dentre outras. Esse empenho em fornecer serviços de segurança se justifica pelo fato de que há um número cada vez mais crescente de atividades criminosas envolvendo criptomoedas. Além disso, a movimentação financeira nas *Blockchains*

cresce a cada ano. De acordo com o relatório sobre crimes criptográficos e lavagem de dinheiro da CipherTrace, no final do terceiro trimestre de 2022, o valor total de mercado de todos os ativos criptográficos, incluindo *stablecoins* e *tokens*, era de aproximadamente US$ 1,1 trilhão [CipherTrace 2023a].

A Chainalysis atua em conjunto com orgãos reguladores e empresas do setor de criptomoedas para fornecer informações cruciais na prevenção de crimes financeiros, como lavagem de dinheiro, fraudes, dentre outros. A empresa já colaborou com autoridades em vários casos, incluindo os que envolvem o mercado da Silk Road[19] [Chainalysis 2020]. Uma das soluções oferecidas aos clientes é o "Chainalysis Reactor". Segundo a empresa, a ferramenta de investigação conecta transações de criptomoedas a entidades do mundo real, examina atividades criminosas, como o movimento de fundos roubados, bem como atividades legítimas, como empréstimos instantâneos e transferências de NFTs [Chainalysis 2023]. O software pode ser visto na Figura 4.11.

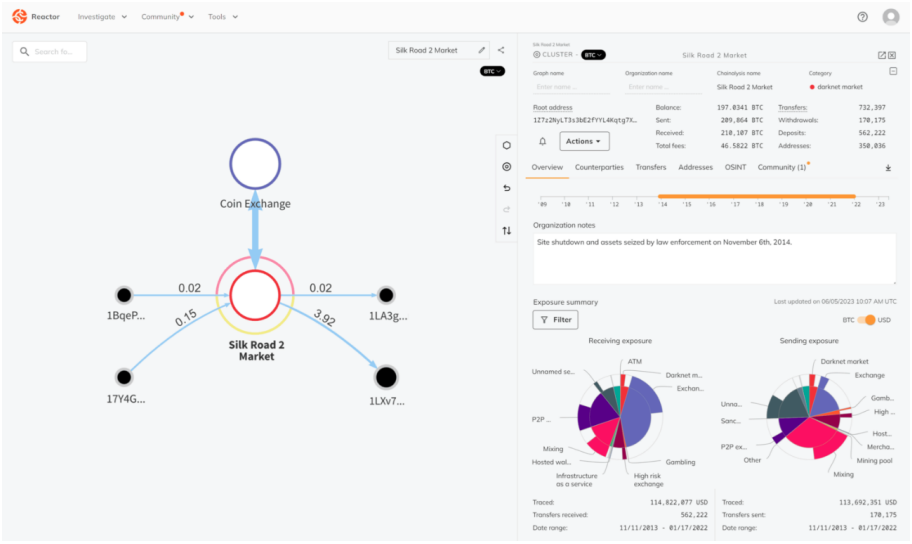

Figura 4.11: Chainalysis Reactor

Assim como a Chainalysis, a CipherTrace também oferece diversas soluções de forense que utilizam algoritmos de inteligência artificial. Uma delas é o Inpector, cuja finalidade é vincular endereços criptográficos a organizações do mundo real, entidades sancionadas, endereços IP e eventos. É possível acompanhar os históricos transacionais completos e há um explorador gráfico que facilita o acompanhamento do dinheiro, além de recursos integrados de gerenciamento de casos que potencializam a investigação colaborativa [CipherTrace 2023b]. Uma visão geral de da ferramenta é mostrada na Figura 4.12.

A Elliptic é proprietária do Investigator. O software utiliza técnicas avançadas de aprendizado de máquina para a identificação de atividades ilícitas. A ferramenta disponibilizada pela empresa utiliza análise de dados e aprendizado de máquina para identificar

[19]O SilkRoad foi um marketplace, ativo entre 2011 e 2013, com foco em produtos e serviços ilegais. Criada por Ross Ulbricht, a Silk Road permitia transações anônimas com Bitcoin e outras criptomoedas, facilitando a compra e venda de produtos ilegais, incluindo drogas, armas e documentos falsificados [Trautman 2014].

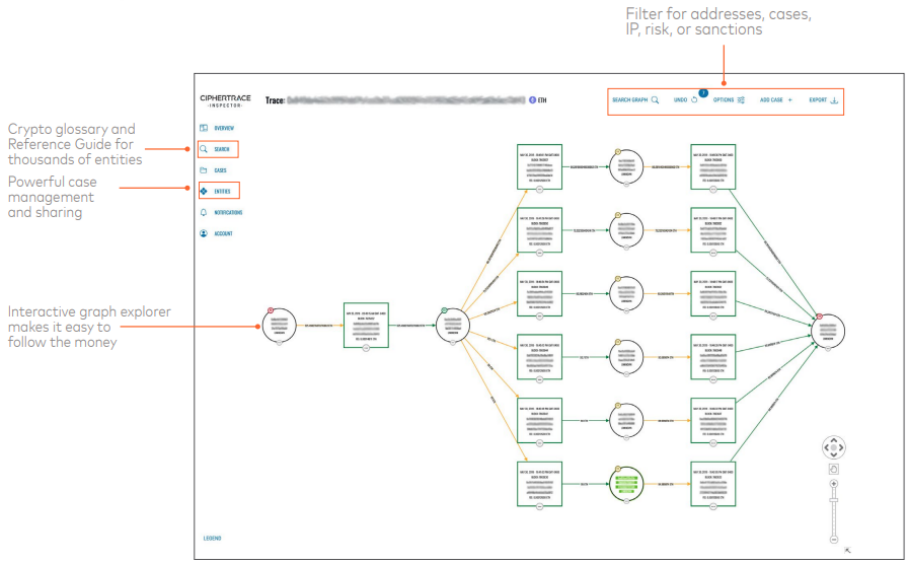

Figura 4.12: CipherTrace Inspector

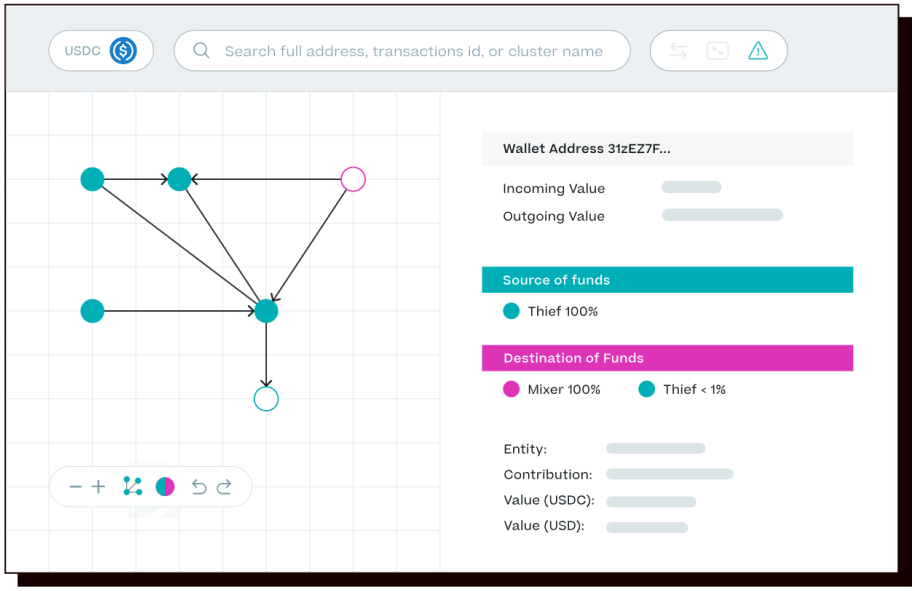

Figura 4.13: Elliptic Investigator

transações suspeitas e rastrear fundos. Segundo a Elliptic 2023, é possível conduzir investigações com um único clique em *blockchains* e ativos com facilidade. Além disso, pode-se visualizar instantaneamente o fluxo de fundos criptográficos através de carteiras, entidades e transações para encontrar rapidamente evidências significativas e reduzir o tempo e os recursos necessários para encerrar casos. Uma tela de exemplo do Investigator é mostrada na Figura 4.13.

O Wallet Explorer, embora não tenha desenvolvido uma ferramenta específica de forense, desempenha um papel crucial no ecossistema de criptomoedas ao manter informações abrangentes sobre o tema. Um dos pontos positivos do site é a classificação de carteiras de criptomoedas. Sua capacidade de categorizar e organizar diferentes tipos

de carteiras contribui para a compreensão e facilita análises mais aprofundadas sobre o comportamento das transações e a dinâmica do mercado de criptoativos [Torrez 2023].

Embora não seja um empresa propriamente dita, ainda pode-se citar o Wallet Explorer. Esta é uma plataforma que fornece um serviço de classificação de carteiras de Bitcoin. Ao inserir um endereço na *search engine*, a plataforma classifica o endereço em várias categorias possíveis, como apostas, serviços de *mixer*[20] e *exchanges*[21]. O Wallet Explorer pode ser considerado um "padrão ouro" na classificação de endereços de Bitcoin, sendo amplamente adotada em diversos trabalhos acadêmicos[22].

Essas empresas desempenham um papel fundamental no fornecimento de ferramentas e tecnologias para análise forense de criptomoedas, apoiando agências governamentais, empresas e reguladores na aplicação de regulamentos e na prevenção de atividades ilegais no espaço das criptomoedas. Porém, a principal desvantagem de ferramentas de código fechado é a falta de transparência e visibilidade do código-fonte. Por serem proprietárias e as empresas não disponibilizarem seu código para o público, os usuários não podem examinar, modificar ou verificar como a ferramenta opera internamente. Isso pode resultar em dependência de fornecedor e falta de personalização para atender às necessidades específicas dos usuários, o que justifica a criação de uma ferramenta *Open Source*.

4.6.4. Caso do Faraó dos Bitcoins

O "Faraó dos Bitcoins" é um empresário acusado de promover pirâmides financeiras usando Bitcoin. Uma busca pelo termo "Faraó dos Bitcoins" nos leva à página Carteira usada pelo Faraó dos Bitcoin é revelada que contém nove endereços que foram obtidos pelos investigadores. Na Listagem 4.9 os endereços de entradas das transações executas pelos endereços contidos na lista `farao` são obtidos e salvos. Usando a H1 é possível fazer uma análise sobre estes endereços e obter mais informações sobre eles. O endereço 1*JawW...xkwo* participa de 263 transações e formou um cluster de tamanho 6287. Dificilmente todos estes endereços pertencem à mesma pessoa. Uma busca pelo endereço no WalletExplorer indica que existem transações destes endereço com um serviço de *mixer*, que justifica o tamanho do cluster. Conforme pode ser observado na Tabela 4.2 a plicação do H1 nos demais endereços produzem três novos endereços:

- *bc*1*qz*30*fyctnylenx*584*w*483*ekxd*9*tyds*07*h*7*pyexq*;
- *bc*1*qp*2*nx*27*jln*57*va*4*cw*55*kmpu*34*h*30*s*6*plzylkmtx*;
- *bc*1*q*6*p*5*pn*7*l*9*n*0*vs*4*v*4*rc*5*vet*3*zs*9*hfhylwh*0*fqm*35

[20] Serviço que mistura transações de múltiplos usuários, aumentando a privacidade ao tornar difícil rastrear a origem específica dos fundos na Blockchain. Ele promove o anonimato ao embaralhar as transações, protegendo a identidade e preservando a fungibilidade das criptomoedas [Wu et al. 2021a].

[21] Uma exchange de criptomoedas é um mercado onde os usuários podem comprar e vender criptomoedas. Muitas delas oferecem apenas serviços de negociação entre criptomoedas, enquanto algumas possibilitam transações entre moedas fiduciárias (por exemplo, Dólar Americano ou Euro) e criptomoedas [Xia et al. 2020].

[22] Uma busca no Google Scholar indica uma extensa lista de trabalhos acadêmicos que referenciam a ferramenta.

Endereço	Cluster
1Jaw...xkwo	–
bc1qlu...atak	bc1qlu...atak
bc1qt...y95w	bc1qt...y95w
bc1q8...mq3d	bc1q8...mq3d,bc1qn...0clv
bc1qq...w85t	bc1qq...w85t,**bc1q6...qm35**
bc1qn...0clv	bc1qn...0clv,bc1q8...mq3d
bc1qu...mynm	bc1qu...mynm,bc1qe...zull,**bc1qz...yexq**
bc1qn...9acq	bc1qn...9acq, **bc1qp...kmtx**
bc1qe...zull	bc1qe...zull,bc1qu...mynm,**bc1qz...yexq**

Tabela 4.2: Resultado da aplicação do H1 em cada um dos nove endereços do Faraó do Bitcoin. Na primeira linha o cluster é composto por 6287 endereços, que foram omitidos. Em destaque os endereços obtidos pela heurística(Endereços apresentados de forma compactada).

, que podemos afirmar com muita certeza também são controlados pelo Faraó dos Bitcoins e devem sofrer o mesmo processo de investigação dos demais endereços.

Na Figura 4.14 é mostrado as transações envolvendo os endereços controlados pelo Faraó (em azul, sem o endereço 1*JawW...xkwo*). Em verde os endereços que fizeram pagamentos e em vermelho os endereços que receberam pagamentos.

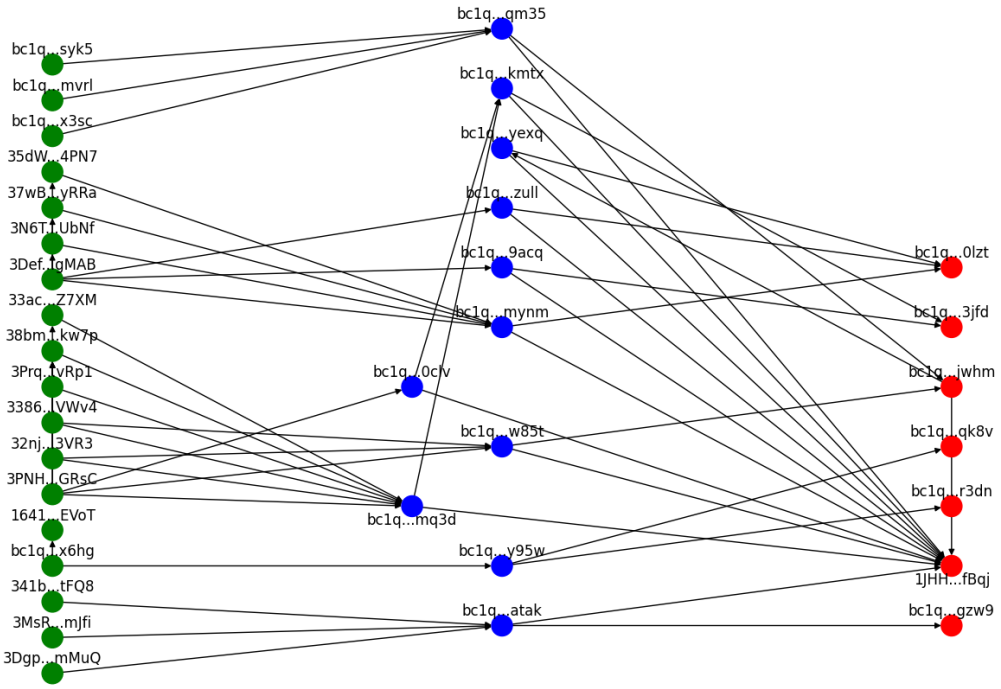

Figura 4.14: Transações envolvendo os endereços do Faraó dos Bitcoins. Em azul os endereços controlados pelo Faraó dos Bitcoins. Em verde os endereços que fizeram pagamento e em vermelho os endereços que receberam pagamentos.

```
import json
import requests
import pickle
import time

farao = ['1JawWE56G5NmnB5iuYbFikbdETs88Fxkwo',
         'bc1qluuy04mjxqj8yc44lgnez8eml4pwulvukfatak',
         'bc1qt7jjpqdfvqhtqkadlnuhzzem2tateg7mm0y95w',
         'bc1q8kmtzc0a43w0cjrzwzwqsa9frxaseyzcg6mq3d',
         'bc1qqgjmxevtn3cyg8cvxfg7yyk6a7n3zudt4hw85t',
         'bc1qn9k6s0lyxgw5mdndta3780md23z9kmu4980clv',
         'bc1qu9tj6kcusrncvm7wm06n2mqq0jtfg26vk9mynm',
         'bc1qnanyweuswqm9sz3d93ag0vrc69mpq4v40g9acq',
         'bc1qehzj8sulj3plzuarzzmdm77d6rd8chvc5hzull']

tocluster = {}
for f in farao:

    url = "https://blockchain.info/rawaddr/{}"
    resp = requests.get(url=url.format(f))
    data = resp.json()

    tocluster[f] = []
    for tx in data['txs']:
        temp = [ i['prev_out']['addr'] for i in tx['inputs']]
        if f in temp:
            tocluster[f].append(temp)

    n_tx =data['n_tx']
    done = len(data['txs'])

    while (done<n_tx):
        url = "https://blockchain.info/rawaddr/{}?offset={}"
        resp =requests.get(url=url.format(f,done))
        data = resp.json()
        for tx in data['txs']:
            temp = [ i['prev_out']['addr'] for i in tx['inputs']]
            if f in temp:
                tocluster[f].append(temp)
        done += len(data['txs'])
        time.sleep(5)

    time.sleep(5)

with open('txsFarao.pkl', 'wb') as sf:
  pickle.dump(tocluster, sf)
```

Listagem 4.9: Obtenção dos endereços de entradas das transações

4.7. Aprendizado de Máquina

O Aprendizado de Máquina (AM) é um campo da Inteligência Artificial que se concentra no desenvolvimento de algoritmos capazes de aprender a partir de dados e tomar decisões baseadas em padrões identificados. Esses algoritmos são projetados para melhorar seu desempenho ao longo do tempo, sem a necessidade de programação explícita para cada

tarefa [Mahesh 2020]. No contexto das criptomoedas, o uso de AM tem se mostrado particularmente promissor, fornecendo ferramentas poderosas para a análise de grandes volumes de dados gerados pela blockchain e auxiliando, por exemplo, na detecção de atividades ilícitas [Weber et al. 2019] e previsão de preços [Phaladisailoed and Numnonda 2018].

Existem duas abordagens principais: o aprendizado supervisionado e o não supervisionado. No aprendizado supervisionado, o algoritmo é treinado com um conjunto de dados rotulados, onde as entradas são mapeadas para as saídas corretas. O objetivo é que o modelo aprenda a generalizar a partir desses exemplos para prever corretamente a saída para novos dados não rotulados [Singh et al. 2016]. Já no aprendizado não supervisionado, o modelo é aplicado a um conjunto de dados sem rótulos, com o objetivo de descobrir estruturas ocultas, como grupos ou padrões, que não eram previamente conhecidos [Khanum et al. 2015]. Ambas as abordagens são amplamente utilizadas no estudo e na aplicação de AM em criptomoedas, cada uma oferecendo vantagens para diferentes tipos de problemas.

Os exemplos discutidos neste texto são didáticos e aplicados ao contexto das criptomoedas, destacando como o AM pode ser utilizado para resolver problemas específicos, como a detecção de transações ilícitas. Embora o campo de AM aplicado às criptomoedas seja relativamente recente, já existe uma vasta literatura acadêmica e uma crescente base de pesquisa que explora este tema. Estudos recentes têm demonstrado o potencial dessas técnicas para enfrentar desafios complexos e, ao mesmo tempo, promover avanços significativos nas investigações forenses realizadas.

4.7.1. Aplicação de algoritmo não-supervisionado

Existem diversos algoritmos de aprendizado de máquina não supervisionado, cada um projetado para identificar padrões ou estruturas nos dados sem a necessidade de rótulos predefinidos. Alguns dos principais algoritmos incluem a Análise de Componentes Principais (PCA), utilizada para redução de dimensionalidade [Shah et al. 2021]; o DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*), que identifica clusters baseado na densidade dos dados [Dokuz et al. 2020]. Entre esses, o K-means se destaca por sua popularidade e eficiência, particionando os dados em k clusters distintos, minimizando a variabilidade interna de cada grupo e facilitando a descoberta de padrões inerentes aos dados [Wang et al. 2018].

O K-means é um dos algoritmos de clustering mais amplamente utilizados no aprendizado de máquina, devido à sua simplicidade e eficiência. O objetivo principal do K-means é particionar um conjunto de dados em k grupos distintos (clusters), de modo que os objetos dentro de um mesmo cluster sejam mais semelhantes entre si do que aos objetos em outros clusters. Essa semelhança é geralmente medida pela distância euclidiana entre os pontos de dados [Chong et al. 2021].

O algoritmo K-means começa com a seleção de k centros (ou centróides) iniciais, que podem ser escolhidos de maneira aleatória ou baseados em uma estratégia específica [Hamerly and Elkan 2003]. A seguir, o algoritmo realiza as seguintes etapas iterativas:

- **Atribuição de clusters:** Cada ponto de dado é atribuído ao cluster cujo centróide

é o mais próximo, com base na distância euclidiana. Essa etapa cria k clusters iniciais.

- **Recalcular os centróides:** Após a atribuição inicial, o centróide de cada cluster é recalculado como a média de todos os pontos de dados pertencentes a esse cluster.
- **Repetição:** As etapas de atribuição e recalculação são repetidas até que os centróides não se movam significativamente, ou até que um número máximo de iterações seja atingido. Esse processo de repetição garante que os clusters se ajustem gradualmente até atingir uma configuração estável.

Para demonstrar o uso do K-means foi escolhido o problema de classificar as transações do *mixer* Wasabi [Ádám Ficsór et al. 2021], para tal, foram utilizadas as seguintes características definidas em [Stütz et al. 2022]:

Valores Únicos de Saída O processo de *mixing*, que esconde a relação de origem e destino dos valores, só tem efeito se houverem valores repetidos;

Razão entre entrada e saída O CoinJoin do Wasabi une n entidades, e cada entidade deve produzir duas saídas: uma do *mixer* e outra de troco. Ainda uma última saída para remuneração do operador do *mixer*. Assim para n entradas espera-se $2n+1$ saídas;

Reuso de entrada O reuso de endereços de entrada aponta uma falta de preocupação com privacidade;

SegWit O serviço de *mixer* usa endereços do tipo SegWit.

Utilizando o *dataset* disponibilizado em [Wu et al. 2021b], foram obtidas 100 transações aleatórias do conjunto com o *label* Wasabi. Outras 100 transações aleatórias foram obtidas da Blockchain, respeitando o mesmo intervalo de tempo das transações Wasabi obtidas. Como resultado, foi obtido um conjunto de 200 transações, sendo 100 transações Wasabi e 100 transações não-Wasabi. Na Figura 4.15 é mostrado o resultado da clusterização do conjunto testado usando o método `KMeans` da biblioteca `Scikit Learn`.

4.7.2. Aplicação de algoritmo supervisionado

Existem diversos algoritmos de aprendizado de máquina supervisionado, cada um com características e aplicações específicas para prever resultados com base em dados rotulados. Alguns dos algoritmos mais comuns incluem o Support Vector Machine (SVM), que busca maximizar a margem entre diferentes classes [Auti et al. 2022]; o Regressão Logística, frequentemente utilizado para problemas de classificação binária [Andi 2021]. Entre esses, a Árvore de Decisão se destaca por sua simplicidade e interpretabilidade, criando um modelo em forma de árvore onde cada decisão é baseada em regras lógicas "se-então", facilitando a compreensão e a aplicação em diversos contextos [Rathan et al. 2019].

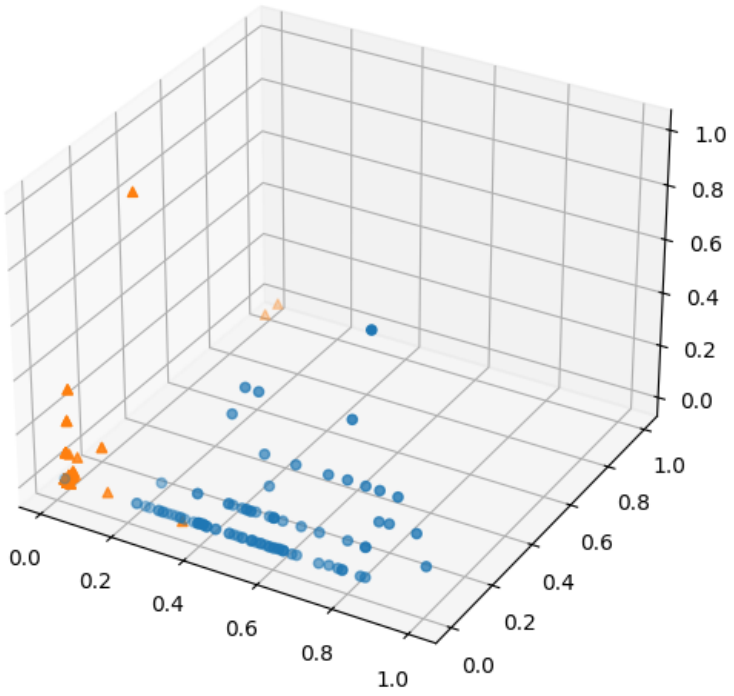

Figura 4.15: Resultado da aplicação do algoritmo K-means em transações. Em azul, as transações Wasabi e em amarelo as transações não-Wasabi.

Uma árvore de decisão é uma estrutura hierárquica utilizada para tomar decisões ou fazer previsões com base em dados. Ela se assemelha a uma árvore invertida, onde cada nó interno representa uma característica ou atributo do conjunto de dados, cada aresta representa uma regra de decisão, e cada nó folha representa o resultado final ou a classe predita. O processo de tomada de decisão segue um caminho da raiz até uma folha, com base nas respostas às regras estabelecidas em cada nó [Somvanshi et al. 2016].

Para construir uma árvore de decisão, o algoritmo divide recursivamente o conjunto de dados em subconjuntos mais homogêneos, utilizando critérios como a entropia ou o índice de Gini, que medem a pureza dos nós. O objetivo é maximizar a separação entre as classes, de modo que as folhas contenham, idealmente, instâncias de uma única classe [Myles et al. 2004].

Uma das principais vantagens das árvores de decisão é a sua simplicidade e clareza. Elas são fáceis de entender e interpretar, mesmo para aqueles que não têm um conhecimento profundo de AM. Elas também lidam bem com variáveis categóricas e numéricas, tornando-as extremamente versáteis. Ao contrário de outros algoritmos, como as redes neurais, as árvores de decisão requerem pouco ou nenhum preprocessamento

de dados, como normalização ou padronização. Elas também não são sensíveis a dados faltantes, pois podem utilizar critérios de substituição ou dividir apenas os dados disponíveis [Somvanshi et al. 2016, Myles et al. 2004].

Um dos principais desafios das árvores de decisão é a sua tendência a se ajustarem excessivamente ao conjunto de treinamento, o que é conhecido como *overfitting*. Isso ocorre quando a árvore se torna muito complexa, aprendendo detalhes e ruídos específicos dos dados de treinamento, o que prejudica a sua capacidade de generalização para novos dados [Ying 2019]. Além disso, pequenas variações nos dados podem resultar em árvores de decisão significativamente diferentes. Isso pode gerar modelos instáveis, especialmente em conjuntos de dados pequenos ou ruidosos. Em casos onde as relações entre os atributos são complexas, as árvores de decisão podem não ser tão eficazes quanto outros algoritmos, como as redes neurais [Li and Belford 2002].

A aplicação de árvores de decisão na identificação de transações ilícitas, como lavagem de dinheiro ou financiamento de atividades ilegais, é uma área de grande interesse e relevância, especialmente no contexto das criptomoedas [Nerurkar et al. 2021]. O processo envolve o treinamento de um modelo de árvore de decisão com dados históricos de transações rotuladas como lícitas ou ilícitas. Os atributos utilizados podem incluir valores transacionados, frequências de transações, endereços de origem e destino, entre outros. Um exemplo prático pode ser feito com o conjunto de dados Elliptic++.

4.7.2.1. Conjunto de dados Elliptic++

Um conjunto de dados é uma coleção estruturada de dados que é utilizada em várias etapas de um projeto de aprendizado de máquina, desde o treinamento até a validação e teste de modelos. Geralmente consiste em linhas e colunas, onde cada linha representa uma instância (ou exemplo) e cada coluna representa uma característica ou variável (também chamada de *feature*). Em um contexto supervisionado, o dataset também inclui um rótulo ou classe associada a cada instância, que serve como a saída desejada que o modelo deve prever [Gong et al. 2023].

Os conjuntos de dados podem ser compostos por diferentes tipos de dados, como números, textos, imagens, ou até mesmo dados temporais. Por exemplo, em um problema de classificação de e-mails como "spam"ou "não spam", cada linha dos dados pode representar um e-mail, com colunas que incluem características como a frequência de certas palavras, a presença de anexos, e um rótulo indicando se o e-mail é ou não spam [Dada et al. 2019].

O sucesso de um modelo de aprendizado de máquina depende em grande parte da qualidade dos dados nos quais ele é treinado. Um conjunto de dados bem construído permite que o modelo aprenda padrões e relações significativas, melhorando sua capacidade de fazer previsões precisas em dados novos. Um bom dataset deve ser representativo do problema real que se deseja resolver. Isso significa que deve incluir uma variedade de exemplos e cobrir diferentes cenários possíveis [Dada et al. 2019].

O Elliptic++ é uma coleção rica e abrangente de transações de Bitcoin fornecida pela Elliptic, uma empresa especializada em análise e inteligência de blockchain. Este

conjunto de dados foi projetado para ajudar pesquisadores e profissionais a estudar e identificar atividades financeiras ilícitas no ecossistema de criptomoedas. O Elliptic++ inclui transações rotuladas que distinguem atividades legítimas de atividades ilegais, fornecendo uma base sólida para a construção de modelos de aprendizado de máquina voltados para a detecção de fraudes e outras atividades suspeitas [Elliptic 2024].

O Elliptic++ contém mais de 200.000 transações, das quais aproximadamente 2% são rotuladas como ilícitas, 21% como lícitas e o restante como não rotuladas. Cada transação é descrita por 166 atributos que capturam uma ampla gama de informações. Esses atributos incluem características temporais, como o tempo de criação da transação, características topológicas que descrevem a posição da transação na rede de transações e características locais que fornecem informações sobre o montante e a natureza das entradas e saídas da transação [Elliptic 2024, Weber et al. 2019].

Uma das vantagens do Elliptic++ é que ele foi cuidadosamente preparado e validado por especialistas em blockchain e finanças, garantindo a qualidade e a confiabilidade dos dados. As transações rotuladas foram identificadas com base em um extenso trabalho de investigação e análise realizado pela Elliptic, utilizando uma combinação de técnicas manuais e automatizadas [Elliptic 2024]. Isso significa que os modelos de AM treinados com este conjunto de dados podem aprender a distinguir padrões complexos de comportamento financeiro que são característicos de atividades ilícitas, como pode ser visto na próxima seção.

4.7.2.2. Classificação de transações usando o Elliptic++

Existem várias formas de trabalhar com algoritmos de aprendizado de máquina. Atualmente, Python se destaca como uma das principais linguagens para essa tarefa, especialmente devido à sua ampla adoção e ao rico ecossistema de bibliotecas voltadas para o aprendizado de máquina [Raschka et al. 2020]. A biblioteca scikit-learn, por exemplo, é amplamente reconhecida por sua simplicidade de uso, extensa documentação e vasta gama de algoritmos implementados, incluindo métodos de classificação, regressão e clustering [Scikit-learn 2024, Pedregosa et al. 2011]. Devido a esses fatores, optou-se por utilizar Python para aplicar algoritmos de árvore de decisão no Elliptic++, aproveitando-se assim das facilidades e robustez que a linguagem e suas bibliotecas proporcionam.

Inicialmente, o conjunto de dados Elliptic++ foi carregado utilizando a biblioteca Pandas. Nele há 4.545 transações ilícitas, 42.019 lícitas e 157.205 sem classificação, totalização 203.769 registros. Dado o grande número de atributos disponíveis, optou-se por simplificar o modelo, restringindo a análise apenas ao atributos relacionados à blockchain, com a exclusão das referentes às características locais e agregadas. Essa escolha foi feita para manter o foco em uma abordagem mais direta e menos complexa. Portanto o modelo foi construído com apenas 19 dos 184 atributos disponíveis.

Em seguida, foi gerada uma matriz de correlação para investigar as relações entre as variáveis selecionadas (Figura 4.16). A matriz de correlação, que mede a intensidade e a direção do relacionamento linear entre variáveis, permitiu identificar potenciais correlações fortes ou fracas, auxiliando na compreensão da estrutura dos dados.

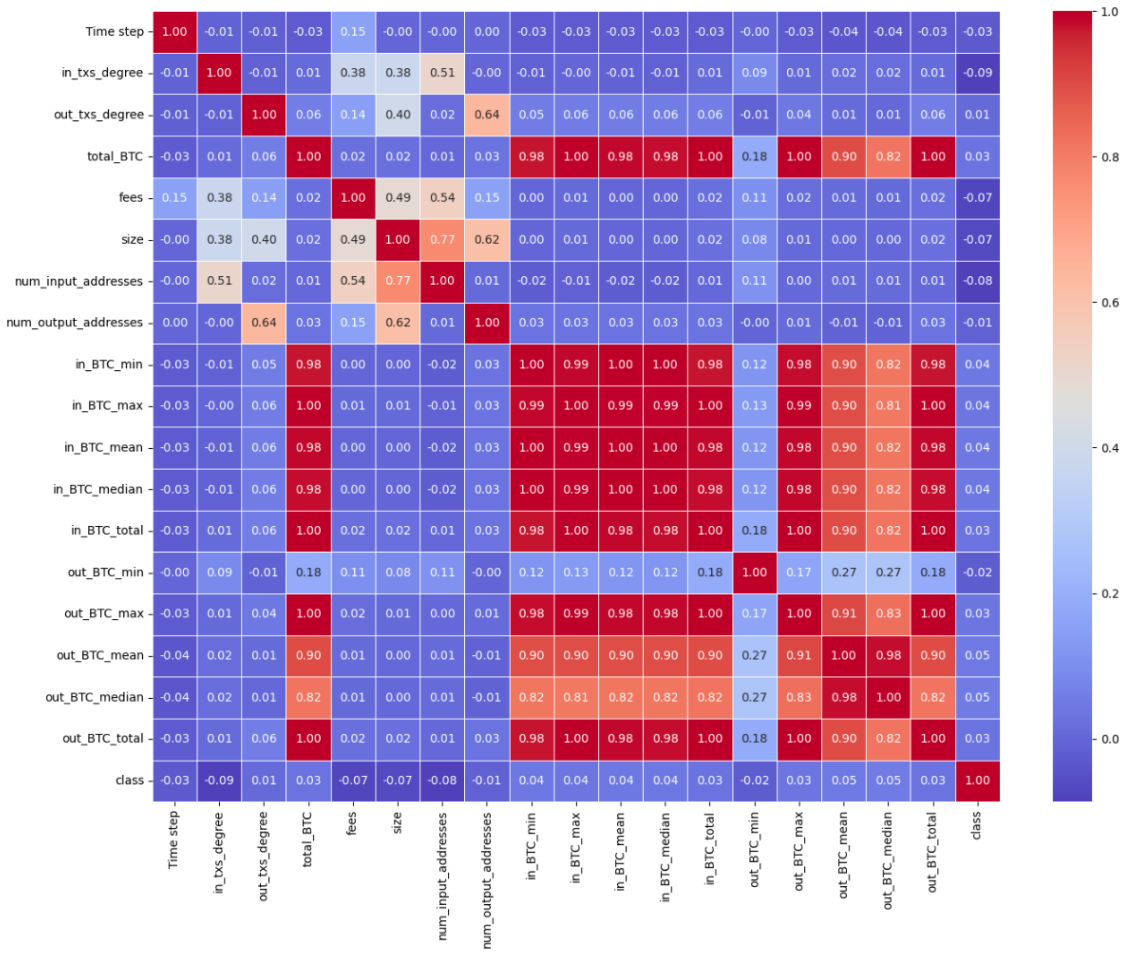

Figura 4.16: Matriz de correlação dos atributos selecionados para o modelo

Em uma matriz de correlação, a diagonal exibe sempre o valor 1 porque representa a correlação de cada variável consigo mesma. Essa auto-correlação é sempre perfeita, resultando em uma correlação de 1, pois qualquer variável tem uma relação linear perfeita com ela mesma. Os valores fora da diagonal mostram a correlação entre diferentes variáveis, variando de -1 a 1, onde valores próximos de 1 indicam uma forte correlação positiva e valores próximos de -1 indicam uma forte correlação negativa.

Também é possível avaliar a importância das características. Ela é uma métrica fundamental para entender como diferentes variáveis contribuem para a previsão feita por um modelo de aprendizado de máquina. Ela indica o grau em que cada característica influencia a decisão do modelo, ajudando a identificar quais variáveis têm maior impacto nas previsões. A Figura 4.17 mostra o resultado obtido para o experimento. O atributo "fees" (taxas), por exemplo, é um dos mais importantes para o modelo gerado.

Posteriormente, os dados foram divididos aleatoriamente em dois subconjuntos: 70% foram destinados ao treinamento do modelo, enquanto os 30% restantes foram reservados para testes. Essa divisão assegura uma avaliação adequada do modelo, garantindo que ele seja testado em dados não utilizados durante o treinamento.

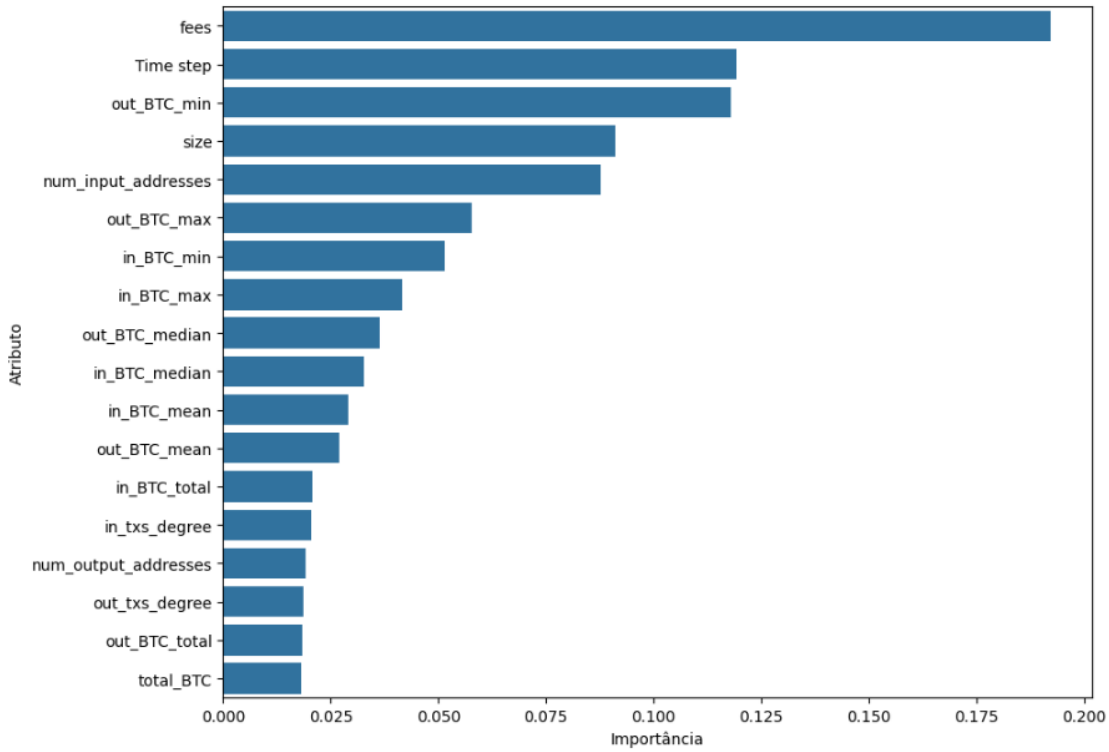

Figura 4.17: Importância dos atributos escolhidos para o modelo

Por fim, aplicou-se o algoritmo de árvore de decisão, disponibilizado pela biblioteca scikit-learn, aos dados de treinamento. O modelo foi avaliado utilizando o conjunto de teste, resultando em uma acurácia de 84,6%. Além da acurácia, outras métricas para o modelo simplificado são mostradas na Tabela 4.3.

Tabela 4.3: Relatório de Classificação

Classe	Precisão	Revocação	F1-Score	# de amostras
1	0.55	0.58	0.57	1289
2	0.66	0.68	0.67	12381
3	0.91	0.90	0.90	47172
Acurácia			0.85	60842
Média aritmética	0.71	0.72	0.71	60842
Média ponderada	0.85	0.85	0.85	60842

A precisão, que indica a proporção de previsões corretas entre todas as previsões feitas para cada classe, varia de 0.55 para a classe 1 (ilícito), 0.66 para a classe 2 (lícito) a 0.91 para a classe 3 (desconhecido), demonstrando uma maior exatidão na previsão da classe 3. A revocação, que mede a capacidade do modelo de identificar corretamente todas as instâncias de cada classe, segue uma tendência similar, com valores de 0.58 para a classe 1, 0.68 para a classe 2 e 0.90 para a classe 3. O F1-score, uma métrica que combina precisão e revocação em uma única medida, também reflete um desempenho mais forte

para a classe 3 (0.90) em comparação com as classes 1 e 2 (lícito), cujos f1-scores são 0.57 e 0.67, respectivamente. A acurácia geral do modelo é de 0.85, indicando que 85% das previsões feitas pelo modelo são corretas. As médias macro e ponderada das métricas de desempenho são 0.71 e 0.85, respectivamente, sugerindo um bom equilíbrio entre as classes e um desempenho robusto geral do modelo.

A matriz de confusão apresentada na Figura 4.18 mostra o desempenho do modelo de classificação. Na primeira classe, o modelo classificou corretamente 752 instâncias, mas cometeu 70 erros ao classificá-las como pertencentes à segunda classe e 467 erros ao classificá-las como pertencentes à terceira classe. Para a segunda classe, o modelo teve 8.384 acertos, mas errou 92 vezes ao classificá-las como primeira classe e 3.905 vezes como terceira classe. Na terceira classe, houve 42.334 acertos, com 513 erros ao classificá-las como primeira classe e 4.325 como segunda classe.

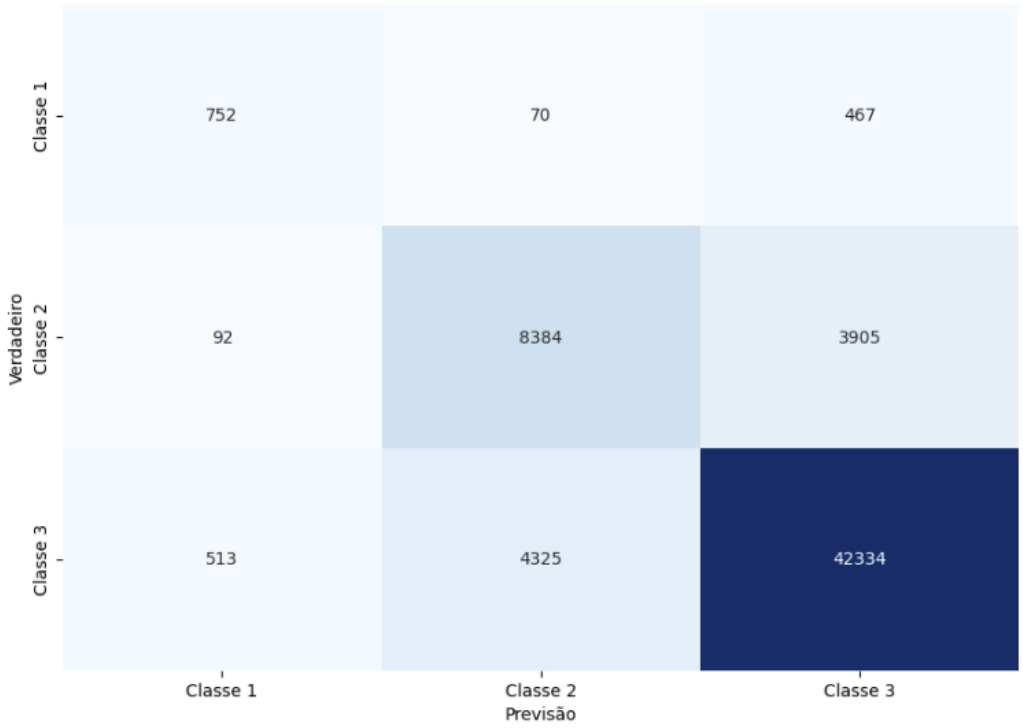

Figura 4.18: Matriz de confusão do modelo

Com base no modelo de previsão de transações ilícitas treinado com os dados do Elliptic++, seria possível desenvolver uma aplicação que recebe transações Bitcoin como entrada e fornece como saída a classificação da transação (lícita ou ilícita). Além disso, dado que o algoritmo escolhido foi uma árvore de decisão, as regras geradas pelo modelo poderiam ser apresentadas ao usuário do sistema, proporcionando transparência e explicabilidade nas decisões tomadas.

Caso o leitor se interesse, o código completo pode ser encontrado no repositório do GitHub "bitcoinforensics"[23], do NUSEC/FACOM - Núcleo de Segurança da Faculdade de Computação da Universidade Federal de Uberlândia.

[23] https://github.com/nusecfacom/bitcoinforensics

4.8. Conclusões

O Bitcoin foi concebido como uma moeda descentralizada que oferece um certo grau de pseudonimato aos seus usuários. No entanto, como a blockchain não fornece confidencialidade, diversas análises podem ser realizadas para identificar usuários ou determinar o tipo de transação efetuada. Usuários que necessitam de maior privacidade - por razões variadas, incluindo, mas não se limitando a atividades ilícitas - podem recorrer a técnicas avançadas para aumentar seu anonimato. Todavia, mesmo com tais precauções, o uso do Bitcoin ainda permite que análises sejam conduzidas, possibilitando a obtenção de informações sobre as transações.

Outras Criptomoedas Após a criação do Bitcoin, inúmeras criptomoedas surgiram e estão surgindo. Cada moeda apresenta suas peculiaridades que precisam serem estudadas e compreendidas. Em especial, a chamadas *privacycoins* - como Monero[24], Zcash[25] e Dash[26] - que apresentam privacidade no próprio projeto são um desafio especial;

Quantidade de Dados O tamanho das blockchain são da ordem de centenas de gigabytes, com milhões de transações e endereços. O uso de um sistema computacional, incluindo a escolha de algoritmos eficientes, é necessária para aplicações de mundo real;

Mixers Neste minicurso, apresentamos os primeiros passos das análises envolvendo *mixers*, em geral identificamos a ocorrências destas práticas. Tentar identificar entidades envolvidas e os possíveis caminhos complexos do fluxo do dinheiro é uma tarefa necessária e mais trabalhosa [Wang et al. 2023, Hong et al. 2018];

Camada 2 As aplicações de Camada 2 são soluções encontradas para resolver problemas de escalabilidade da blockchain [Gangwal et al. 2022, Seres et al. 2019]. Como o nome sugere, esta camada fica acima da blockchain e a suas operações não são registradas na blockchain, dificultando análise;

Contratos Inteligentes Os Contratos Inteligentes disponibilizados por algumas criptomoedas permitem que programa complexos sejam executados de forma descentralizada, compartilhando informações e fazendo transações financeiras [Szabo 1997, Wood et al. 2014]. Essas novas funcionalidades trazem inúmeras necessidades nas análises forense.

Referências

[els] El salvador adopted bitcoin as an official currency. https://insights.som.yale.edu/insights/el-salvador-adopted-bitcoin-as-an-official-currency-salvadorans-mostly-shrugged. Accessado: 2024-05-26.

[24] https://www.getmonero.org/
[25] https://z.cash/
[26] https://www.dash.org/pt-br/

[Ampel et al. 2023] Ampel, B., Otto, K., Samtani, S., and Chen, H. (2023). Disrupting ransomware actors on the bitcoin blockchain: A graph embedding approach. In *2023 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6.

[Andi 2021] Andi, H. K. (2021). An accurate bitcoin price prediction using logistic regression with lstm machine learning model. *Journal of Soft Computing Paradigm*, 3(3):205–217.

[Antonopoulos 2014] Antonopoulos, A. M. (2014). *Mastering Bitcoin: Unlocking Digital Crypto-Currencies*. O'Reilly Media, Inc., 1st edition.

[Aponte-Novoa et al. 2021] Aponte-Novoa, F. A., Orozco, A. L. S., Villanueva-Polanco, R., and Wightman, P. (2021). The 51*IEEE Access*, 9:140549–140564.

[Auti et al. 2022] Auti, A., Patil, D., Zagade, O., Bhosale, P., and Ahire, P. (2022). Bitcoin price prediction using svm. *Int. J. Eng. Appl. Sci. Technol*, 6(11):226–229.

[Bean 2011] Bean, H. (2011). Is open source intelligence an ethical issue? In *Government Secrecy*, volume 19, pages 385–402. Emerald Group Publishing Limited.

[Benford 1938] Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4):551–572.

[Block 2023] Block, L. (2023). The long history of osint. *Journal of Intelligence History*, pages 1–15.

[Böhm and Lolagar 2021] Böhm, I. and Lolagar, S. (2021). Open source intelligence: Introduction, legal, and ethical considerations. *International Cybersecurity Law Review*, 2:317–337.

[Chainalysis 2020] Chainalysis (2020). Tchainalysis in action: Us government agencies seize more than $1 billion in cryptocurrency connected to infamous darknet market silk road.

[Chainalysis 2023] Chainalysis (2023). *Chainalysis Reactor*. Data de Acesso: 27/11/2023.

[Chong et al. 2021] Chong, B. et al. (2021). K-means clustering algorithm: a brief review. *vol*, 4:37–40.

[Christin 2012] Christin, N. (2012). Traveling the silk road: A measurement analysis of a large anonymous online marketplace.

[CipherTrace 2023a] CipherTrace (2023a). Crypto crimes & anti-money laundering (aml) report march 2023.

[CipherTrace 2023b] CipherTrace (2023b). *Inspector*. Data de Acesso: 27/11/2023.

[Costa et al. 2012] Costa, P. R., Souza, F. F., Times, V. C., and Benevenuto, F. (2012). Towards integrating online social networks and business intelligence. In *Proceedings of the international conferences web based communities and social media*, pages 21–32.

[Dada et al. 2019] Dada, E. G., Bassi, J. S., Chiroma, H., Adetunmbi, A. O., Ajibuwa, O. E., et al. (2019). Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, 5(6).

[Dalal et al.] Dalal, S., Wang, Z., and Sabharwal, S. Identifying ransomware actors in the bitcoin network.

[Dalal et al. 2021] Dalal, S. R., Wang, Z., and Sabharwal, S. (2021). Identifying ransomware actors in the bitcoin network. *ArXiv*, abs/2108.13807.

[Dokuz et al. 2020] Dokuz, A. Ş., Çelik, M., and Ecemiş, A. (2020). Anomaly detection in bitcoin prices using dbscan algorithm. *European Journal of Science and Technology*, 2020:436–443.

[Dolejška et al. 2023] Dolejška, D., Koutenský, M., Veselý, V., and Pluskal, J. (2023). Busting up monopoly: Methods for modern darknet marketplace forensics. *Forensic Science International: Digital Investigation*, 46.

[dos Reis et al. 2023] dos Reis, E. F., Teytelboym, A., ElBahraw, A., Loizaga, I. D., and Baronchelli, A. (2023). Identifying key players in dark web marketplaces.

[Dudani et al. 2023] Dudani, S., Baggili, I., Raymond, D., and Marchany, R. (2023). The current state of cryptocurrency forensics. *Forensic Science International: Digital Investigation*, 46:301576.

[Elliptic 2023] Elliptic (2023). *Investigator*. Data de Acesso: 27/11/2023.

[Elliptic 2024] Elliptic (2024). The elliptic dataset: Cryptocurrency and financial crime.

[Ermilov et al. 2017] Ermilov, D., Panov, M., and Yanovich, Y. (2017). Automatic bitcoin address clustering. *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 461–466.

[Eyal and Sirer 2014] Eyal, I. and Sirer, E. G. (2014). Majority is not enough: Bitcoin mining is vulnerable. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8437:436–454.

[Gangwal et al. 2022] Gangwal, A., Gangavalli, H. R., and Thirupathi, A. (2022). A survey of layer-two blockchain protocols.

[Gong et al. 2023] Gong, Y., Liu, G., Xue, Y., Li, R., and Meng, L. (2023). A survey on dataset quality in machine learning. *Information and Software Technology*, 162:107268.

[Hamerly and Elkan 2003] Hamerly, G. and Elkan, C. (2003). Learning the k in k-means. *Advances in neural information processing systems*, 16.

[Harrigan and Fretter 2017] Harrigan, M. and Fretter, C. (2017). The unreasonable effectiveness of address clustering. *Proceedings - 13th IEEE International Conference on Ubiquitous Intelligence and Computing, 13th IEEE International Conference on Advanced and Trusted Computing, 16th IEEE International Conference on Scalable Computing and Communications, IEEE Internationa*, pages 368–373.

[Hiramoto and Tsuchiya 2020] Hiramoto, N. and Tsuchiya, Y. (2020). Measuring dark web marketplaces via bitcoin transactions: From birth to independence. *Forensic Science International: Digital Investigation*, 35.

[Hong et al. 2018] Hong, Y., Kwon, H., Lee, J., and Hur, J. (2018). A practical demixing algorithm for bitcoin mixing services. In *BCC 2018 - Proceedings of the 2nd ACM Workshop on Blockchains, Cryptocurrencies, and Contracts, Co-located with ASIA CCS 2018*, pages 15–20. Association for Computing Machinery, Inc.

[Hwang et al. 2022] Hwang, Y.-W., Lee, I.-Y., Kim, H., Lee, H., Kim, D., et al. (2022). Current status and security trend of osint. *Wireless Communications and Mobile Computing*, 2022.

[Hyman 2015] Hyman, M. (2015). Bitcoin atm: A criminal's laundromat for cleaning money. . *Thomas L. Rev.*, 27:296.

[Ishikawa 2017] Ishikawa, M. (2017). Designing Virtual Currency Regulation in Japan: Lessons from the Mt Gox Case. *Journal of Financial Regulation*, 3(1):125–131.

[Johnson et al. 2001] Johnson, D., Menezes, A., and Vanstone, S. (2001). The elliptic curve digital signature algorithm (ecdsa). *Int. J. Inf. Secur.*, 1(1):36–63.

[Juodis et al. 2024] Juodis, M., Filatovas, E., and Paulavičius, R. (2024). Overview and empirical analysis of wealth decentralization in blockchain networks. *ICT Express*, 10(2):380–386.

[Kanemura et al. 2019] Kanemura, K., Toyoda, K., and Ohtsuki, T. (2019). Identification of darknet markets' bitcoin addresses by voting per-address classification results. pages 154–158. Institute of Electrical and Electronics Engineers Inc.

[Khanum et al. 2015] Khanum, M., Mahboob, T., Imtiaz, W., Ghafoor, H. A., and Sehar, R. (2015). A survey on unsupervised machine learning algorithms for automation, classification and maintenance. *International Journal of Computer Applications*, 119(13).

[Li and Belford 2002] Li, R.-H. and Belford, G. G. (2002). Instability of decision tree classification algorithms. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 570–575.

[Li et al. 2020a] Li, S.-N., Yang, Z., and Tessone, C. J. (2020a). Mining blocks in a row: A statistical study of fairness in bitcoin mining. In *2020 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, pages 1–4.

[Li et al. 2020b] Li, S.-N., Yang, Z., and Tessone, C. J. (2020b). Proof-of-work cryptocurrency mining: a statistical approach to fairness. In *2020 IEEE/CIC International Conference on Communications in China (ICCC Workshops)*, pages 156–161.

[Loporchio 2023] Loporchio, Matteo; Bernasconi, A. D. F. M. D. R. L. (2023). Is bitcoin gathering dust? an analysis of low-amount bitcoin transactions. *Applied Network Science*, 8(2364-8228).

[Mahesh 2020] Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9(1):381–386.

[Meiklejohn et al. 2013] Meiklejohn, S., Pomarole, M., Jordan, G., Levchenko, K., McCoy, D., Voelker, G. M., and Savage, S. (2013). A fistful of bitcoins: Characterizing payments among men with no names. pages 127–140.

[Milad et al. 2024] Milad, M., Ovezik, C., Karakostas, D., and Woods, D. W. (2024). Statistical confidence in mining power estimates for pow blockchains. In *Companion Proceedings of the ACM on Web Conference 2024*, WWW '24, page 1752–1760, New York, NY, USA. Association for Computing Machinery.

[Molitor et al. 2023] Molitor, D., Raghupathi, W., Raghupathi, V., and Saharia, A. (2023). Understanding cryptocurrency: A descriptive analytics study of bitcoin. *International Journal of Blockchain Applications and Secure Computing*, 1:1–25.

[Myles et al. 2004] Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., and Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18(6):275–285.

[Nakamoto 2009] Nakamoto, S. (2009). Bitcoin: A peer-to-peer electronic cash system.

[Narayanan et al. 2016] Narayanan, A., Bonneau, J., Felten, E., Miller, A., and Goldfeder, S. (2016). *Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction*. Princeton University Press.

[Narayanan and Clark 2017] Narayanan, A. and Clark, J. (2017). Bitcoin's academic pedigree. *Communications of the ACM*, 60:36–45.

[Nerurkar et al. 2021] Nerurkar, P., Bhirud, S., Patel, D., Ludinard, R., Busnel, Y., and Kumari, S. (2021). Supervised learning model for identifying illegal activities in bitcoin. *Applied Intelligence*, 51:3824–3843.

[Noll 2023] Noll, F. (2023). The controversial business of cash-to-crypto bitcoin atms. *Payments System Research Briefing*.

[P et al. 2016] P, R., CS, P., and M., B. (2016). Common pitfalls in statistical analysis: The perils of multiple testing. *Perspect Clin Res.*, 7:106–107.

[Pedregosa et al. 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

[Phaladisailoed and Numnonda 2018] Phaladisailoed, T. and Numnonda, T. (2018). Machine learning models comparison for bitcoin price prediction. In *2018 10th international conference on information technology and electrical engineering (ICITEE)*, pages 506–511. IEEE.

[Raschka et al. 2020] Raschka, S., Patterson, J., and Nolet, C. (2020). Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *Information*, 11(4):193.

[Rathan et al. 2019] Rathan, K., Sai, S. V., and Manikanta, T. S. (2019). Crypto-currency price prediction using decision tree and regression techniques. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 190–194. IEEE.

[Salisu et al. 2023] Salisu, S., Filipov, V., and Pene, B. (2023). Blockchain forensics: A modern approach to investigating cybercrime in the age of decentralisation. In *Proceedings of the 18th International Conference on Cyber Warfare and Security*.

[Schnoering and Vazirgiannis 2023] Schnoering, H. and Vazirgiannis, M. (2023). Heuristics for detecting coinjoin transactions on the bitcoin blockchain.

[Scikit-learn 2024] Scikit-learn (2024). Scikit-learn: Machine learning in python.

[Sendin 2018] Sendin, I. d. S. (2018). On detecting cold storage transactions on bitcoin's blockchain. In *Anais do XVIII Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*, pages 155–166, Porto Alegre, RS, Brasil. SBC.

[Seres et al. 2019] Seres, I. A., Gulyás, L., a. Nagy, D., and Burcsi, P. (2019). Topological analysis of bitcoin's lightning network. pages 1–7.

[Shah et al. 2021] Shah, A., Chauhan, Y., and Chaudhury, B. (2021). Principal component analysis based construction and evaluation of cryptocurrency index. *Expert systems with applications*, 163:113796.

[Singh et al. 2016] Singh, A., Thakur, N., and Sharma, A. (2016). A review of supervised machine learning algorithms. In *2016 3rd international conference on computing for sustainable global development (INDIACom)*, pages 1310–1315. Ieee.

[Somvanshi et al. 2016] Somvanshi, M., Chavan, P., Tambade, S., and Shinde, S. (2016). A review of machine learning techniques using decision tree and support vector machine. In *2016 international conference on computing communication control and automation (ICCUBEA)*, pages 1–7. IEEE.

[Steele 1990] Steele, R. D. (1990). Intelligence in the 1990's: Recasting national security in a changing world. *American Intelligence Journal*, 11(3):29–36.

[Stütz et al. 2022] Stütz, R., Stockinger, J., Moreno-Sanchez, P., Haslhofer, B., and Maffei, M. (2022). Adoption and actual privacy of decentralized coinjoin implementations in bitcoin. pages 254–267. Association for Computing Machinery (ACM).

[Szabo 1997] Szabo, N. (1997). Formalizing and securing relationships on public networks. *First Monday*, 2.

[Tabatabaei and Wells 2016] Tabatabaei, F. and Wells, D. (2016). Osint in the context of cyber-security. *Open Source Intelligence Investigation: From Strategy to Implementation*, pages 213–231.

[Team 2024] Team, C. (2024). Money laundering and cryptocurrency. trends and new techniques for detection and investigation. Technical report, Chainalysis.

[Tironsakkul et al. 2022] Tironsakkul, T., Maarek, M., Eross, A., and Just, M. (2022). The unique dressing of transactions: Wasabi coinjoin transaction detection. In *Proceedings of the 2022 European Interdisciplinary Cybersecurity Conference*, EICC '22, page 21–28, New York, NY, USA. Association for Computing Machinery.

[Torrez 2023] Torrez, A. S. (2023). El rastro virtual de las criptomonedas.

[Trautman 2014] Trautman, L. J. (2014). Virtual currencies; bitcoin & what now after liberty reserve, silk road, and mt. gox? *Richmond Journal of Law and Technology*, 20(4).

[Vidal-Tomás 2022] Vidal-Tomás, D. (2022). Which cryptocurrency data sources should scholars use? *International Review of Financial Analysis*, 81:102061.

[Vičič and Tošić 2022] Vičič, J. and Tošić, A. (2022). Application of benford's law on cryptocurrencies. *Journal of Theoretical and Applied Electronic Commerce Research*, 17.

[Wang et al. 2018] Wang, Y., Li, F., Hu, J., and Zhuang, D. (2018). K-means algorithm for recognizing fraud users on a bitcoin exchange platform. In *Proceedings of the 18th International Conference on Electronic Business*.

[Wang et al. 2023] Wang, Z., Chaliasos, S., Qin, K., Zhou, L., Gao, L., Berrang, P., Livshits, B., and Gervais, A. (2023). On how zero-knowledge proof blockchain mixers improve, and worsen user privacy. In *ACM Web Conference 2023 - Proceedings of the World Wide Web Conference, WWW 2023*, pages 2022–2032. Association for Computing Machinery, Inc.

[Weber et al. 2019] Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., and Leiserson, C. E. (2019). Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics. *arXiv preprint arXiv:1908.02591*.

[Wood et al. 2014] Wood, G. et al. (2014). Ethereum: A secure decentralised generalised transaction ledger. *Ethereum project yellow paper*, 151(2014):1–32.

[Wu et al. 2021a] Wu, L., Hu, Y., Zhou, Y., Wang, H., Luo, X., Wang, Z., Zhang, F., and Ren, K. (2021a). Towards understanding and demystifying bitcoin mixing services. In *Proceedings of the Web Conference 2021*, pages 33–44.

[Wu et al. 2021b] Wu, L., Hu, Y., Zhou, Y., Wang, H., Luo, X., Wang, Z., Zhang, F., and Ren, K. (2021b). Towards understanding and demystifying bitcoin mixing services. In *Proceedings of the World Wide Web Conference*. The Web Conference.

[Xi et al. 2020] Xi, H., Fan, Z., Shenwen, L., Hongliang, M., and Ketai, H. (2020). A review on data analysis of bitcoin transaction entity. In *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pages 159–164.

[Xi et al.] Xi, H., Ketai, H., Shenwen, L., Jinglin, Y., and Hongliang, M. Bitcoin address clustering method based on multiple heuristic conditions. *IET Blockchain*, 2:44–56.

[Xia et al. 2020] Xia, P., Wang, H., Zhang, B., Ji, R., Gao, B., Wu, L., Luo, X., and Xu, G. (2020). Characterizing cryptocurrency exchange scams. *Computers & Security*, 98:101993.

[Yeboah-Ofori and Brimicombe 2018] Yeboah-Ofori, A. and Brimicombe, A. (2018). Cyber intelligence and osint: Developing mitigation techniques against cybercrime threats on social media. *International Journal of Cyber-Security and Digital Forensics (IJCSDF)*, 7(1):87–98.

[Ying 2019] Ying, X. (2019). An overview of overfitting and its solutions. In *Journal of physics: Conference series*, volume 1168, page 022022. IOP Publishing.

[Zhang et al. 2020] Zhang, Y., Wang, J., and Luo, J. (2020). Heuristic-based address clustering in bitcoin. *IEEE Access*, 8:210582–210591.

[Ádám Ficsór et al. 2021] Ádám Ficsór, Kogman, Y., Ontivero, L., and Seres, I. A. (2021). WabiSabi: Centrally coordinated CoinJoins with variable amounts. Cryptology ePrint Archive, Paper 2021/206. https://eprint.iacr.org/2021/206.