SmartSet: Um Dataset de Contratos Inteligentes da Rede Blockchain Polygon

Josué N. Campos[1], **Ronan M. Dutra**[1], **Alex B. Vieira**[2], **José A. M. Nacif**[1]

[1]Instituto de Ciências Exatas e Tecnológicas, *Campus* UFV-Florestal
Universidade Federal de Viçosa (UFV) – Florestal, MG – Brazil

[2]Departamento de Informática, Universidade Federal de Juiz de Fora
(UFJF) – Juiz de Fora, MG – Brazil

{josue.campos,ronan.dutra,jnacif}@ufv.br, alex.borges@ufjf.edu.br

Abstract. *Blockchain networks have been gaining popularity in recent years. With the arrival of smart contracts and the Ethereum network, several sidechains, such as the Polygon network, are being used as alternatives for low transaction costs and speed gains in block mining. Because of this, several studies are emerging to ensure that existing smart contracts in Blockchain networks are free of vulnerabilities. In this paper, we develop a dataset of smart contracts written in the Solidity language, taken directly from the Polygon network. As a result, we obtained a total sample of 156,250 contracts, of which 4,708 are verified by the network and 151,542 are unverified. Furthermore, we extracted these contracts during 1.2 days, which can serve as a basis for analysis and creation of models that can improve the search for vulnerabilities and facilitate the development of secure smart contracts.*

Resumo. *As redes Blockchain vêm ganhando popularidade nos últimos anos. Com a chegada dos contratos inteligentes e da rede Ethereum, diversas sidechains, como a rede Polygon, estão sendo utilizadas como alternativas por baixo custo de transações e ganho de velocidade na mineração de blocos. Dessa forma, diversas pesquisas estão surgindo para garantir que os contratos inteligentes existentes nas redes Blockchain sejam livres de vulnerabilidades. Neste trabalho, nós desenvolvemos um conjunto de dados de contratos inteligentes escritos na linguagem Solidity, retirados diretamente da rede Polygon. Nós obtivemos uma amostra de 4.708 contratos inteligentes verificados pela rede e 151.542 contratos não verificados, totalizando um número de 156.250 códigos extraídos durante 1,2 dias que poderão servir como base de análise e criação de modelos que possam aprimorar a busca por vulnerabilidades e facilitar o desenvolvimento de contratos inteligentes seguros.*

1. Introdução

Com o advento da criptomoeda Bitcoin, a tecnologia Blockchain ganhou visibilidade nos últimos anos como uma maneira descentralizada de registrar transações imutáveis entre usuários anônimos [Nakamoto 2008]. A partir do Bitcoin, diversas outras criptomoedas e tecnologias surgiram, principalmente com a chegada da rede Ethereum e da expansão do uso da Blockchain para além do registro de transações financeiras ponto-a-ponto [Buterin et al. 2014]. Por meio dos contratos inteligentes, diversas aplicações

voltadas para a Blockchain foram criadas, como os Tokens hospedados na rede Ethereum e os Tokens Não-Fungíveis (NFTs). Além disso, pelo fato do custo por transações atingirem, na maioria das vezes, valores altos, estão sendo criadas Blockchains associadas à Ethereum, mas com seus próprios mecanismos de consenso, tokens e contratos inteligentes, que são as chamadas Sidechains [Singh et al. 2020].

Apesar destas tecnologias emergentes possuírem um alto nível de criptografia e anonimidade, os contratos inteligentes podem estar suscetíveis a vulnerabilidades de segurança, assim como contratos tradicionais. Dessa maneira, diversos pesquisadores estão criando soluções para garantir que, uma vez que um contrato inteligente esteja na Blockchain, ele possa estar livre dos ataques já mapeados ao longo dos anos. Por exemplo, por meio do site *Smart Contract Weakness Classification Registry*[1] são elencadas as principais vulnerabilidades de segurança relacionadas aos contratos inteligentes, servindo como base para desenvolvedores e pesquisadores. Adicionalmente, ferramentas de análise estática e dinâmica dos códigos de contratos inteligentes estão sendo desenvolvidas e aprimoradas, visando automatizar a procura por vulnerabilidades e auxiliando desenvolvedores durante a criação de novas aplicações [Kushwaha et al. 2022].

Nesse sentido, o objetivo deste trabalho é apresentar um conjunto de dados de contratos inteligentes retirados diretamente da rede Polygon [Kanani et al. 2021], uma das Sidechains da rede Ethereum utilizadas principalmente pelo fato de possuir baixo custo monetário por transação. O *dataset* proposto pode ser útil para auxiliar a criação e aprimoramento das ferramentas de análise de contratos inteligentes, e também pode servir como base para modelos de aprendizado de máquina de classificação de contratos. Dessa maneira, os desenvolvedores poderão consultar os códigos já existentes na rede e criar soluções cada vez mais seguras para estas redes descentralizadas.

O restante deste artigo foi organizado da seguinte forma: Na seção 2 são apresentados os principais conceitos associados ao conjunto de dados. Já na seção 3 são elencados os principais trabalhos relacionados na literatura a respeito da criação de conjunto de dados a partir de redes Blockchain e análises de segurança de contratos inteligentes. Na seção 4 é apresentado todo o processo de desenvolvimento do trabalho, para que nas seções 5 e 6, respectivamente, sejam elencados os resultados obtidos e a conclusão do trabalho.

2. Visão Geral

Esta seção apresenta os principais conceitos sobre o conjunto de dados construído, em uma visão geral sobre Blockchain, sidechains e contratos inteligentes. Inicialmente, são discutidas as definições dos conceitos abordados e, posteriormente, suas associações com o presente trabalho.

2.1. Blockchain

A blockchain é um livro-razão imutável e distribuído capaz de armazenar registros de transações de maneira descentralizada [Bhutta et al. 2021]. Estas transações são armazenadas em blocos interligados, de modo que o *hash* do próximo bloco é calculado com base no bloco anterior. Sendo assim, a blockchain tem a capacidade de gerar uma espécie

[1] `https://swcregistry.io/`

de histórico das transações. O primeiro bloco da cadeia é chamado de bloco gênesis, pois a partir dele os próximos blocos podem ser anexados na cadeia.

A Figura 1 apresenta a estrutura de um bloco. Cada bloco é dividido entre cabeçalho e corpo. No cabeçalho constam os dados principais de identificação de um bloco, como, por exemplo, o *hash* do bloco e o *hash* do bloco anterior, servindo como um apontador e criando a cadeia. Além disso, têm-se também o *timestamp* em que o bloco foi anexado na cadeia, o valor de *nonce* que representa a comprovação do esforço realizado pelo algoritmo de consenso para anexar o bloco na blockchain. Por fim, têm-se a *merkle root* que atua armazenando o valor de *hash* de todas as transações validadas do bloco. Já no corpo de um bloco constam, principalmente, a lista de transações realizadas, validadas e armazenadas permanentemente.

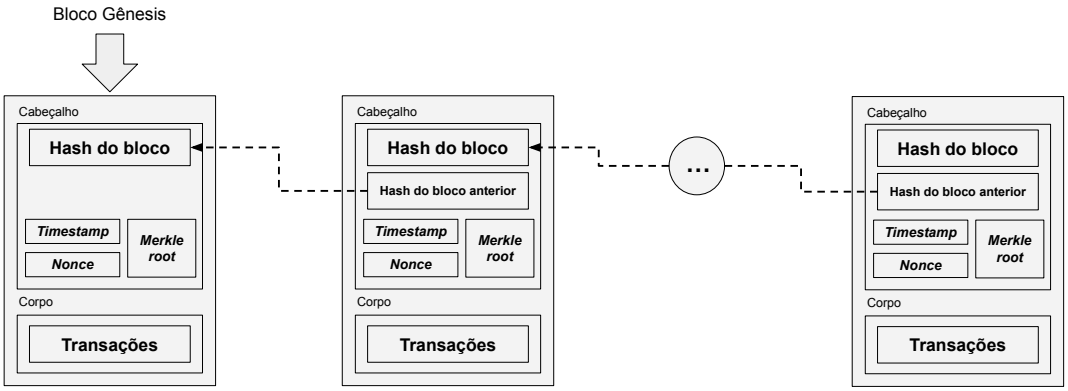

Figura 1. Estrutura básica de uma blockchain.

2.2. Sidechains

As sidechains são redes blockchains que validam dados de outras blockchains [Back et al. 2014]. À medida que novas blockchains e criptomoedas foram criadas, o mercado e o desenvolvimento entrou em um estado de fragmentação. Por tal motivo, as sidechains atuam integrando blockchains já existentes, isto é, ao invés de alterar uma blockchain inteiramente por conta de uma funcionalidade, passa a ser mais vantagem criar uma sidechain com esta funcionalidade que comunica com a blockchain pai. Cada sidechain pode possuir sua arquitetura independentemente da blockchain pai, ou seja, cada blockchain pode possuir seu próprio algoritmo de consenso, seus próprios blocos e seus próprios tokens.

Neste trabalho utilizamos a rede Polygon para extração dos contratos e construção do *dataset*, visto que é uma das sidechains mais utilizadas atualmente. Conforme os gráficos da Figura 2 retirados diretamente dos *scans* das blockchains[2][3], podemos observar o crescimento desta sidechain em comparação com a rede Ethereum ao longo do ano de 2022. Além disso, a rede Polygon destaca-se pelo seu baixo custo de *gas* em relação a

[2]`https://polygonscan.com/`
[3]`https://etherscan.io/`

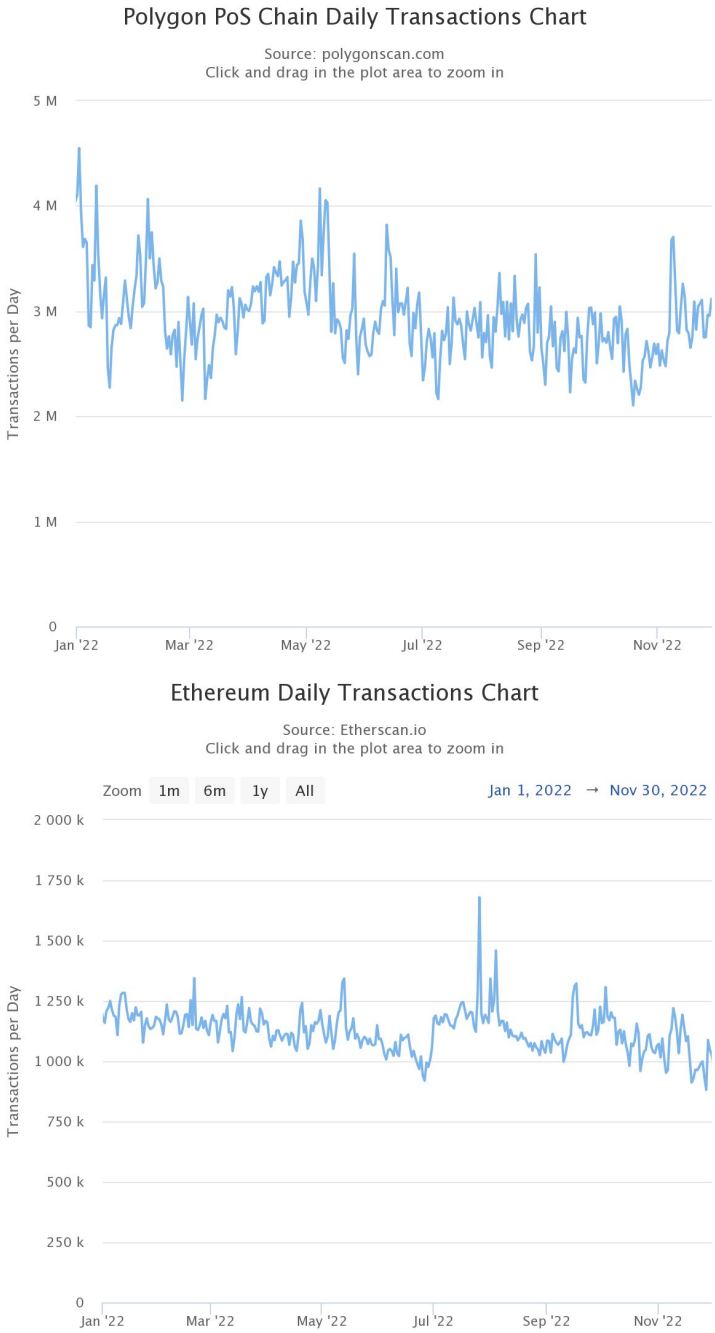

Figura 2. Gráfico de transações diárias das redes Polygon e Ethereum.

Ethereum, uma espécie de taxa cobrada por cada transação. A razão deste menor custo é devido a uma das vantagens das sidechains que é a utilização de um próprio token que, no caso da Polygon, faz uso da criptomoeda MATIC.

2.3. Contratos Inteligentes

Os contratos inteligentes (*smart contracts*) são programas que são executados automaticamente na blockchain, desde que as condições para uma de suas operações sejam previamente atendidas [Hewa et al. 2021]. Pelo fato de utilizarem a blockchain, os contratos

inteligentes possuem a característica de dispensarem a utilização de uma terceira entidade regulamentadora, isto é, desde que o contrato seja acordado entre ambas as partes, o seu uso se dará de forma descentralizada e automática.

```
contract MyToken{
    address owner;

    constructor() {
        owner = msg.sender;
    }

    function getOwner() external view returns (address) {
        return owner;
    }
}
```

Código 1: Estrutura básica de um contrato inteligente escrito na linguagem Solidity.

Por meio da popularização da rede Ethereum, diversas aplicações foram criadas utilizando os contratos inteligentes. Pelo fato da sua natureza possibilitar a construção de operações mais complexas na blockchain em vista de uma simples troca de moedas, aplicações financeiras e médicas, monitoramento e acordos contratuais puderam ser desenvolvidos, estendendo até os Tokens Não-Fungíveis. No entanto, a construção de um contrato inteligente deve ser bem definida tanto pelas partes envolvidas, quanto pelos desenvolvedores que o codificam. Em seu desenvolvimento, é feito o uso da linguagem Solidity (Código 1) e, portanto, assim como em qualquer sistema desenvolvido, questões relacionadas à segurança e prevenção de ataques devem ser levadas em consideração.

3. Trabalhos Relacionados

Recentemente, [Oliva et al. 2020] propuseram um estudo sobre os contratos inteligentes da rede Ethereum, classificando-os nos aspectos de frequência de acesso, funcionalidade e complexidade. Neste trabalho, nós propomos um conjunto de dados de contratos inteligentes capaz de ser analisado e classificado, porém levando em consideração a sidechain Polygon e seu baixo custo por transação.

Mesmo os contratos verificados pelas redes blockchains podem estar suscetíveis a vulnerabilidades de segurança e, por tal motivo, diversas ferramentas de análise estática e dinâmica de contratos inteligentes podem ser utilizadas para garantir um código bem construído. Em [Durieux et al. 2020] é proposto uma revisão empírica destas ferramentas de análise, por meio de um *dataset* de 47.587 contratos extraídos da rede Ethereum. Com a construção de um conjunto de dados expandidos para diversas blockchains e sidechains, as ferramentas podem ser aprimoradas para diferentes casos de uso, auxiliando um número significativo de desenvolvedores de contratos inteligentes.

Além do viés de segurança, os contratos inteligentes podem levantar questões éticas e legais que devem ser analisadas e pensadas antes de serem colocados nas redes blockchain. Consequentemente, em [Gregoriadis et al. 2022] são analisados os conteúdos de sistemas baseados nas blockchains do Bitcoin e da Ethereum utilizando a plataforma

BigQuery. Por meio do conjunto de dados proposto neste trabalho, por se tratar de uma *sidechain* pública, diversas análises nesse sentido podem ser realizadas e testadas.

4. Metodologia

Nesta seção é relatado o processo de construção do *dataset*, desde a obtenção da base de dados até a extração dos códigos da rede Polygon. A obtenção dos dados baseia-se na utilização da ferramenta Google Cloud BigQuery para serem retornados os endereços dos contratos inteligentes alocados na rede. Por sua vez, a extração dos códigos utiliza-se destes endereços para serem realizadas requisições na API da *sidechain*, obtendo-se assim o código-fonte do contrato.

4.1. Google Cloud BigQuery

A plataforma de armazenamento em nuvem Google BigQuery[4] é um *data warehouse* que permite a análise massiva de dados. Por meio desta ferramenta, pessoas e empresas podem facilmente consultar e analisar dados de diversos locais diferentes. Neste trabalho nós utilizamos o conjunto de dados da Polygon hospedado na plataforma BigQuery. Este conjunto de dados reúne diversas tabelas contendo dados reais e atualizados em tempo real de acordo com as transações que ocorrem na blockchain Polygon.

Dentre as tabelas presentes no conjunto de dados da Polygon, nós extraímos a que armazena as informações dos contratos inteligentes. Nela podem ser obtidos dados como: o endereço do contrato, seu *bytecode*, data e hora do bloco, código *hash* do bloco e indicadores que dizem se o contrato inteligente enquadra nos padrões de contratos Ethereum ERC-20 e ERC-721. Para exportar os dados necessários, a plataforma BigQuery permite consultas escritas na linguagem SQL, conforme o Código 2. A partir da amostra coletada, retiramos apenas o endereço do contrato, já que é o atributo necessário para extração do código-fonte na API da blockchain.

```
SELECT address
FROM `public-data-finance.crypto_polygon.contracts`;
```

Código 2: Consulta realizada na plataforma BigQuery.

4.2. Polygon API

A API da rede Polygon está intimamente ligada à plataforma de Scan da sidechain. O site PolygonScan provê recursos de visualização de todas as transações da rede, exibições analíticas por meio de gráficos, bem como sua API[5] para extração de dados da rede. Nesse sentido, utilizamos o método da API que retorna os dados de contratos inteligentes verificados na blockchain.

Para a requisição ser realizada, passamos o endereço de cada contrato obtido pela plataforma BigQuery, de maneira que os endereços de contratos verificados retornem o código-fonte e os endereços de contratos não verificados retornem uma requisição vazia, já que a API não disponibiliza o código-fonte destes contratos inteligentes. No Código 3 consta todo o processo de obtenção dos códigos utilizando a linguagem Python para realizar as requisições.

[4] https://cloud.google.com/bigquery
[5] https://docs.polygonscan.com/

```
def getContractSourceCode(apiKey, address):
    params = {
        "module": "contract",
        "action" : "getsourcecode",
        "address" : address,
        "apikey" : apiKey
    }
    response = requests.get(
        "https://api.polygonscan.com/api",
        params=params
    )
    data = response.json()
    if data["message"] == "OK":
        result = data["result"]
        sourceCode = result[0]["SourceCode"]
        return sourceCode
    return None
```

Código 3: Código que requisita na API o código-fonte de um contrato inteligente.

5. Resultados

Nós extraímos os códigos dos contratos inteligentes verificados pela Sidechain Polygon, utilizando os endereços dos contratos na rede como índice de busca. Os contratos verificados são aqueles que passam por uma avaliação do código-fonte, realizada pela própria Blockchain. O código compilado do contrato inteligente é comparado com o código presente na rede e, posteriormente, torna-se público para todos. Já os contratos inteligentes não verificados são aqueles que não passam por este processo, logo, a API das Blockchains não são capazes de retornarem o código-fonte dos mesmos. Na Tabela 1 é possível observar a amostra de endereços extraídos da plataforma Google BigQuery. A amostra extraída representa apenas 0,094% de toda a base de endereços de contratos disponível e toda a extração durou 1,2 dias.

Qtde. Amostra	Qtde. Disponível	Porcentagem (%)
156.250	166.443.932	0,094

Tabela 1. Relação entre a amostra de endereços coletadas e endereços disponíveis para extração.

Em relação aos contratos inteligentes resgatados pela API da Polygon, dos 156.250 endereços da amostra, foram retornados 4.708 contratos inteligentes verificados pela rede. Por sua vez, os 151.542 endereços restantes foram marcados como contratos não verificados e, portanto, não puderam ter seus códigos extraídos. Não foram levados em consideração contratos sem transações realizadas, o critério de seleção foi extrair o máximo de contratos inteligentes possíveis para compor o conjunto de dados. Sendo assim, pode-se constatar quais contratos desenvolvidos compõem a rede principal da Polygon. No gráfico da Figura 3 é exibida a proporção entre contratos verificados e não verificados, sendo que os contratos inteligentes verificados representam cerca de 3,01%

da amostra estudada. Todo o *dataset* pode ser consultado de maneira aberta através do GitHub[6].

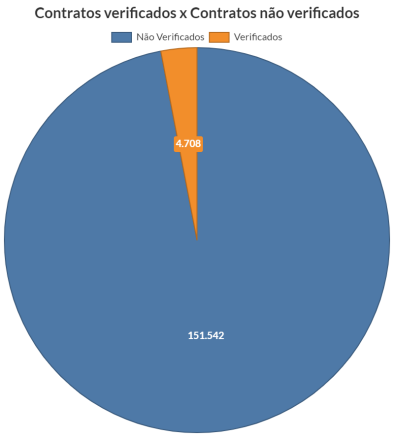

Figura 3. Gráfico relacionando contratos inteligentes verificados e não verificados.

6. Conclusão

As plataformas Blockchain, juntamente com os contratos inteligentes ainda possuem muitos desafios. O presente trabalho propôs a construção de um conjunto de dados de contratos inteligentes com o objetivo de auxiliar o desenvolvimento de códigos seguros e a criação de modelos de classificação. Através dos 4.708 contratos extraídos, esperamos permitir desenvolvedores de contratos e ferramentas de análise aprimorarem suas técnicas.

Além disso, o trabalho projetou a diferença entre contratos verificados e não verificados na rede Polygon dentre a amostra estudada, constatando que os usuários de contratos não verificados devem confiar na transação realizada, sem a possibilidade de visualizar o que ele está assinando. Em contrapartida, usuários dos contratos inteligentes verificados possuem a capacidade de obter mais informações acerca do contrato que está sendo utilizado, bem como há possibilidade do próprio usuário verificar o código e atestar que o contrato executa o que ele realmente foi designado para fazer.

Ademais, como trabalhos futuros esperamos aprimorar o conjunto de dados com mais contratos inteligentes e com técnicas mais ágeis de extração de código, isto é, incrementar a amostra extraindo mais endereços de contratos por meio da plataforma BigQuery e utilizar a API de maneira otimizada. Em adição, planejamos realizar trabalhos sobre o *dataset* construído, de maneira que possamos classificar os contratos extraídos em classes que favoreçam os estudos de contratos inteligentes e redes Blockchain.

Referências

Back, A., Corallo, M., Dashjr, L., Friedenbach, M., Maxwell, G., Miller, A., Poelstra, A., Timón, J., and Wuille, P. (2014). Enabling blockchain innovations with pegged si-

[6]https://github.com/lesc-ufv/smarset-polygon

dechains. *URL: http://www. opensciencereview. com/papers/123/enablingblockchain-innovations-with-pegged-sidechains*, 72:201–224.

Bhutta, M. N. M., Khwaja, A. A., Nadeem, A., Ahmad, H. F., Khan, M. K., Hanif, M. A., Song, H., Alshamari, M., and Cao, Y. (2021). A survey on blockchain technology: evolution, architecture and security. *IEEE Access*, 9:61048–61073.

Buterin, V. et al. (2014). A next-generation smart contract and decentralized application platform. *white paper*, 3(37):2–1.

Durieux, T., Ferreira, J. F., Abreu, R., and Cruz, P. (2020). Empirical review of automated analysis tools on 47,587 ethereum smart contracts. In *Proceedings of the ACM/IEEE 42nd International conference on software engineering*, pages 530–541.

Gregoriadis, M., Muth, R., and Florian, M. (2022). Analysis of arbitrary content on blockchain-based systems using bigquery. *arXiv preprint arXiv:2203.09379*.

Hewa, T. M., Hu, Y., Liyanage, M., Kanhare, S. S., and Ylianttila, M. (2021). Survey on blockchain-based smart contracts: Technical aspects and future research. *IEEE Access*, 9:87643–87662.

Kanani, J., Nailwal, S., and Arjun, A. (2021). Matic whitepaper. *Polygon, Bengaluru, India, Tech. Rep., Sep*.

Kushwaha, S. S., Joshi, S., Singh, D., Kaur, M., and Lee, H.-N. (2022). Ethereum smart contract analysis tools: A systematic review. *IEEE Access*.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review*, page 21260.

Oliva, G. A., Hassan, A. E., and Jiang, Z. M. J. (2020). An exploratory study of smart contracts in the ethereum blockchain platform. *Empirical Software Engineering*, 25(3):1864–1904.

Singh, A., Click, K., Parizi, R. M., Zhang, Q., Dehghantanha, A., and Choo, K.-K. R. (2020). Sidechain technologies in blockchain networks: An examination and state-of-the-art review. *Journal of Network and Computer Applications*, 149:102471.