Explorando o Algoritmo Energy-based Flow Classifier (EFC) na Identificação de Transações Suspeitas Para Lavagem de Dinheiro

Kevin de Santana Araujo

k.santanaraujo@gmail.com

https://www.mercadobitcoin.com.br

Mercado Bitcoin (MB)

EFC Aplicado à Técnicas de Lavagem de Dinheiro

Brasília, Outubro 2024

Sumário

Elliptic Dataset 1/5

Sobre o Dataset

- Dispobilizado[1] pela Elliptic, uma empresa dedicada a detecção de crimes financeiros em criptomoedas [1].
- Inclui 49 grafos amostrados da *blockchain* do *Bitcoin* em diferentes momentos sequenciais no tempo (*time-steps*), conforme apresentado na Figura 1
- Cada grafo é um grafo acíclico direcionado, que começa com uma transação e inclui transações relacionadas subsequentes na *blockchain*, contendo aproximadamente duas semanas de dados.

[1] Disponível em https://www.kaggle.com/datasets/ellipticco/elliptic-data-set

Elliptic Dataset 2/5

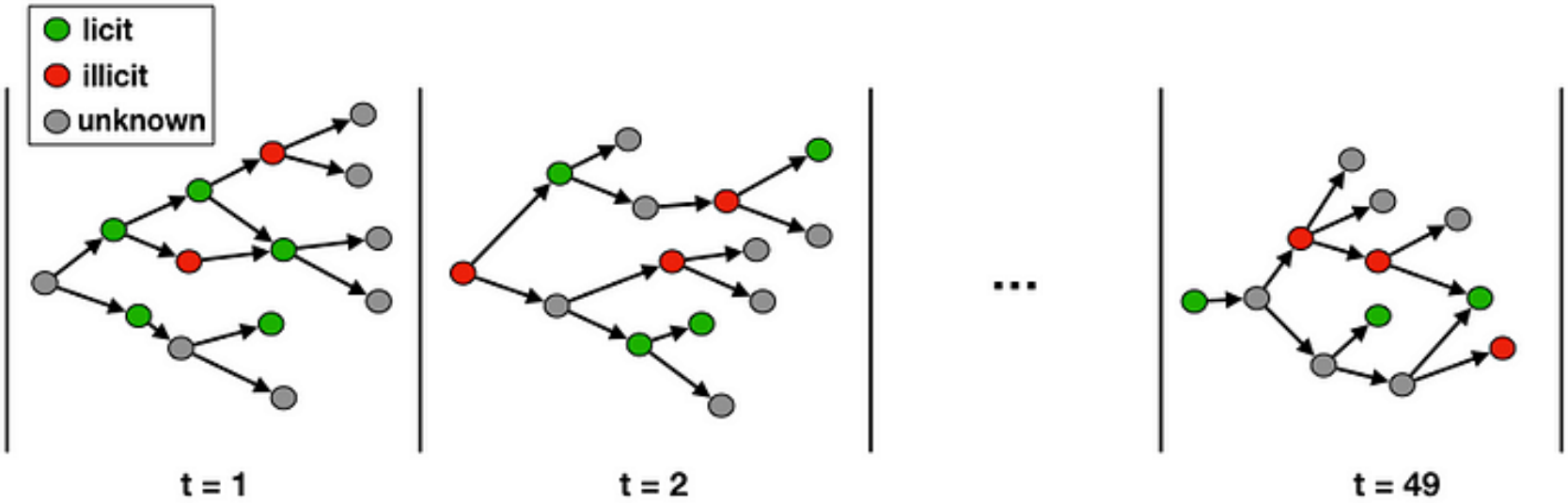

Figura: Estrutura do dataset (retirada de [2])

Elliptic Dataset 3/5

Características do Grafo

- As transações de Bitcoins são transferências de um endereço de Bitcoin (por exemplo, uma pessoa ou empresa) para outro, representado como nós no grafo;
- Cada transação consome o resultado de transações anteriores e gera resultados que podem ser gastos por transações futuras;
- As vértices do grafo representam o fluxo de Bitcoins entre as transações;

Elliptic Dataset 4/5

Sobre os dados

- O dataset consiste em 203.769 transações, das quais 21% são rotuladas como lícitas e 2% como ilícitas, com base na categoria do endereço de bitcoin que criou a transação;
- Cada transação tem 166 *features*, dessas 94 representam informações sobre a própria transação
- Os recursos restantes foram criados por Weber et al. [1] usando informações de um salto para trás/para frente da transação, como o mínimo, o máximo e o desvio padrão de cada recurso da transação.
- Todos as *features*, exceto *timestamp*, são totalmente anônimos e padronizados com média zero e variância unitária.

Elliptic Dataset 5/5

Como o Dataset é Construído

- Métodos e funções utilizadas do artigo [3] (dataset de treinamento e de teste);
- O *split* dos dados são feitos a partir do *time-step* dos grafos, de 1 a 49. Conforme Figura 3

Elliptic Dataset 5.1/5

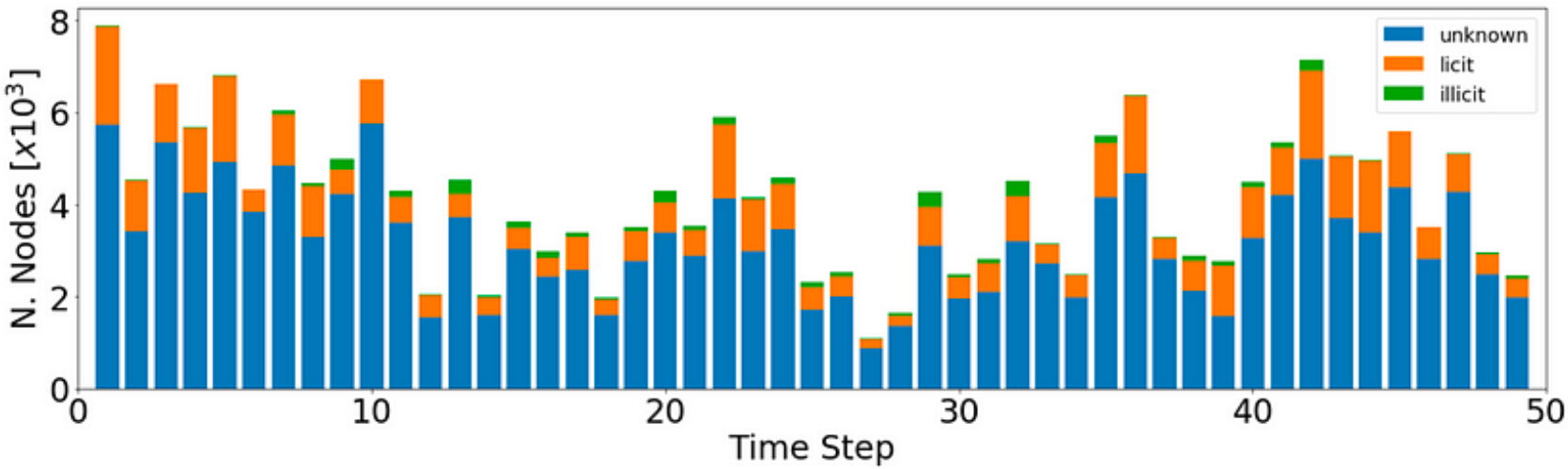

Figura: Tamanho e composição dos componentes conectados vs. *time step*. (retirada de [2])

Re-escrevendo o EFC

Algoritmo

- API antiga (2021) escrita em Cpython. Implementação em baixo nível que faz chamadas a funções puras em C.
- Bibliotecas e dependências desatualizadas, exigingo a re-escrita de muitos métodos;
- Código não documentadao e confuso, tornando difícil o processo de melhoria e correção de bugs;
- Dependente de recompilação a cada alteração/correção;
- Optei por reescrever todo o código de Cpython para Python puro, melhorias ainda precisam ser feitas
- Tempo de execução, cache de chamada de funções, documentação, organização.

Aplicando o EFC ao Elliptic Data Set

Métodos

1. Não supervisionado, 0% da base rotulada;
2. 5% da base rotulada;
3. 10% da base rotulada;
4. 100% da base rotulada;

Métricas

```
df_metrics
```

	contamination	accuracy	f1	f1_micro	f1_macro	precision	recall	roc_auc
0	0%	0.630114	0.068299	0.630114	0.418776	0.040831	0.208680	0.434038
1	5%	0.627415	0.058226	0.627415	0.412998	0.034833	0.177285	0.417988
2	10%	0.628254	0.058350	0.628254	0.413382	0.034922	0.177285	0.418437
3	100%	0.628254	0.058350	0.628254	0.413382	0.034922	0.177285	0.418437

Figura: Métricas Utilizando o Elliptic Data Set e EFC

Discussões

- Construção do dataset, utilizamos o mesmo lapso temporal. Time step 49 3;
- O artigo [3] utiliza 0.05% 0.1%, 0.15% e 0.2% de contaminação;

Próximos Passos

Futuro

- Defesa?
- Dedicamos mais tempo aos experimentos?
- Outros algoritmos? LLM? Temos tempo?

[1] Mark Weber et al. "Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics". Em: *arXiv preprint arXiv:1908.02591* (2019).

[2] Claudio Bellei. *The Elliptic Data Set: opening up machine learning on the blockchain.* Aug. 2019. URL: https://medium.com/elliptic/the-%20elliptic-%20data-%20set-%20opening-%20up-%20machine-%20learning-%20on-%20the-%20blockchain-%20e0a343d99a14.

[3] Joana Lorenz et al. "Machine learning methods to detect money laundering in the bitcoin blockchain in the presence of label scarcity". Em: *Proceedings of the first ACM international conference on AI in finance*. 2020, pp. 1–8.

Agradecimentos

O projeto conta com o apoio do Mercado Bitcoin.

Obrigado! Perguntas?