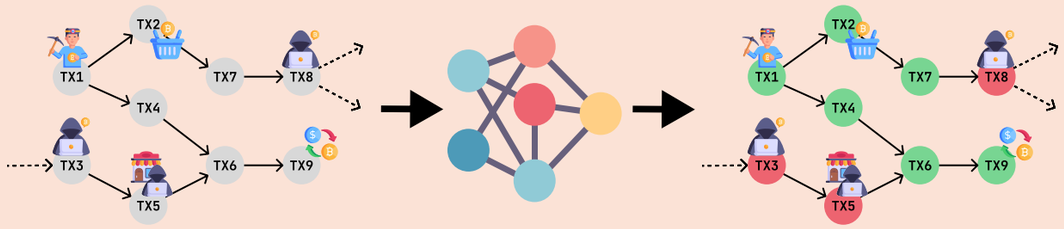
VYUŽITÍ NEURONOVÝCH SÍTÍ PRO ANALÝZU SÍTĚ BITCOIN

ING. MIROSLAV ŠAFÁŘ
VEDOUCÍ: ING. JAN PLUSKAL, PH.D.

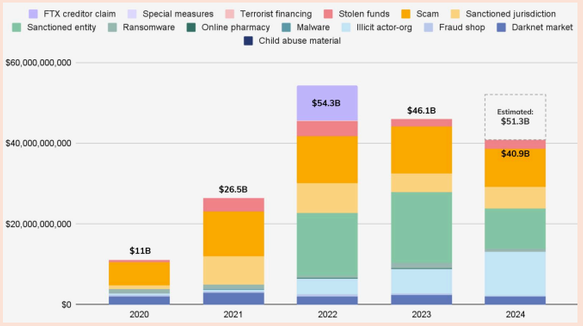
IT SPY

MOTIVACE



Vizualizace detekce nezákonných transakcí za pomoci neuronové sítě

Pseudonymita a absence centrální kontrolní entity u kryptoměny Bitcoin přitahuje pozornost zlomyslných aktérů, kteří chtějí zneužít blockchainové technologie k praní špinavých peněz, financování terorismu, k podvodům anebo k nákupům nelegálního zboží na darknet tržištích. Společnost Chainanalysis [1] odhaduje, že v roce 2024 bylo adresami spojenými s ilegálními aktivitami přijato více jak 51 miliard dolarů. Identifikace podvodného jednání a nákupů nelegálního zboží představuje klíčový úkol bezpečnostních složek, přičemž detekce praní špinavých peněz a financování terorismu pak patří mezi hlavní priority institucí zaměřených na boj s finanční kriminalitou.



Celková hodnota prostředků přijatá adresami spojenými s ilegálními aktivitami

Avšak identifikace prostředků pocházejících z trestné činnosti je současným problémem i ostatních komerčních institucí, které jsou ze zákona povinné identifikovat své klienty, sledovat transakce a hlásit podezřelé aktivity Finančnímu analytickému úřadu. Automatické řešení identifikace takovýchto nelegálních transakcí je tak přirozeným cílem vědecké komunity, který se v poslední dekádě dostal do popředí.

ÚVOD

Moje práce posuzuje dosavadní možnosti využití strojového učení a stanovuje hlavní výzvy, které je zapotřebí adresovat před jejich možným využitím ve forenzní praxi. Práce dále analyzuje obsah nejpoužívanější anonymizované datové sady *Elliptic* [2] využívané k trénování modelů na identifikaci transakcí spojených s ilegální aktivitou. Za pomoci reverzního inženýrství se mi podařilo deanonymizovat 47 % původních příznaků, a díky tomu identifikovat 100 % transakcí obsažených v této datové sadě, což řeší nemožnost rozšíření této datové sady o další transakce. Společně s tím i odhaluje zásadní nedostatky této datové sady v podobě špatné izolace testovacích dat a chybějící grafové informace.

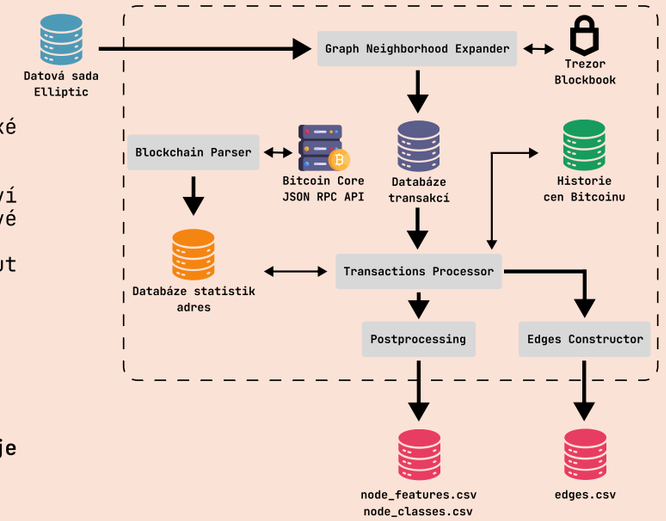
Tato diplomová práce přináší novou otevřenou datovou sadu *OpenElliptic*, která opravuje tyto zásadní nedostatky původní anonymizované datové sady *Elliptic*, a poskytuje tak potřebný základ pro nasazení metod hlubokého učení v praktických forenzních aplikacích. Nakonec tato práce experimentálně vyhodnocuje a porovnává tradiční metody strojového učení s nejmodernějšími metodami hlubokého učení založenými na grafových neuronových sítí, a to jak na původní anonymizované datové sadě, tak na nově představené otevřené datové sadě.

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DATOVÁ SADA

Jednou z klíčových výzev využití strojového učení pro automatickou detekci ilegálních transakcí je sběr a zpracování trénovacích a testovacích dat. Ačkoliv blockchain kryptoměny Bitcoin obsahuje všechny historické transakce, získání validních označených dat použitelných k trénování neuronových sítí je komplikované. Jak ukázalo reverzní inženýrství datové sady *Elliptic*, izolace testovací datové sady v prostředí jednoho velkého grafu je komplexní problém a je třeba správně navrhnout vzorkovací algoritmus, který z blockchainu extrahuje potřebné datové sady.

Moje práce takovýto algoritmus představuje, společně i s nástrojem pro samotnou tvorbu datové sady a datovou sadou *OpenElliptic*, která je otevřenou alternativou k anonymizované sadě *Elliptic* a která opravuje její zásadní nedostatky v podobě chybějících informací a špatné izolace testovacích dat.



Architektura systému pro tvorbu otevřené datové sady OpenElliptic

POROVNÁNÍ ML PŘÍSTUPŮ

Model	Precision	Recall	Illicit F1
XGBoost	0,78 / 0,89	0,68 / 0,62	0,73 / 0,73
Random Forest	0,81 / 0,85	0,72 / 0,71	0,76 / 0,77
Vícevrstvý perceptron	0,50 / 0,60	0,72 / 0,53	0,59 / 0,56
Grafová attention síť	0,83 / 0,28	0,61 / 0,70	0,71 / 0,40
GAT-Residual	0,92 / 0,67	0,66 / 0,66	0,77 / 0,66
DGA-GNN	0,74 / 0,53	0,73 / 0,70	0,73 / 0,60

Výsledky jednotlivých ML modelů na datové sadě Elliptic vs. OpenElliptic

V rámci mé práce porovnávám dostupné modely tradičního strojového učení jako je model XGBoost, Random Forest a vícevrstvý perceptron s nejmodernějšími modely hlubokého učení založeného na grafových neuronových sítích v podobě grafové attention sítě (GAT), grafové konvoluční sítě a grafové sítě DGA-GNN, která využívá dynamického seskupování jednotlivých uzlů grafu.

Z výsledků provedených experimentů lze konstatovat, že tradiční model strojového učení Random Forest založený na rozhodovacích stromech překonává výkonově komplexnější modely grafových neuronových sítí a to díky své lepší odolnosti vůči přetrénování.

REFERENCE

[1] Chainanalysis. *2025 Crypto Crime Trends: Illicit Volumes Portend Record Year as On-Chain Crime Becomes Increasingly Diverse and Professionalized* online. 2025. Dostupné z: https://www.chainalysis.com/blog/2025-crypto-crime-report-introduction/.

[2] Weber, M.; Domeniconi, G.; Chen, J.; Weidele, D. K. I.; Bellei, C. et al. *Anti-Money Laundering in Bitcoin: Experimenting with Graph Convolutional Networks for Financial Forensics*. arXiv, červenec 2019. Dostupné z: http://arxiv.org/abs/1908.02591.