“Breaking Bad: De-Anonymising Entity Types on the Bitcoin Blockchain Using Supervised Machine Learning”

HICSS–51, January 2018
Internet and the Digital Economy
Distributed Ledger Technology: The Blockchain Minitrack

Awa Sun Yin
@awasunyin
About Me

Haohua (Awa) Sun Yin

Academic Background & Current Position

- BSc. in Business & Statistics
- MSc. in Information Systems
- Data Scientist at Chainalysis, spec. in ML/DL & Blockchain
- Research director at the Interchain Foundation*

Research Areas of Interest

- Application of ML/DL to Blockchain data for clustering, de-anonymization, etc.
- 2nd and 3rd Generation Blockchains: Ethereum, Cøsmos*
- Privacy Coins & Cryptography: Monero, ZCash
Adoption of Cryptocurrencies

- 2.9 to 5.8 million unique users (mostly Bitcoin), and increasing
- Accepted as a payment method by over 100,000 merchants (~2014)

Affiliation with Illicit Activities

- Used for: Money laundering, scamming, terror financing
- Used as payment method for: cyber-extortion (ransom payments), thievery, trading illegal goods in the Darknet

Need for Investigation & Compliance Tools

- Businesses: Required by AML and KYC regulations, need tools to assess the risk of each of their customers
- Law Enforcement: Need for domain specific analysis and investigation tools
Background & Motivations

Problem Formulation

Anyone can create anytime a BTC “account”

BTC “Account”

Public Key / Address
Receiving Address

Transaction Hash
Amount BTC

Phone Number
Personal ID

Email Address
Physical Address

Other E-Payment Methods

Issued by centralized organizations

Issued by centralized organizations

Personal ID

Email Address

Physical Address

Phone Number

BTC

Anyone can create anytime a BTC “account”

BTC

Public Key / Address

Receiving Address

Transaction Hash

Amount BTC

Phone Number

Personal ID

Email Address

Physical Address

Phone Number

Personal ID

Email Address

Physical Address

BTC

Anyone can create anytime a BTC “account”

BTC

Public Key / Address

Receiving Address

Transaction Hash

Amount BTC

Phone Number

Personal ID

Email Address

Physical Address

Other E-Payment Methods

Issued by centralized organizations
To what extent can we predict the category of a yet-unidentified cluster on the Bitcoin Blockchain?
Basic Concepts

Category: Unknown
Basic Concepts

Category: Personal Wallet

Alice

72cme94f19a47ea0d118e80bef56fbd1e
85d134472f18214ed8e6b236d2701f8

10 BTC

Kraken

1PxoZtvmsb6LnYS5iNERuJBNR5oQZxuXJ
1K4EqxSVd6NndkmVAMsSBd1n8VJNJLwr2

Category: Exchange

B13HjAyJyAX6SYc6hrEqLf4naoKCWdeS
1FaEn6fGeDfDkHZvvSLsVRVxqgQt95JeyD
1JQFWqx9VFMagd8UrjqiJriLmCvCuJLyf5
Methodology: Dataset

Dataset Description

- 434 Observations
- 10 categories
- + 200 M transactions
- Period: January 2009 – May 2017
Background & Motivations

Problem Formulation

Research Question & Objectives

Data Preparation

Data Provider's Tool

Raw data → Clustering & manual labelling → Enriched data → Preprocessing → Feature engineering → Input data

BTC Node or Core APIs → Co-spend, custom heuristics, etc

Data Preparation (Python & Numpy)

API

Cleaning & Dimensionality Reduction → 76 Features → [434, 77]
Analysis: Supervised Learning

Data Analysis (Python & Scikit-Learn)

1. Selection of algorithms
2. Find optimal parameters
3. Find optimal parameters

Input data

- kNN
- **RF
- ET
- **AB
- DT
- **BG
- ***GB

Oversample two minority classes

SMOTE

CV-Random Search

Train each model with default parameters

Train each model with optimal parameters

Assess & compare results

Conf, ROC & Class. Report
ROC Curve and Classification Report from Gradient Boosted Trees classifier, average confidence of 77%

<table>
<thead>
<tr>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exchange</td>
<td>0.79</td>
<td>0.94</td>
<td>0.86</td>
<td>201</td>
</tr>
<tr>
<td>Gambling</td>
<td>0.74</td>
<td>0.83</td>
<td>0.78</td>
<td>89</td>
</tr>
<tr>
<td>Hosted Wallet</td>
<td>0.25</td>
<td>0.11</td>
<td>0.15</td>
<td>9</td>
</tr>
<tr>
<td>Merchant Services</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>13</td>
</tr>
<tr>
<td>Mining Pool</td>
<td>0.96</td>
<td>0.84</td>
<td>0.90</td>
<td>31</td>
</tr>
<tr>
<td>Mixing</td>
<td>0.50</td>
<td>0.25</td>
<td>0.33</td>
<td>4</td>
</tr>
<tr>
<td>Other</td>
<td>0.43</td>
<td>0.15</td>
<td>0.22</td>
<td>20</td>
</tr>
<tr>
<td>Ransomware</td>
<td>0.91</td>
<td>0.77</td>
<td>0.83</td>
<td>13</td>
</tr>
<tr>
<td>Scam</td>
<td>0.68</td>
<td>0.59</td>
<td>0.63</td>
<td>22</td>
</tr>
<tr>
<td>Tor Market</td>
<td>0.79</td>
<td>0.59</td>
<td>0.68</td>
<td>32</td>
</tr>
<tr>
<td>Avg / Total</td>
<td>0.74</td>
<td>0.77</td>
<td>0.75</td>
<td>434</td>
</tr>
</tbody>
</table>
Implications
- It is possible to categorize unidentified clusters on Bitcoin using supervised learning
- Further challenging Bitcoin’s true level of anonymity
- Applicability to compliance, investigation tools

Limitations
- Dataset limited to 434 observations
- Low performance with under-sampled categories
- Features not reflecting all available data
- Lack of test set
Multiclass Classification on the Bitcoin Blockchain

Goal: Predict the category of unidentified clusters

Methodology: Using a dataset of already identified clusters (a total of 434 observations across 10 categories)

Results: It is possible to classify with a confidence of 77% using Gradient Boosted Trees

Implications of the Research

A degree of de-anonymization can be achieved using this approach

Considering the limitations: Paving the way for future research
### Future Research

#### Expanding the Dataset of Identified Clusters
- Number of observations per category
- Number of categories

#### Refining & Testing Alternative Methodologies
- Automatic feature engineering and extraction
- Testing more classification algorithms
- Binary Classification

#### Applying the Model
Use the tested methodology to uncover the Bitcoin Blockchain for multiple purposes: cybercrime investigations, compliance tools, etc.

**Future Research**

![Pie charts showing Bitcoin ecosystem (Number of Clusters) by BGC and GBC with various categories and percentages.](image)
Proposed Questions

- How can we increase the dataset in both number of observations and categories?
- If Bitcoin is not truly anonymous, why has it been used for nefarious activities?
- Are there alternatives to Bitcoin that offer higher privacy?
Let’s stay in touch!

awasunyin@gmail.com

Get my pub PGP key here:

awasunyin.com