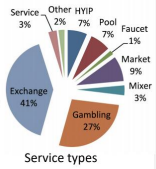
BitScope: Scaling Bitcoin Address De-anonymization using Multi-Resolution Clustering

Zhen Zhang <zgzhen@uw.edu>, Tianyi Zhou <tianyz32@uw.edu>, Zhitong Xie <xzhitong@uw.edu>

Introduction

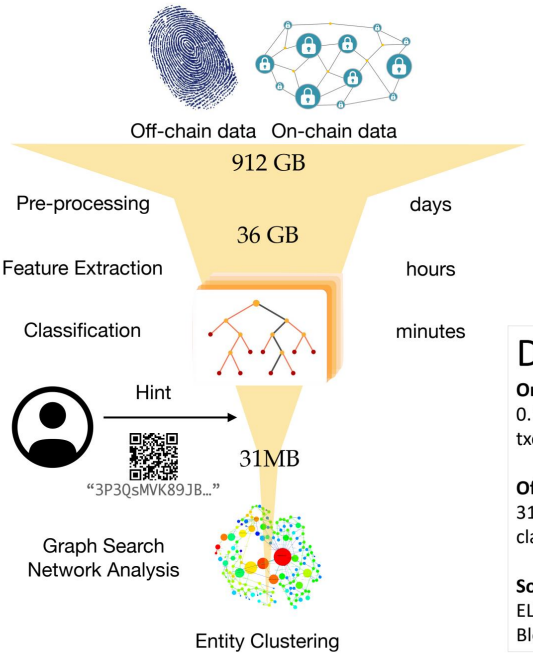
Three **rules of anonymity** in Bitcoin:
(1) Avoid address reuse
(2) Avoid forming central hub or community
(3) Avoid posting your addresses online



Service types

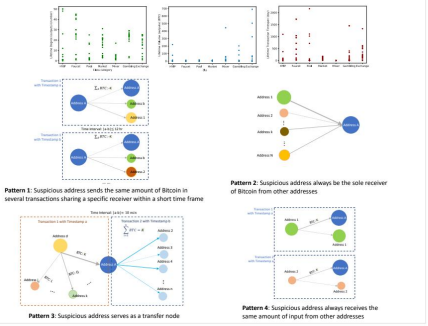
Observations:
(1) Behavioral difference between human and bot
(2) Symmetrical P2P structure vs. asymmetrical structure

BitScope resolution layers:
- **Address Classification**: "Gambling", "Exchange" etc.
- **Service-user Community Detection**
- **Service Address Clustering**



Exploratory Data Analysis

Features: lifetime volume (inputs & outputs); lifetime degree (inputs & outputs); lifetime transaction timespan



Algorithms

```
def iterative_search(seed_nodes, H, p):
  S = seed_nodes
  Q = [ (s, 0) for s in S] # element and depth
  while Q.length:
    a, depth = Q.pop()
    txns = find_in_tx(a) + find_out_tx(a)
    for tx in H(a, txns, p):
      for b in find_tx_neighbors(tx)
        if b not in S and depth < MAX_DEPTH:
          Q.add((b, depth + 1))
          S.add(b)
  return S
```

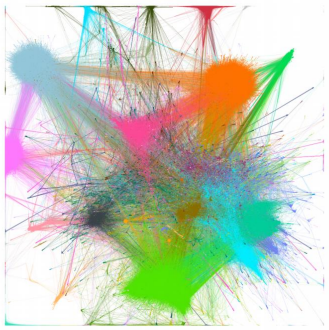
Iterative Search: Find a sub-graph that is most relevant for clustering purpose

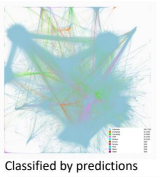
Clustering:
(1) Take advantage of domain-specific knowledge and design corresponding metrics
(2) Use off-the-self clustering/community algorithms for networks

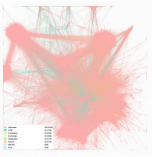
$$J(\text{Users}(a_1), \text{Users}(a_2)) \triangleq \frac{|\text{Users}(a_1) \cap \text{Users}(a_2)|}{|\text{Users}(a_1) \cup \text{Users}(a_2)|}$$

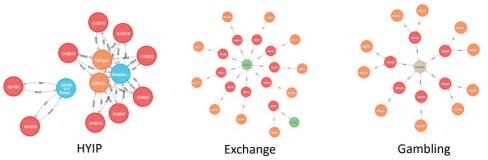
Degree of shared users

Distributions formalization


Partitioned by modularity


Classified by predictions


Classified by ground-truth



Micro Transaction Network Structures for a) **HYIP**: Multiply edges and frequent transactions; b) **Gambling**: More 'txout' than 'txin' and rare mutual transactions; c) **Exchange**: normal pattern while some addresses have extremely high centrality

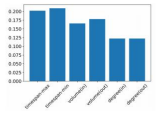
Datasets

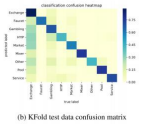
On-chain data (transactions)
0.1B address; 0.1B tx; 0.2B txin; 0.3B txout

Off-chain (labels and entity names)
31,422 rows: addr, tag, link, name, class, owner

Source:
ELTE, BigQuery, WalletExplorer, Blockchain.com

Evaluation


Feature Importance


(b) KFold test data confusion matrix

	Exchange	Faucet	Gambling	HYIP	Market	Mixer	Other	Pool	Service
precision	0.78	0.85	0.76	0.73	0.70	0.84	0.62	0.87	0.77
recall	0.89	0.62	0.69	0.43	0.59	0.76	0.44	0.76	0.71
F-1 score	0.83	0.72	0.72	0.54	0.64	0.80	0.51	0.81	0.74

Classification Results

Dataset: Combine both current and historic data produced by *time-traveling*
Evaluation Method: random shuffling + 10-Fold cross validation
Number of features: 6

name	edges	nodes	density
iter #0	114841	175576	0.000007
iter #1	726484	183319	0.000043
iter #2	1727496	501619	0.000014

(a) Address graph statistics

Method	AR	AMI	V
SharedUsers	0.0003	0.0013	0.7484
Contraction	0.0013	0.0041	0.7463
Baseline	0.0000	-0.0000	0.7664
K-Clique	0.0300	0.2394	0.5229

(b) Clustering results

(1) **Baseline** assigned nodes to different groups (2) **Contraction** use two simple, deterministic inference rules for de-anonymization (3) **SharedUsers** is clustering based on domain-specific network structures. (3) **K-Clique** is a general clustering algorithm.