Detecting Illicit Activity on Crypto-Exchanges

Muhammad Najati al-Imam

Ladislaus von Bortkiewicz Professor of Statistics
BRC Blockchain Research Center
International Research Training Group
Humboldt-Universität zu Berlin
lvb.wiwi.hu-berlin.de
blockchain-research-center.de
irtg1792.hu-berlin.de

What is Money Laundering?

“Money laundering is the illegal process of concealing the origins of money obtained illegally by passing it through a complex sequence of banking transfers or commercial transactions. The overall scheme of this process returns the "clean" money to the launderer in an obscure and indirect way.” -Wikipedia

Why:

- Terrorism
- Drug and Arms Smuggling
- Human Trafficking
- Stolen Funds / Hacks
- Tax Evasion

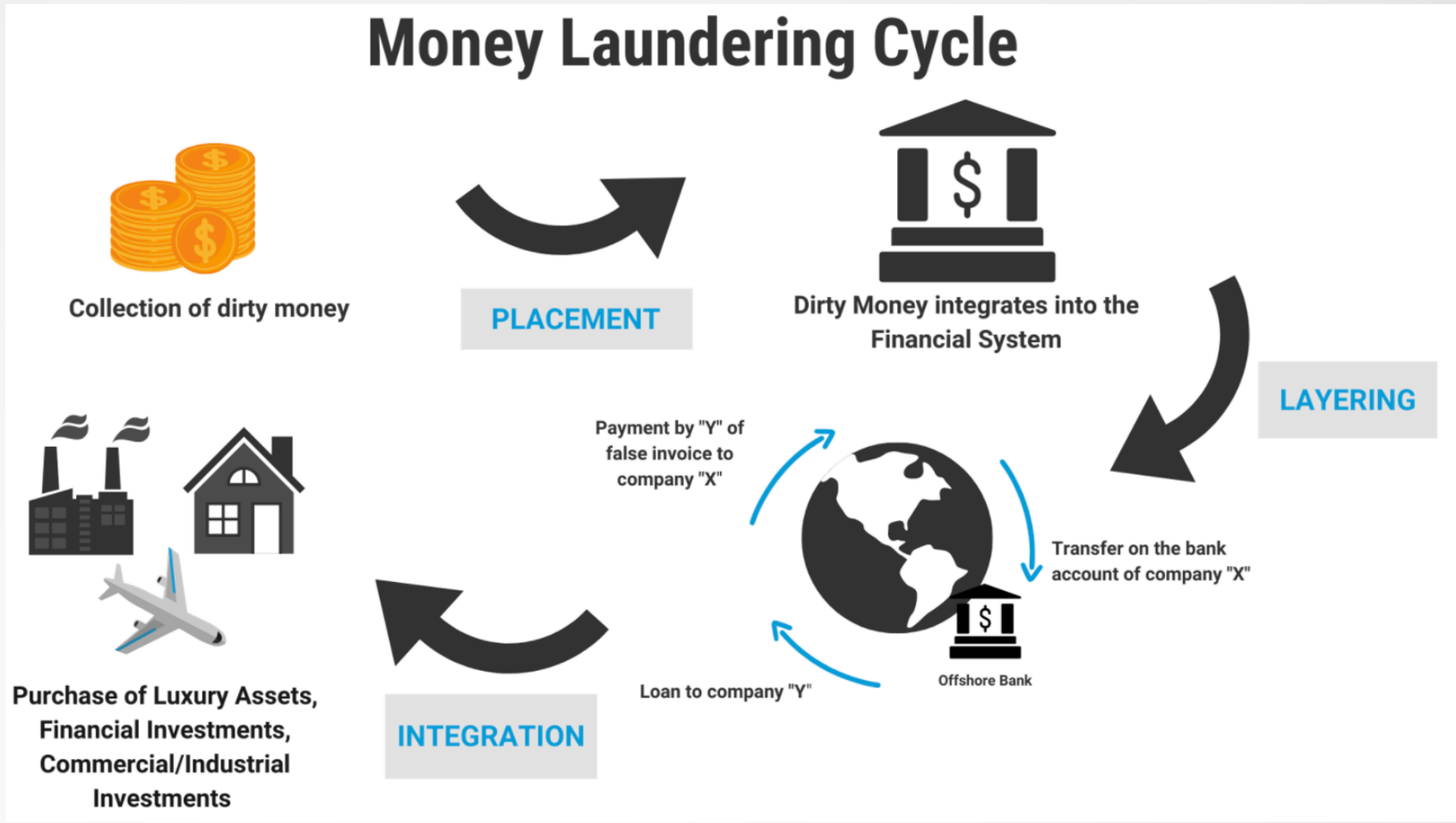
Cryptocurrency as a facilitator

- The estimated amount of total money laundered annually around the world is 2-5% of the global GDP (USD 800 Billion – 2 Trillion)
- In 2019, illicit activity represented 2.1% of all cryptocurrency transaction volume or roughly $21.4 billion worth of transfers.
- In 2020, the illicit share of all cryptocurrency activity fell to just 0.34%, or $10.0 billion in transaction volume

Stages of money laundering

As in the case of cash-based money laundering, there are three main stages in money laundering using cryptos.

1. Placement
2. Layering
3. Integration



Who are the money laundering service providers?

- high-risk exchanges
- mixers
- gambling platforms
- services headquartered in high-risk jurisdictions

How do exchanges protect themselves?

- Know Your Customer (KYC) procedures - AML technology
- Is It required by law? Kind of…
- About a third of crypto exchanges have little or no KYC
- Top 5 of which are Binance, Kraken, Shapeshift, Changelly, Bitcoin ATM (2021)

kraken

ShapeShift

changelly

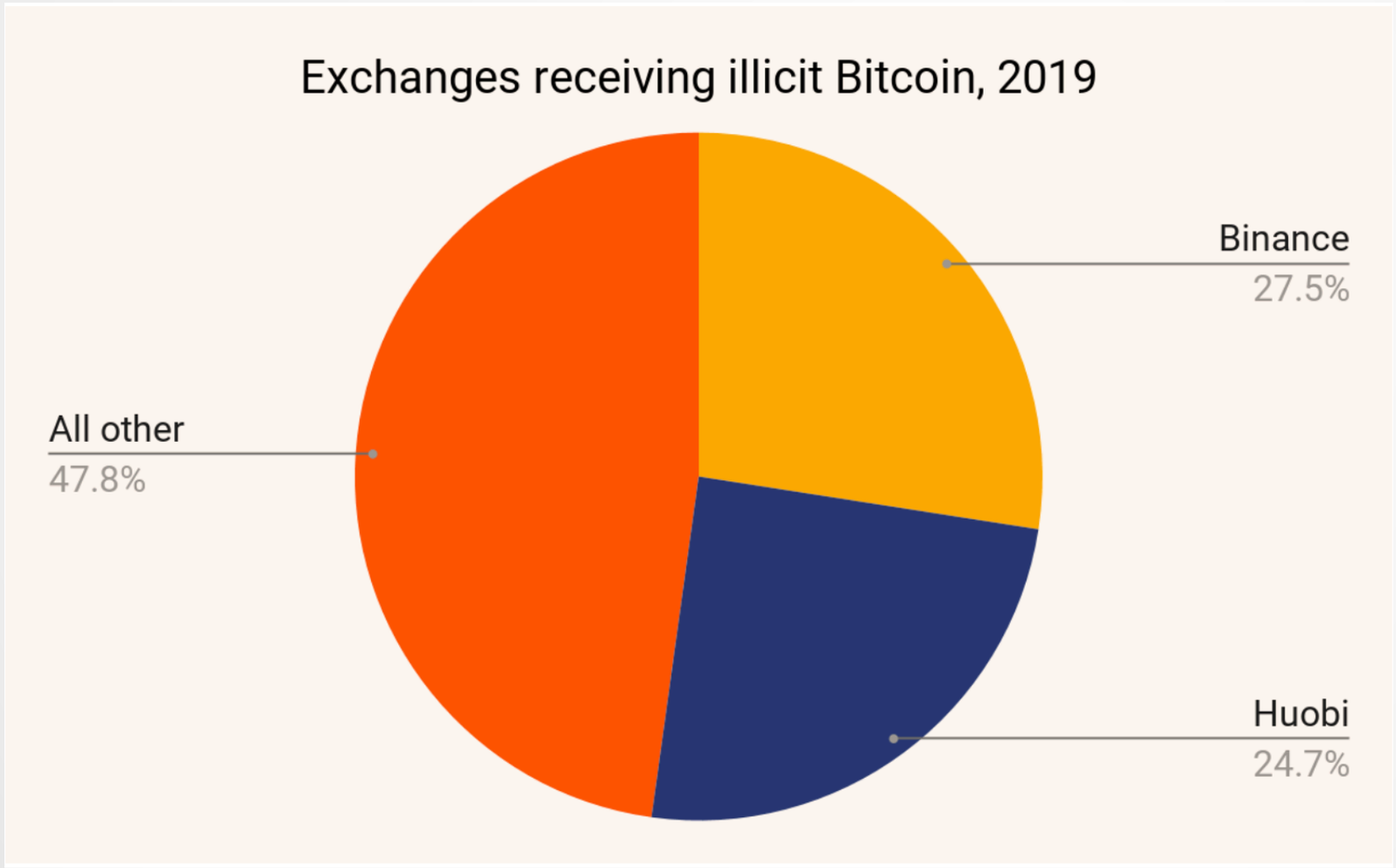
Common KYC requirements

- Full name
- Date of birth
- Phone number and/or email address
- Physical address and/or country of residence
- Photo/scan of government-issued ID
- Copy of utility bill
- Photo of yourself with your ID

Recent Cases - Binance

Binance Holdings Ltd. is under investigation by the Justice Department and Internal Revenue Service (IRS) - (13 May, 2021)

According to Chain Analysis 2020 Crypto Crime Report - $2.8 billion in Bitcoin that moved from criminal entities to exchanges

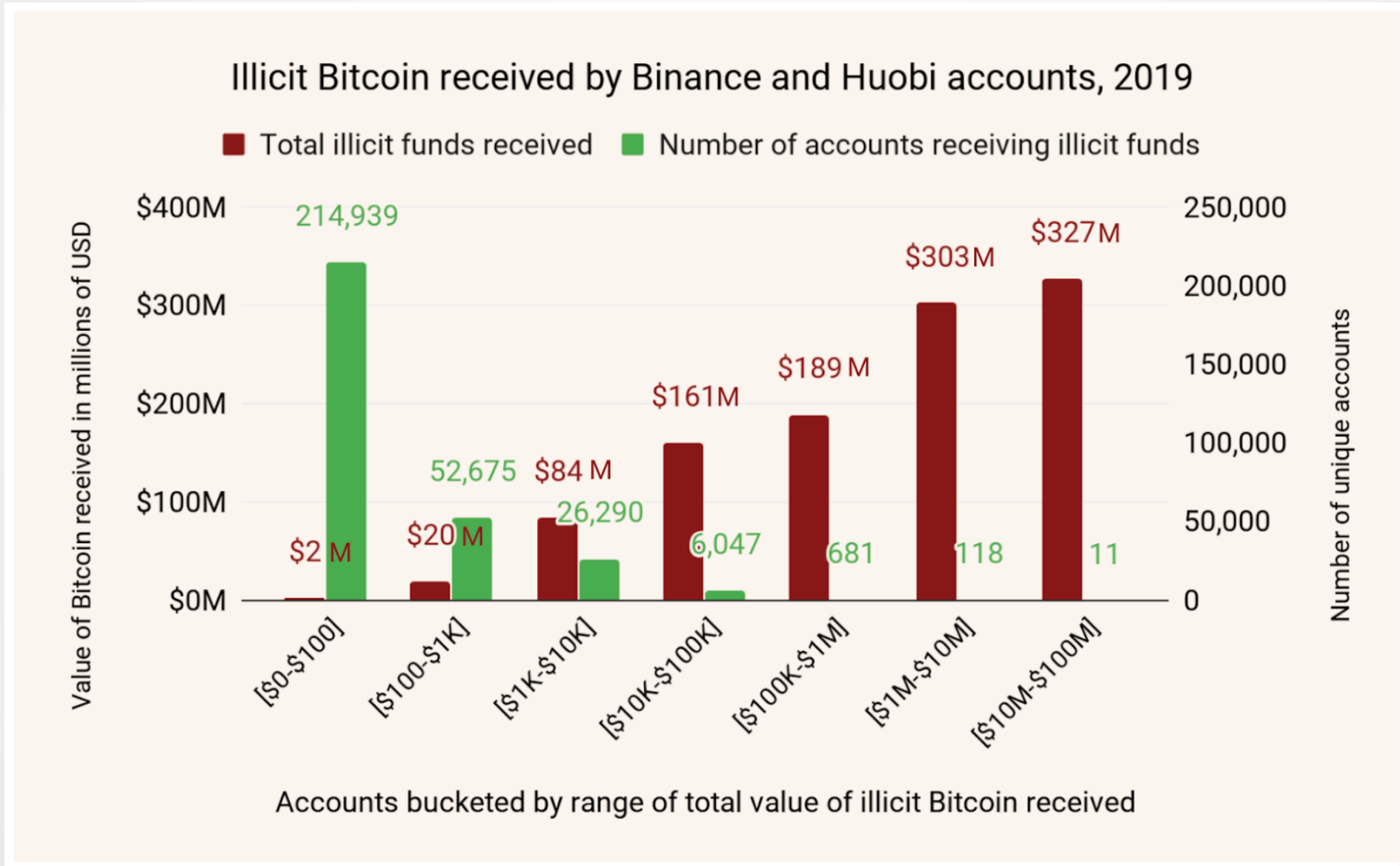


Illicit services include ransomware addresses, sanctioned entities, darknet markets, and addresses associated with scams and stolen funds.

Illicit Bitcoin Funds

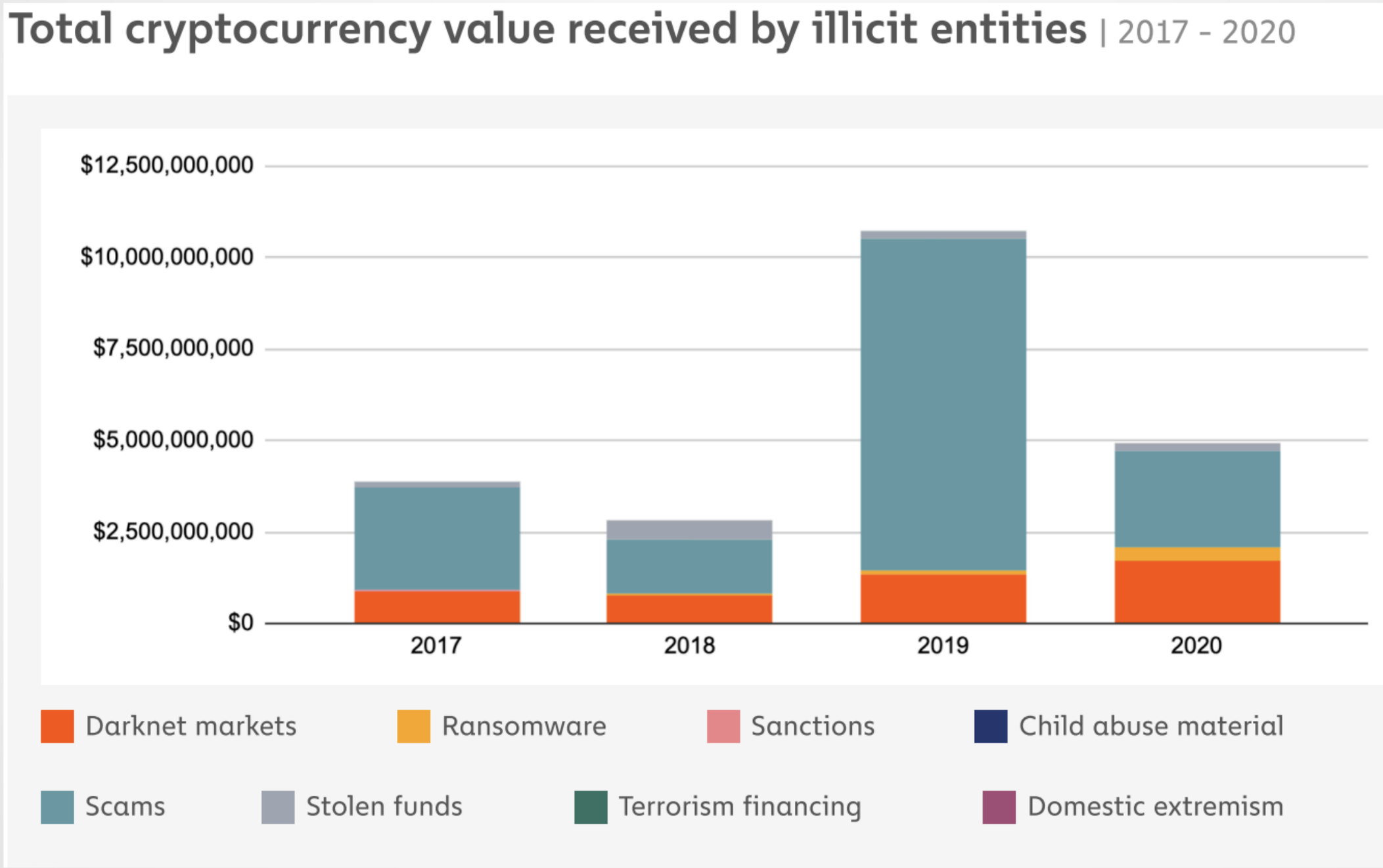
The 810 (out of ~300k) accounts in the three highest-receiving buckets took in a total of over $819 million in Bitcoin from criminal sources, representing 75% of the total - possibly OTC brokers

According to Chain Analysis 2020 Crypto Crime Report



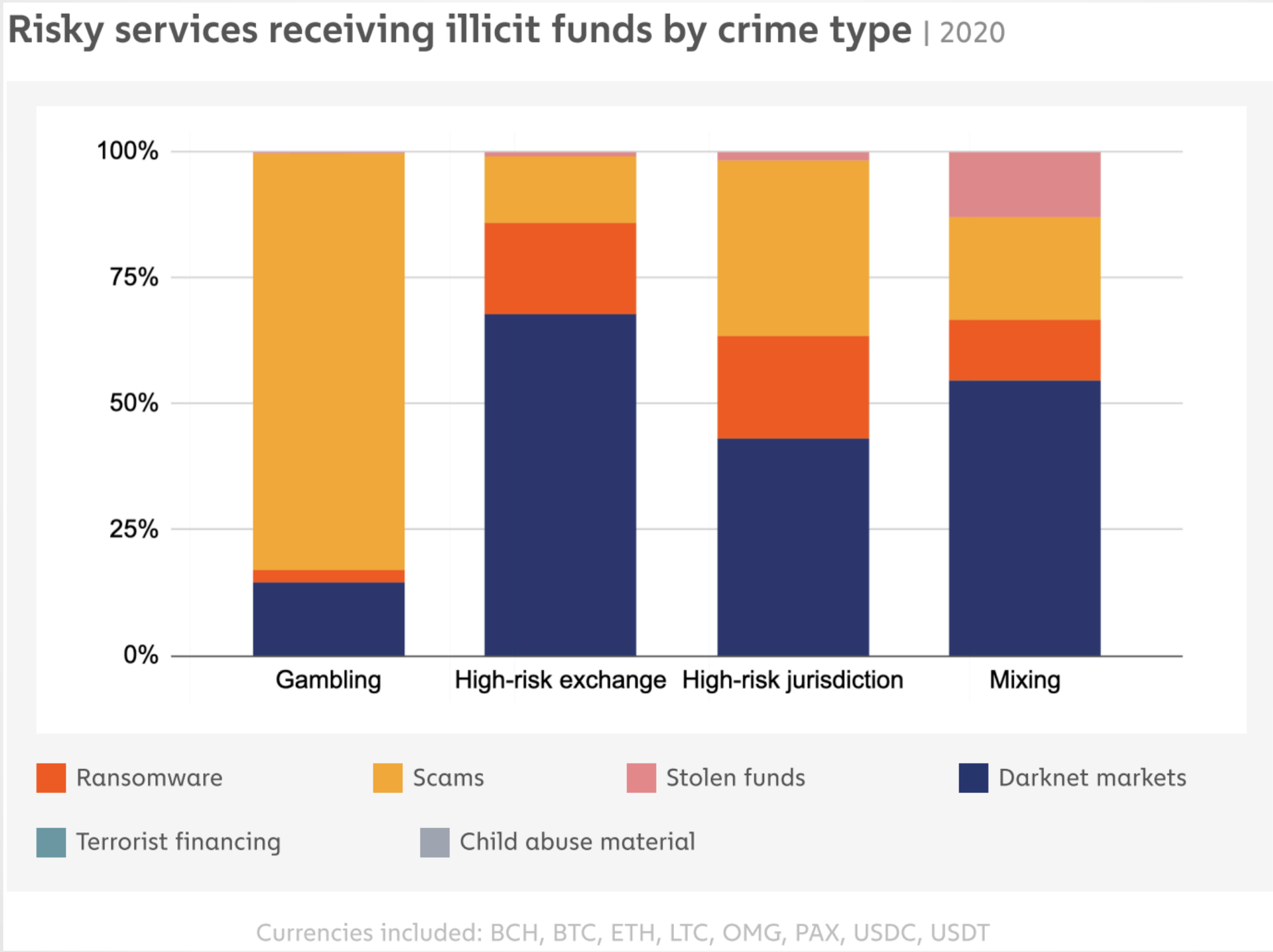
Illicit activity categories

According to Chain Analysis 2021 Crypto Crime Report



Illicit activity categories - laundering methods

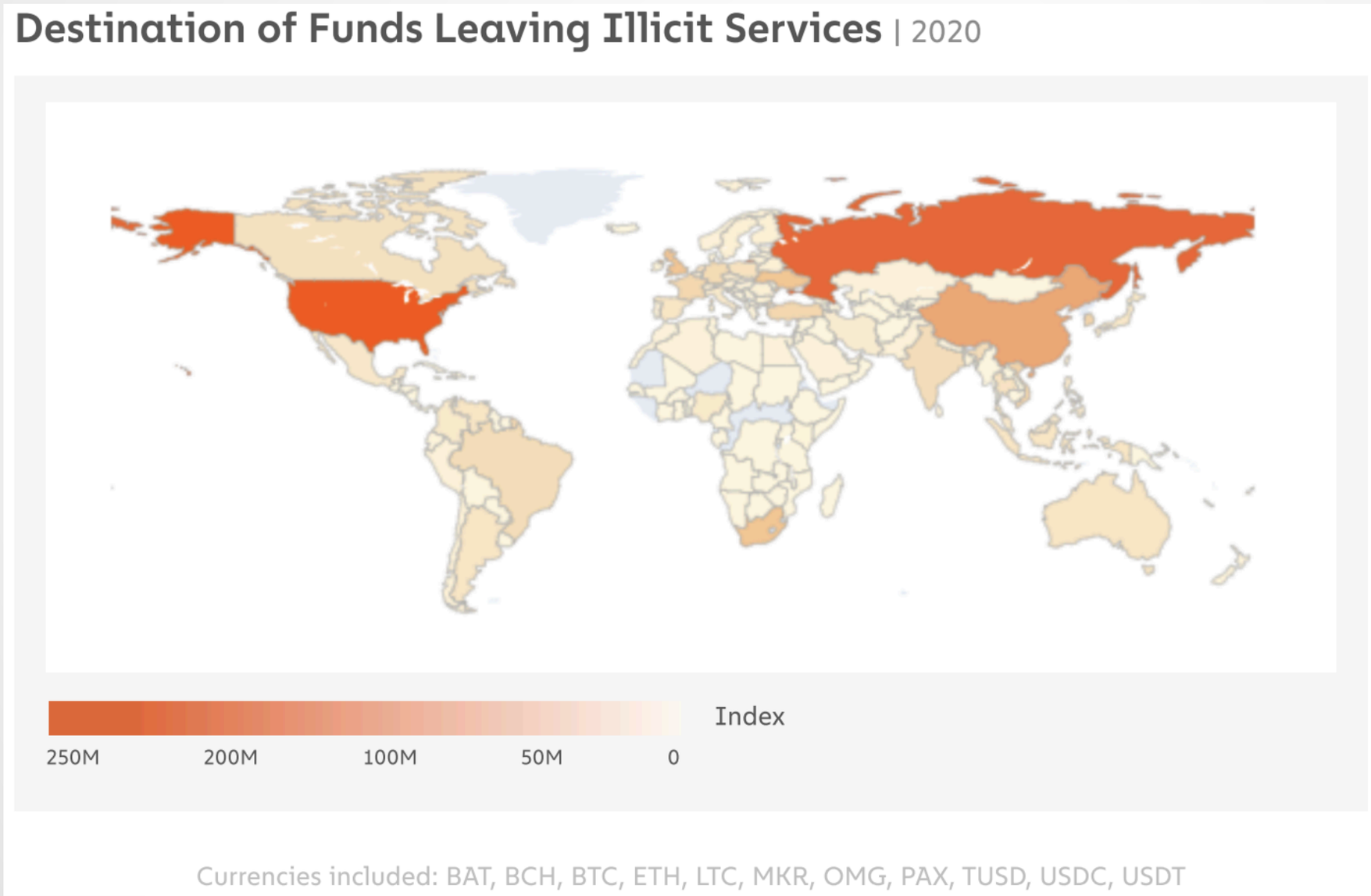
According to Chain Analysis 2021 Crypto Crime Report



Countries that receive the most illicit funds

The following countries receive the highest volume of cryptocurrency from illicit addresses, based on the breakdowns of the locations of the users for the services receiving those funds:

According to Chain Analysis 2021 Crypto Crime Report



- United States
- Russia
- China
- South Africa
- United Kingdom
- Ukraine
- South Korea
- Vietnam
- Turkey
- France

FYI

The **Basel AML Index** is an annual ranking assessing country risk regarding money laundering/ terrorism financing. It focuses on anti-money laundering and counter terrorist financing (AML/ CTF) frameworks and other related factors such as financial/public transparency and judicial strength.

Ranking overall (not specifically tied to crypto)

Ranking	Country	Score
1	Afghanistan*	**8.16**
2	Haiti	**8.15**
3	Myanmar	**7.86**
4	Laos*	**7.82**
5	Mozambique*	**7.81**
6	Cayman Islands	**7.64**
7	Sierra Leone*	**7.51**
8	Senegal	**7.30**
9	Kenya*	**7.18**
10	Yemen*	**7.12**
11	Cambodia	**7.10**
12	Vietnam*	**7.04**
13	Angola*	**7.02**
14	Nigeria*	**6.88**
15	Benin*	**6.85**
16	Nicaragua	**6.78**
17	Côte d'Ivoire*	**6.78**
18	China	**6.76**
19	Algeria*	**6.74**
20	Venezuela*	**6.56**
21	Zimbabwe	**6.54**
22	Cape Verde	**6.52**
23	Sri Lanka	**6.52**
24	Paraguay*	**6.45**
25	Bahamas	**6.43**
26	Tanzania*	**6.39**

What could be a sign of money laundering?

- Velocity of transactions.
- Volume of transactions.
- Sender - Recipient
- Source of funds
- Other indicators?

Can it be detected - What are the challenges?

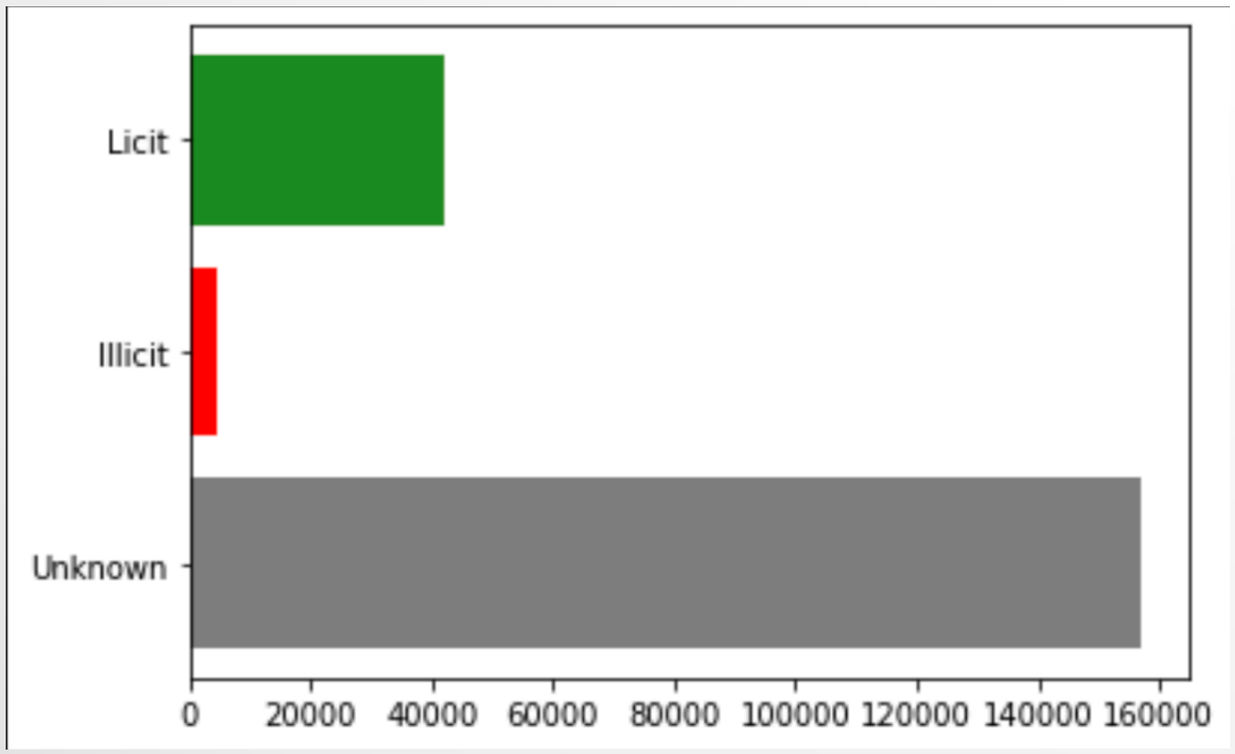
- There are not enough resources or attempts.
- There is a lack of labelled data
- There is a significant imbalance between licit and illicit transactions

Elliptic Dataset

ELLIPTIC

Elliptic: Blockchain analytics, certification, crypto compliance.

- The world's largest labeled transaction dataset publicly available in any cryptocurrency
- 200,000 transactions with a total value of $6 billion (out of 653M as of July, 7, 2021)
- Licit: exchanges, wallet providers, miners, licit services, etc.
- Illicit: scams, malware, terrorist organisations, ransomware, Ponzi schemes, etc.



Licit: 21% ~42k rows (0)

Illicit: 2% ~4.5k rows (1)

Unknown: 77% ~157k rows

Quantlet

Elliptic Dataset - Features

- 94 Local Features: Time Step, number of inputs/outputs, transaction fee, output volume, average number of incoming (outgoing) transactions associated with the inputs/outputs
- 72 Graph Features: obtained using transaction information one-hop backward/ forward from the center node — giving the maximum, minimum, standard deviation and correlation coefficients of the neighbour transactions

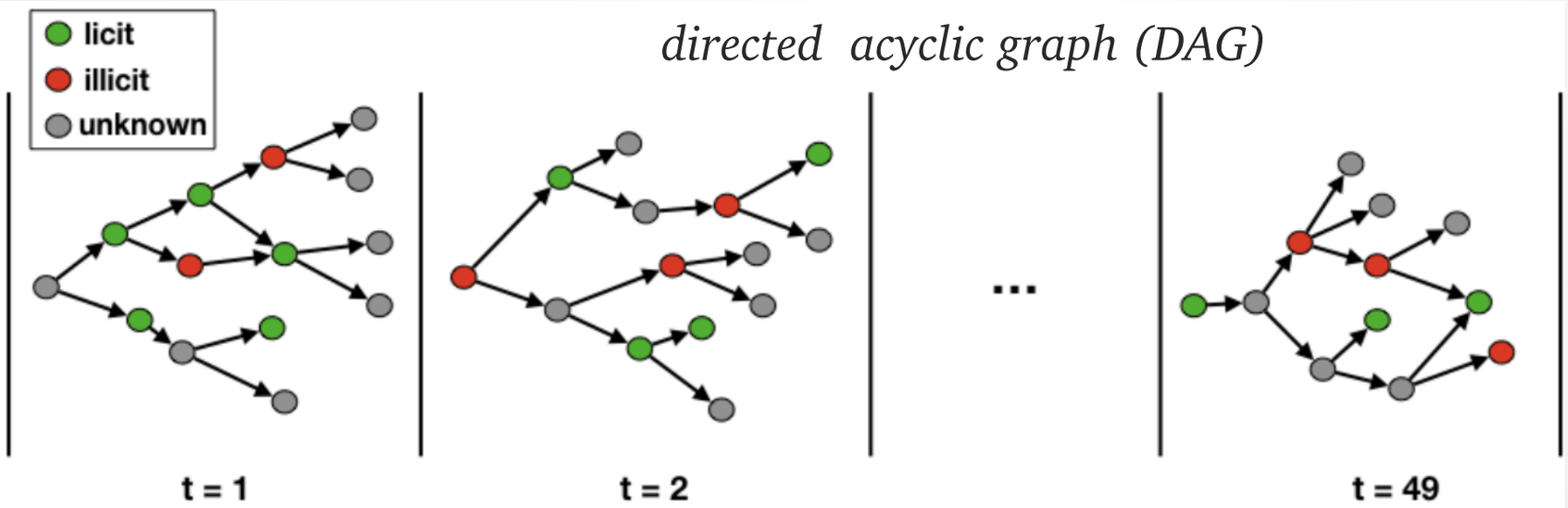


Fig. 2 — The Elliptic Data Set contains 49 connected components, from the earliest one denoted by the time step t = 1 to the later one with a time step t = 49.

Quantlet

Elliptic Dataset - Overview

Due to intellectual property issues, Elliptic cannot provide an exact description of the features.

After dropping class 3 “Unknown”, the percentage of illicit transactions is 10%. No data cleaning required

	txId	class	time step	local_feat_0	local_feat_1	local_feat_2	local_feat_3	local_feat_4
0	232438397	2	1	0.163054	1.963790	-0.646376	12.409294	-0.063725
1	232029206	2	1	-0.005027	0.578941	-0.091383	4.380281	-0.063725
2	232344069	2	1	-0.147852	-0.184668	-1.201369	-0.121970	-0.043875
3	27553029	2	1	-0.151357	-0.184668	-1.201369	-0.121970	-0.043875
4	3881097	2	1	-0.172306	-0.184668	-1.201369	0.028105	-0.043875

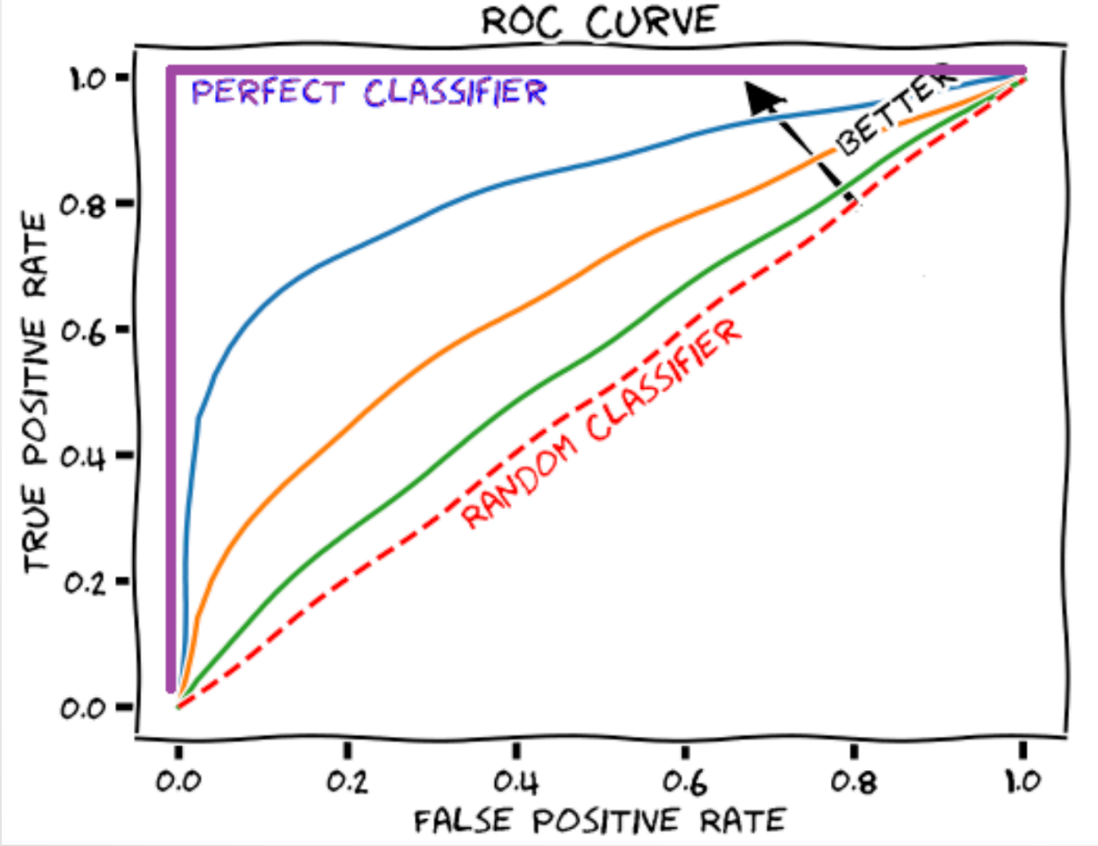
5 rows × 168 columns

Quantlet

Elliptic Dataset - Model Performance

Random Forest					XGBoost					Feed Forward Neural Network				
ACC	Percision	Recall	F1 Score	AUC	ACC	Percision	Recall	F1 Score	AUC	ACC	Percision	Recall	F1 Score	AUC
0.99	0.99	0.95	0.97	0.995	0.99	0.99	0.96	0.97	0.996	0.96	0.87	0.67	0.76	0.945

- ACC = Number of correct predictions / Total Number of correct predictions
- Precision = True Positive / Total predicted positive
- Recall = True Positive / Total actual positive
- F1 = 2 X Precision * Recall / Precision + Recall



- Y Axis is Sensitivity (same as recall): how good a test is at detecting the positives. A test can cheat and maximize this by always returning “positive”
- X Axis is Specificity: how good a test is at avoiding false alarms. A test can cheat and maximize this by always returning “negative”.
- ROC Curve: *Receiver Operating Characteristics* (ROC) curve is a graph showing the performance of a classification model at all thresholds.
- AUC Score: *Area Under Curve* (AUC) score represents the degree or measure of separability. A model with higher AUC is better at predicting True Positives and True Negatives

Quantlet

Random Forest - Parameter Tuning

- n_estimators = number of trees in the forest
- max_features = max number of features considered for splitting a node
- max_depth = max number of levels in each decision tree
- min_samples_split = min number of data points placed in a node before the node is split
- min_samples_leaf = min number of data points allowed in a leaf node

Approaches

- Randomized Search
- Grid Search

Parameter	Value
n_estimators	100
max_features	auto (=sqrt(n_features))
max_depth	none (nodes are expanded until all leaves are pure) - Overfitting problem
min_samples_split	2
min_samples_leaf	1

Quantlet

XGBoost - Parameter Tuning

- learning_rate = makes the model more robust by shrinking the weights on each step
- max_depth = max depth of a tree (Similar to RF)
- min_child_weight = defines the minimum sum of weights of all observations required in a child.
- gamma = A node is split only when the resulting split gives a positive reduction in the loss function. Gamma specifies the minimum loss reduction required to make a split.
- colsample_bytree = max number of features considered for splitting a node (similar to RF)
- n_estimators: number of trees (or rounds)

Approaches

- Exhaustive Grid Search (GS)
- Coordinate Descent (CD)
- Genetic Algorithm (GA)

Parameter	Value
learning_rate	0.01
max_depth	3
min_child_weight	1
gamma	10
colsample_bytree	0.4
n_estimators	1000

Quantlet

Neural Network - Parameter Tuning

- Topology: type, layers, neurons
- Learning Rate: controls the weight at the end of each batch
- Optimizer and loss function: optimizers are algorithms or methods used to change the attributes of the neural network such as weights and learning rate to reduce the losses
- Batch Size: number of samples that will be propagated through the network.
- Number of Epochs: defines the number times that the learning algorithm will work through the entire training dataset

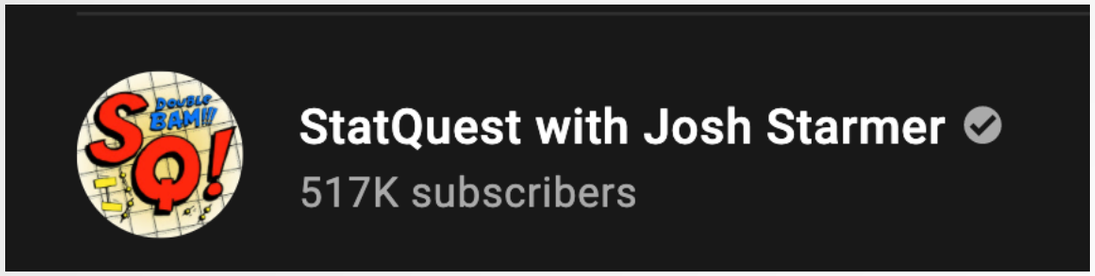
Approaches

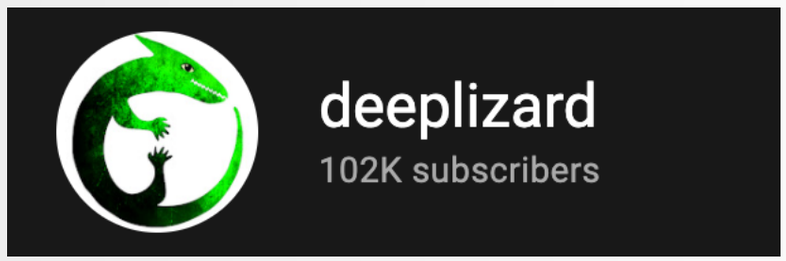
- Randomized Search
- Grid Search

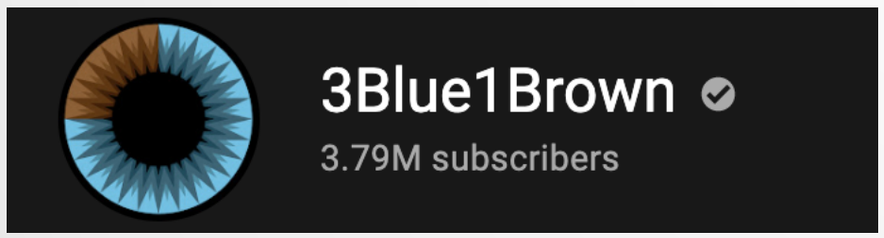
Parameter	Value
topology	Simple FF, 3 layers and output layer (150, 50, 10, 1)
learning rate	0.01
optimzer	adam
batch size	10
epochs	10

Quantlet

Very Good Youtube Channels







Sources

- Money Laundering Definition - https://en.wikipedia.org/wiki/Money_laundering
- Common Crypto Red Flags - https://bitaml.com/2019/07/29/5-red-flags-crypto/
- Money Laundering Estimates - https://www.technologyreview.com/2020/01/16/130843/cryptocurrency-money-laundering-exchanges/
- Binance under investigation - https://www.bloomberg.com/news/articles/2021-05-13/binance-probed-by-u-s-as-money-laundering-tax-sleuths-bore-in
- Chain Analysis 2020 Crypto Crime Report - https://blog.chainalysis.com/reports/money-laundering-cryptocurrency-2019
- Chain Analysis 2021 Crypto Crime Report - https://go.chainalysis.com/2021-Crypto-Crime-Report.html
- Crypto-Exchanges with no KYC requirements - https://www.ikream.com/best-anonymous-cryptocurrency-exchanges-without-kyc-verification-27250
- Machine learning methods to detect money laundering in the Bitcoin blockchain in the presence of label scarcity (Paper) - https://arxiv.org/pdf/2005.14635.pdf
- Money Laundering via Cryptocurrencies: All You Need to Know - https://www.tookitaki.ai/news-views/moneylaundering-via-cryptocurrencies/
- What is Know Your Customer (KYC) for Cryptocurrency? - https://www.cryptovantage.com/guides/know-your-customer/
- Basel AML Index 2020 - https://baselgovernance.org/sites/default/files/2020-07/basel_aml_index_2020_web.pdf

Sources

- The Elliptic Dataset: https://www.elliptic.co/blog/elliptic-dataset-cryptocurrency-financial-crime
- The Elliptic Data Set: opening up machine learning on the blockchain - https://medium.com/elliptic/the-elliptic-data-set-opening-up-machine-learning-on-the-blockchain-e0a343d99a14
- Download the Elliptic Dataset: https://www.kaggle.com/ellipticco/elliptic-data-set
- Total number of Bitcoin transactions: https://www.blockchain.com/charts/n-transactions-total
- Random Forest Classifier Parameters: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- XGBoost parameters complete guide: https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/
- XGBoost Hyperparameter Tuning: https://towardsdatascience.com/doing-xgboost-hyper-parameter-tuning-the-smart-way-part-1-of-2-f6d255a45dde
- Neural Network Hyperparameter Tuning: https://towardsdatascience.com/simple-guide-to-hyperparameter-tuning-in-neural-networks-3fe03dad8594