2020 / 9 / 1

NBiS 2020 @ Online Session

Address Usage Estimation Based on Bitcoin Traffic Behavior

Hiroki Matsumoto, Shusei Igaki, Hiroaki Kikuchi

Meiji University

Background

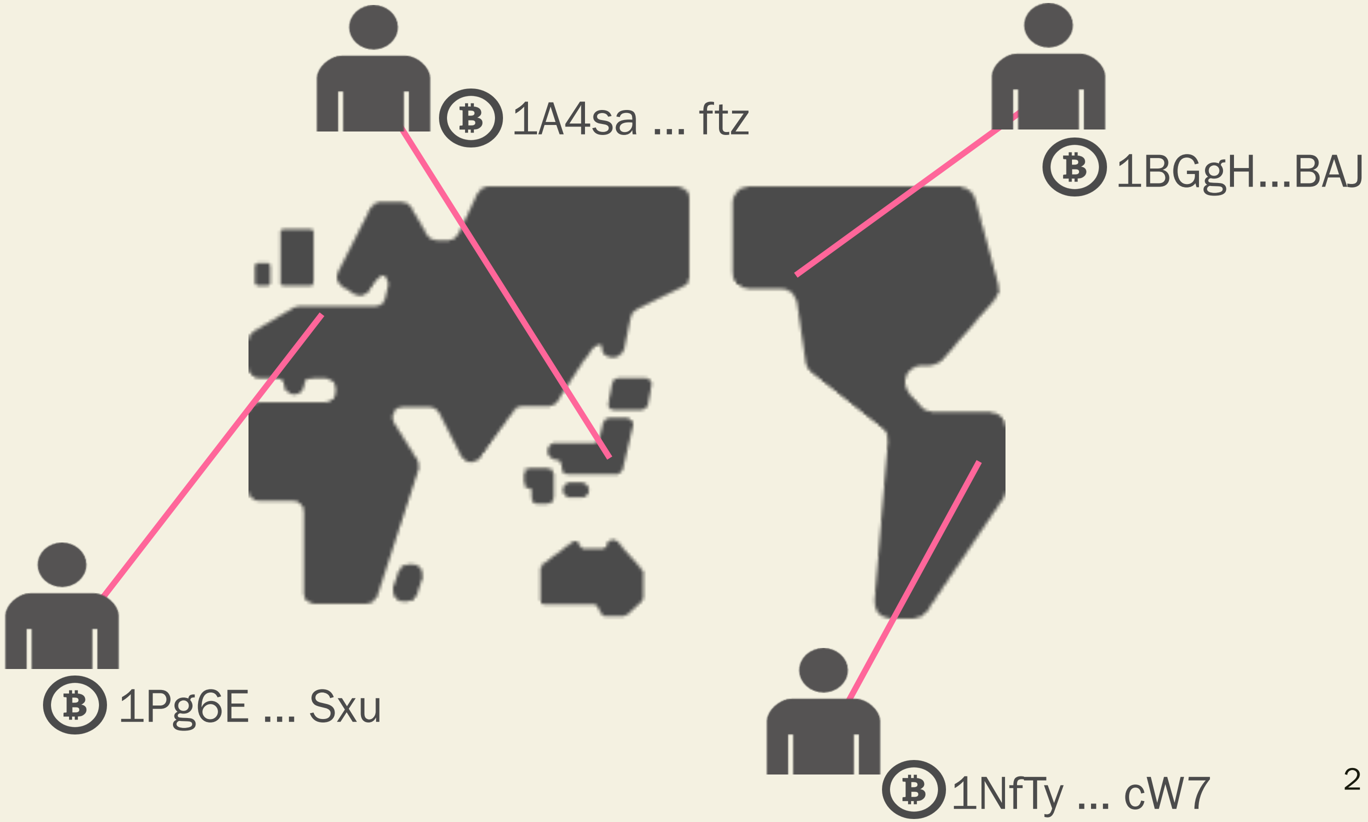
- A large amount of cryptocurrencies were stolen by crackers.
  - Exchange Coincheck; Jan. 2018 (about 58 billion dollar)
  - Exchange Zaif; Sep. 2018 (about 7 billion dollar)
  - Exchange BITPoint; Jul. 2019 (about 3.5 billion dollar)

- Where had these money gone?
  - **It is difficult to trace these money.**
    - E.g.) Mixing service, Trading coin for another cryptocurrency, money laundering service.

Why so hard to track ?

- One-time bitcoin address.
  - Used at pseudonym.

- Involved world wide users.



Previous studies

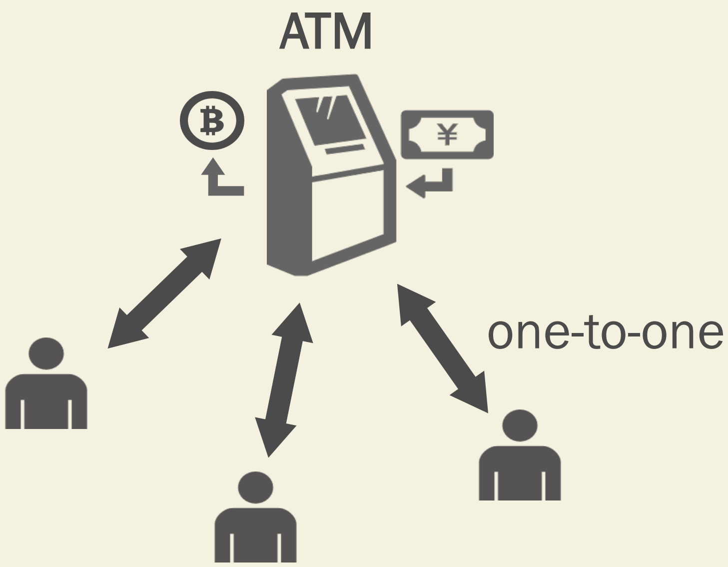
- Identification of the Bitcoin addresses.
  - Heuristics: Combined input addresses are managed by the same user. [Meiklejohn, 2013]
  - Identifying from features of output addresses associated to a target address. [Nagata, 2018]

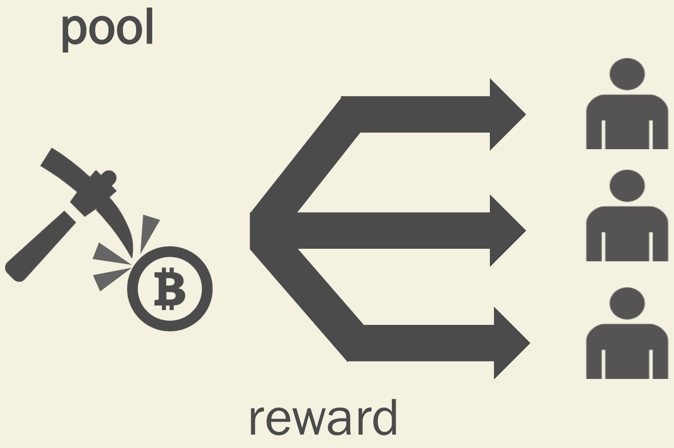
- The estimation user’s attribute.
  - Predicting the time zone where a user lives. [Dupont, 2015]

Our study

- We found that **transaction behavior depends on its usage**.
  - Previous studies didn't consider these behavior.

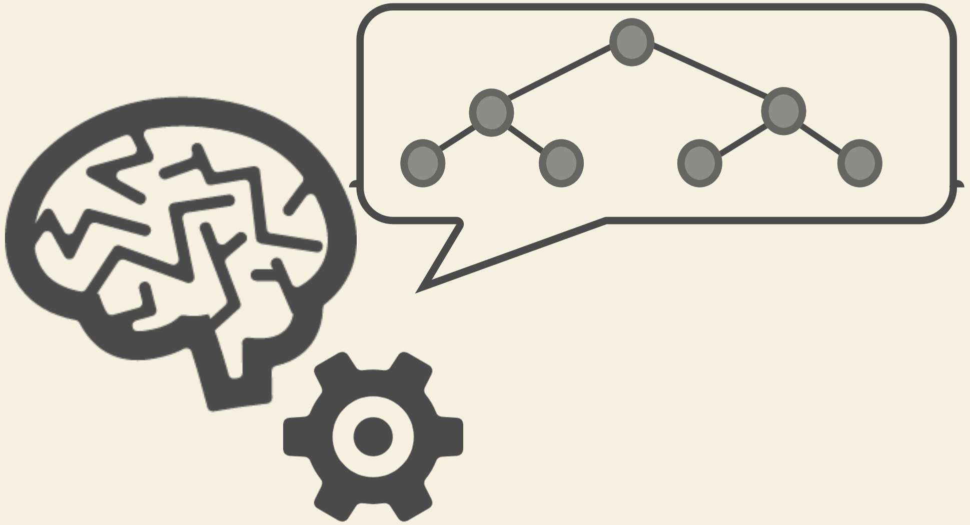
E.g.) Bitcoin ATM VS Mining pool





Our contributions

- We study major seven usages addresses and transactions and examine the characteristics of usages.
- We propose a new algorithm for classifying a set of unknown addresses into seven classes using the decision tree learning.
- We show the experimental results on some useful characteristics of bitcoin traffic.



Seven usages

Bitcointalk (BBS)

(User) These addresses are published in the profile pages of BBS.

Bitcoin ATM (in Toronto, Canada)

(Provider) ATM provider has fixed addresses used to transactions with customer.

(User) Customer deposit read money in an ATM.

Dark web

(Provider) The dark websites that provide illegal service and product publish their addresses.

(User) The dark website publishes some customer’s addresses for their promotion.

Exchange

(User) These addresses are specified in any transactions with known exchange addresses labeled by WalletExplorer[1].

Mining Pool

(Provider) Mining pool use a fixed address to receive a reward for mining bitcoin blocks.

[1] WalletExplorer.com: smart Bitcoin block explorer (https://www.walletexplorer.com/)

Research Questions

- Which is the easiest usage to classify for the seven usages?
- What is the most significant features to estimate the usage of Bitcoin addresses?

Proposed Method

- Original features in this study
  1. Datasets
  2. Transaction pattern
  3. Features

- Experiment
  i. Randomly sample addresses for the dataset.
  ii. Classify seven usage addresses into two groups (training and test).
  iii. Perform threefold cross-validation to evaluate the accuracy of classification for avoiding distortion.
  iv. Record accuracy of the model in precision and recall.
  v. Repeat steps i. to iv. 100 times.

1. Dataset (# addresses, # transactions)

- Collected transactions data from ”Blockchain Explorer”[2].
  - We exclude duplicated addresses that were used for more than one usage.

usage	# addresses		# transactions	duration
	provider	user		
Bitcointalk BBS		2,391	29,638	Apr. 1, 2019 – Sep. 30
Bitcoin ATM	3	452	26,849	
Dark web	26	67	35,076	
Exchange		1,012	33,351	
Mining Pool	98		24,876	
Total	4,049		149,790	

[2] BLOCKCHAIN.COM (https://www.blockchain.com/explorer)

2. Definition of transaction patterns

	Single input address	Multiple input addresses
Any input address (A_1) is specified again at output addresses	S1 Tx Input: A_1 ⇨ Tx Output: B_1, A_1 E.g.) Deposit bitcoin with Bitcoin ATM.	M1 Tx Input: A_1, A_2 ⇨ Tx Output: B_1, A_1 E.g.) Withdraw bitcoin in exchange.
No input addresses are used again at the output addresses	S2 Tx Input: A_1 ⇨ Tx Output: B_1, C_1 E.g.) Specific wallet applications.	M2 Tx Input: A_1, A_2 ⇨ Tx Output: B_1, C_1 E.g.) Mining pool provider pays a mining reward to miners.

3. Features

- Statistics; average, minimum, maximum, median, and standard deviation

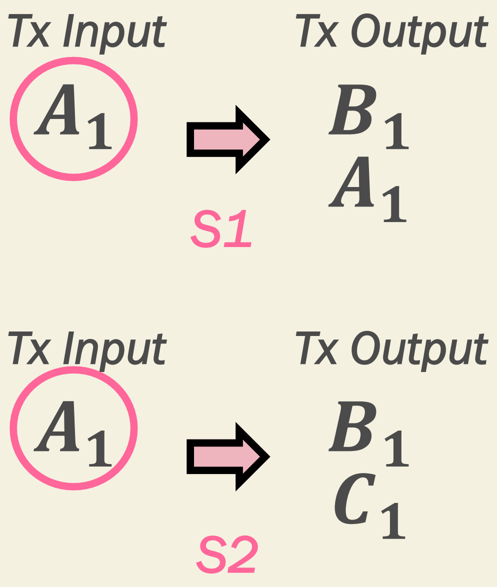
feature	# statistics	description
Txs count	5	Total number of transactions for usages
Txs sending count	5	Total number of sending transactions for usages
Txs receiving count	5	Total number of receiving transactions for usages
Txs input address count	5	Total number of input addresses specified in transaction
Txs output address count	5	Total number of output addresses specified in transaction
Txs address count	1	Total number of addresses in transaction
Reused input address count	1	Total number of reused input address
Reused output address count	1	Total number of reused output address

Result 1. Definition of transaction pattern

More than 90 % provider's addresses were classified transaction pattern $S1$ or $S2$.

- Tx pattern rate: seven usages address

$S1, S2$ ··· Single input address



usage		transaction pattern [%]			
		$S1$	$S2$	$M1$	$M2$
Bitcoin ATM	provider	98.5	0.6	0.8	0.1
Dark web	provider	64.4	28.9	0.2	6.6
Mining Pool	provider	78.7	11.4	0.2	6.6
Bitcointalk BBS	user	23.5	36.1	5.0	35.4
Bitcoin ATM	user	33.3	39.9	0.9	25.9
Dark web	user	23.0	37.8	3.8	35.3
Exchange	user	26.2	33.8	8.7	31.3

Result 2. BBS

- The estimated usages with the decision tree learning algorithm.
  - True Positive score is highest in seven usages. (about 88%)
  - False Positive score is highest in seven usages. (112 addresses)

usage			ATM	Dark web	Mining	BBS	ATM	Dark web	Exchange	total
			provider			user				
			Predicted							
Bitcoin ATM	provider	Actual	0	0	0	1	0	0	0	1
Dark web			0	0	0	8	0	0	0	8
Mining Pool			0	0	2	19	8	0	0	29
Bitcointalk BBS	user		0	0	0	633	31	0	53	717
Bitcoin ATM			0	0	0	16	119	0	1	136
Dark web			0	0	0	12	3	2	3	20
Exchange			0	0	0	56	9	0	239	304

Result 3. Accuracy

Exchange users are classified with 80% accuracy, precision and recall

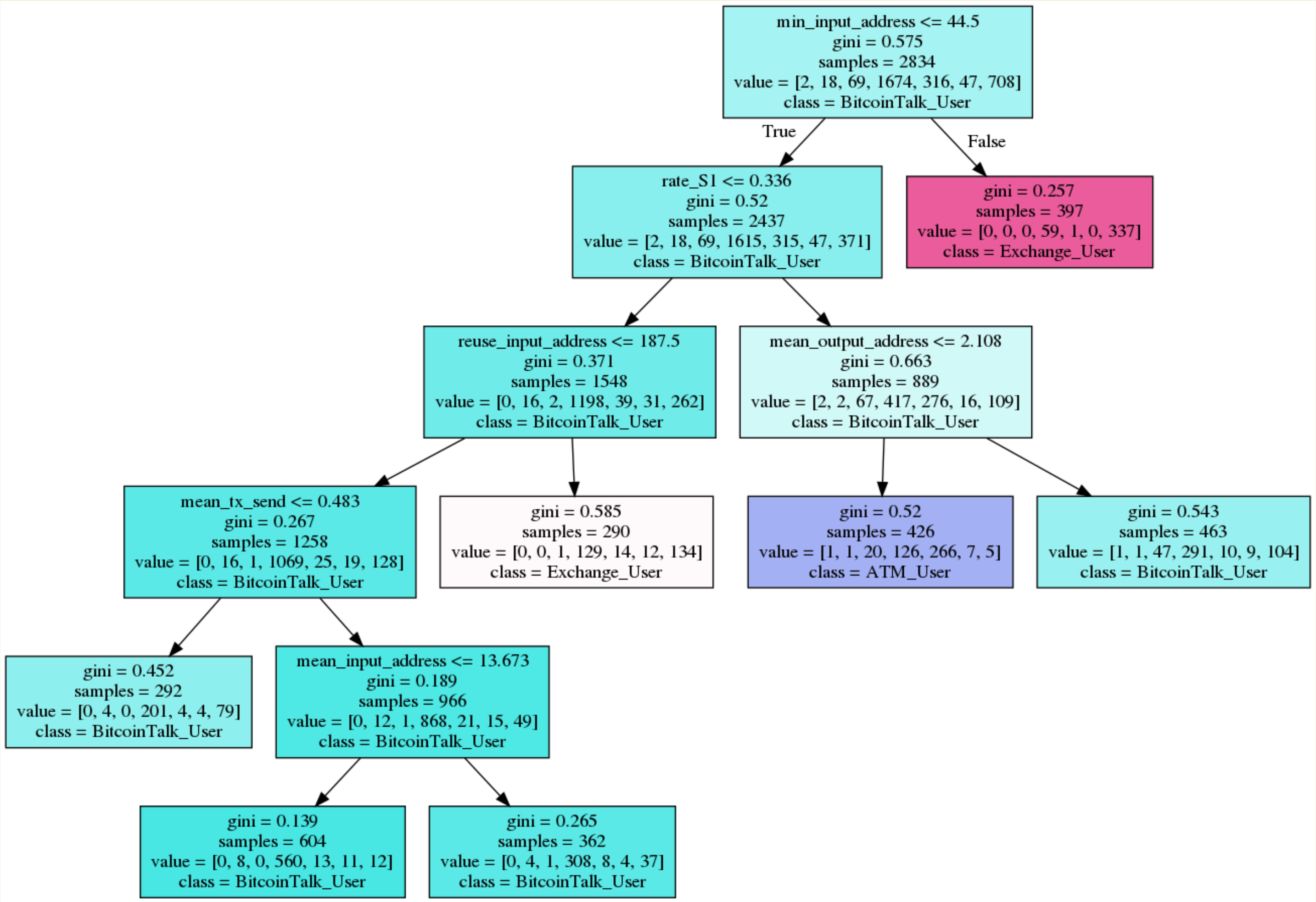
- Results of classification

usage	accuracy[%]		precision[%]		recall[%]	
	provider	user	provider	user	provider	user
Bitcointalk BBS		77		65		63
Bitcoin ATM	99	91	16	45	22	40
Dark web	98	93	6	49	9	36
Exchange		85		80		79
Mining Pool	92		70		65	
Total	81		49		39	

Result 4-1. model of the decision tree

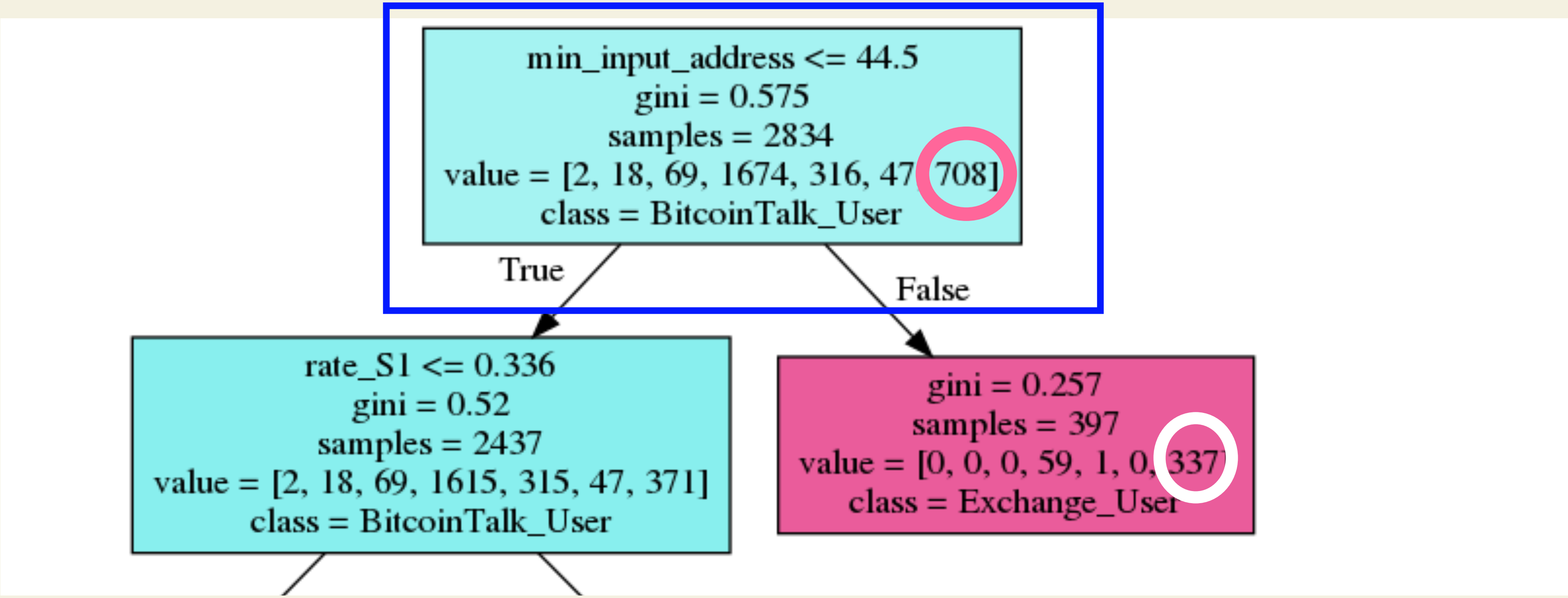
- Performed pruning
  - The highest depth is five.
  - No minor node consists of 10% of all instances.

- It is one of the sample models created 100 times.



Result 4-2. Root node feature

- Root node feature: The number of minimum input addresses
  - This feature is selected on almost all models.
  - In this model, about 48% (337/708) of Exchange addresses were classified as "Exchange users".



The number of minimum input addresses

- Address that "40 or more input addresses" are classified as "Exchange users" with probability of 48 % (486/1012).

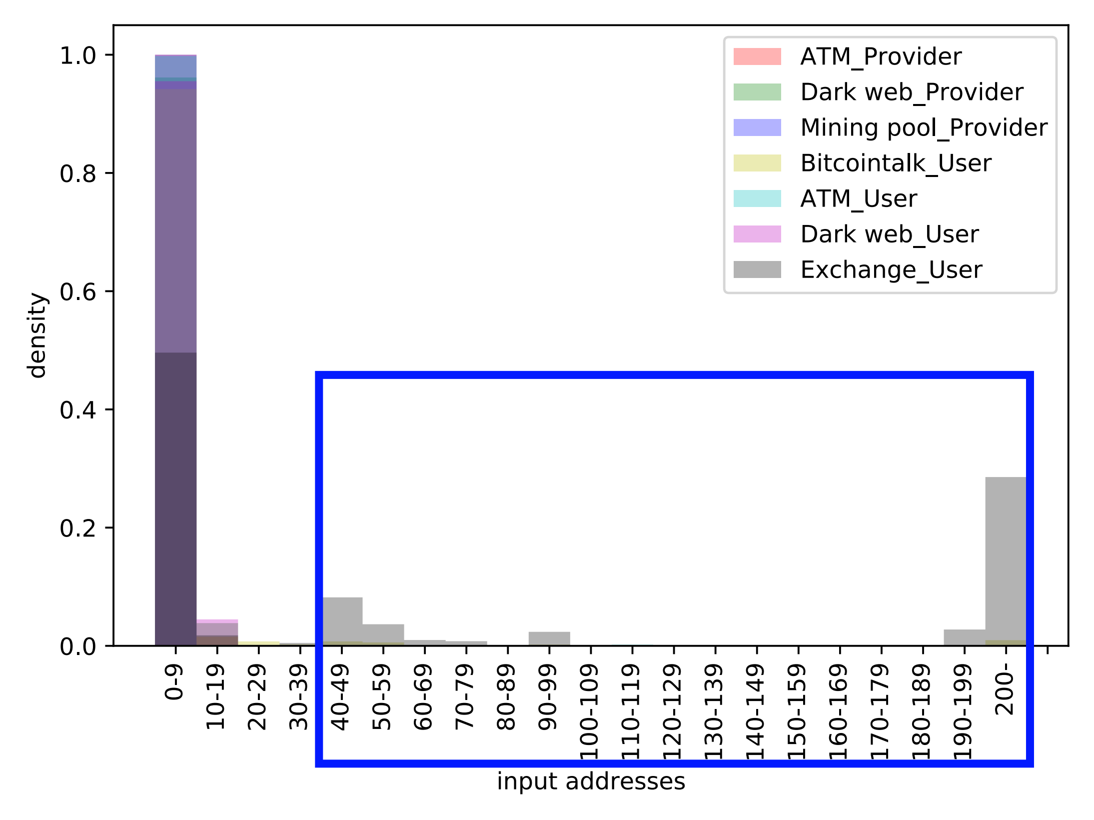


Fig. Histogram of features of the number of minimum input addresses in the seven usages

Table. Number of addresses indicating the number of minimum input addresses in the seven usages

usage		Avg.	Min.	Median	Max.	SD.
Bitcoin ATM	provider	1	1	1	1	0
Dark web		1.9	1	1	17	3.2
Mining Pool		1	1	1	1	0
Bitcointalk BBS	user	7	1	1	676	40.1
Bitcoin ATM		1.3	1	1	112	5.2
Dark web		1.7	1	1	12	2.3
Exchange		137.9	1	10.5	662	190

Research Questions

- Which is the easiest usage to classify for the seven usages?
  - Exchange user **is the highest of seven usages.**
  - accuracy 85%, precision 80%, recall 79%

- What is the most significant features to estimate the usage of Bitcoin addresses?
  - One of the most useful characteristics is **"the number of minimum input addresses".**
  - In this feature, about 48% of Exchange addresses were classified as "Exchange users".

Conclusion

- We found different transaction structures between providers and users.

- Our proposed algorithm estimates precisely the usages of unknown addresses with a accuracy of 80%.

- Future works
  - Our dataset addresses balanced in seven usages.
  - We consider to solve unbalanced and creating new learning model without dependence number of addresses.