

---

# Risk of Bitcoin Addresses to be Identified from Features of Output Addresses

Kodai Nagata<sup>1</sup> Hiroaki Kikuchi<sup>1</sup> Chun-I Fan<sup>2</sup>

1.Meiji University

2.National Sun Yat-sen University

# Cryptocurrency and Anonymity of Bitcoin

---

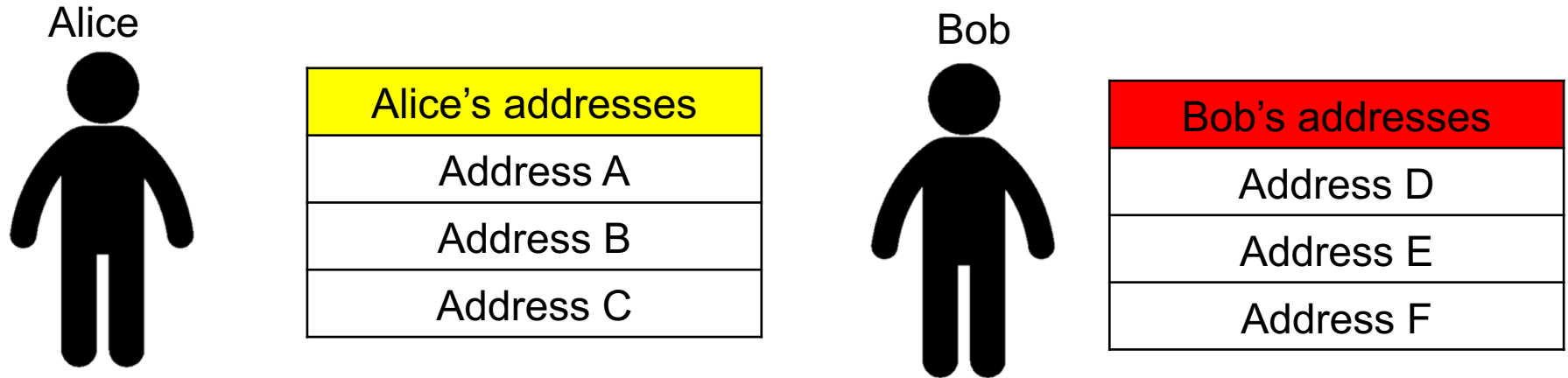
- Outflow incident of NEM
  - Do not arrest the criminal
  - Difficult to trace the stolen NEM



**Coincheck: World's biggest ever digital currency 'theft'**

<https://www.bbc.co.uk/news/world-asia-42845505>

# Bitcoin user and address



## Transaction



- Addresses are pseudonyms
- User are not be identified from an address(**anonymity**)

# Previous studies & Objectives

---

- Several studies have been done for the deanonymization of bitcoin addresses
  - Combined addresses managed by the same user[Sarah,2013]
  - Revealed a target user's **time zone** by analyzing the time distribution of transactions[Dupont,2015]
- How much anonymous Bitcoin address is?

# Problem

---

- We have no ground truth
  - Nobody knows who has which address
  - It is impossible to identify the owner from an address

# Our Solutions

- Collected the ground truth in two ways

1. Addresses that have been published via website **Bitcointalk**
2. Addresses that have been specified by the **coinbase**

The screenshot shows a user profile for 'macbook-air' on Bitcointalk. The profile includes the following information:

- Name:** macbook-air
- Posts:** 324
- Activity:** 324
- Merit:** 250
- Position:** Sr. Member
- Date Registered:** May 30, 2011, 01:02:02 AM
- Last Active:** September 02, 2017, 08:29:08 AM
- ICQ:**
- AIM:**
- MSN:**
- YIM:**
- Email:** hidden
- Website:** F2Pool
- Current Status:** Offline
- Bitcoin address:** 1KFHE7w8BhaENAswwryaoccDb6qcT6DbYY
- Gender:** Male
- Age:** N/A
- Location:** China
- Local Time:** February 05, 2018, 02:20:59 PM
- Trust:** 0: -0 / +0

Addr	Name	Location
1KFHE7w8BhaENAswwryaoccDb6qcT6DbYY	macbook-air	China
1DNNERMT5MMusfYnCBfcKCBjBKZWB C5Lg2	BitHits	None
1Anduck6bsXBXH7fPHzePJSXdC9AEsRmt4	Anduck	None

Profile page in Bitcointalk

# What is coinbase?

## Example of block

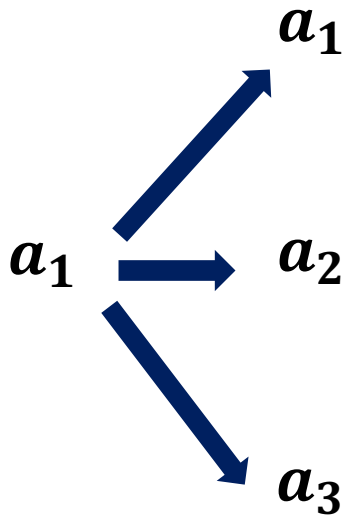
ID	Input	Output	Remittance[10 <sup>-8</sup> ]
<b>Coinbase</b> → $Tx_1$	$N/A$	$a_2$	2500000000 ← <b>Reward</b>
$Tx_2$	$a_2$	$a_4$	900000
$Tx_3$	$a_3$	$a_2, a_3$	6000000
$Tx_4$	$a_2, a_2, a_5$	$a_1, a_2$	110000000
$Tx_5$	$a_3$	$a_1, a_2, a_3, a_5$	40000000

# Propose method(Jaccard re-identification)

---

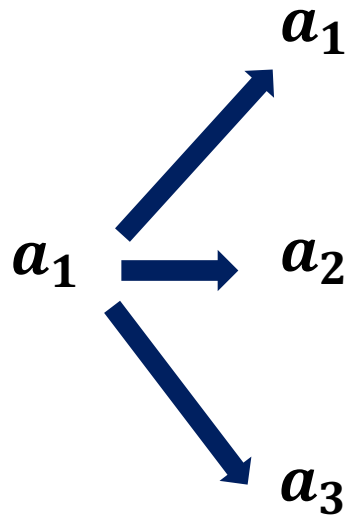
- Note that an output addresses

March



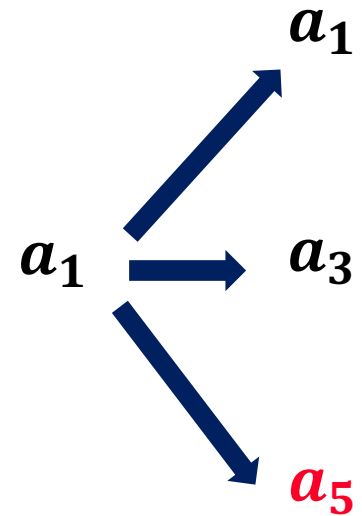
$\{a_1, a_2, a_3\}$

April



$\{a_1, a_2, a_3\}$

May



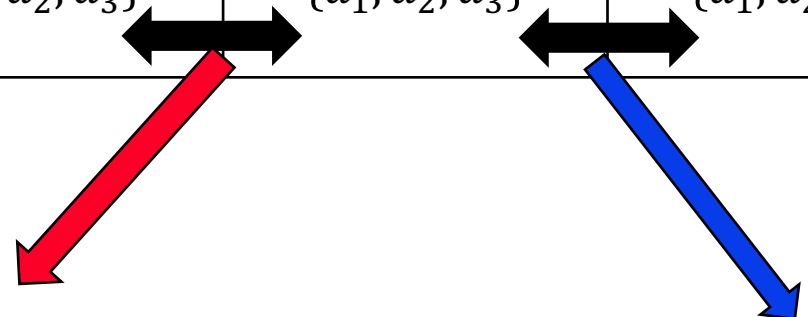
$\{a_1, a_3, a_5\}$



# Jaccard coefficient

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} : \text{Jaccard coefficient}$$

	March	April	May
$a_1$	$\{a_1, a_2, a_3\}$	$\{a_1, a_2, a_3\}$	$\{a_1, a_2, a_5\}$

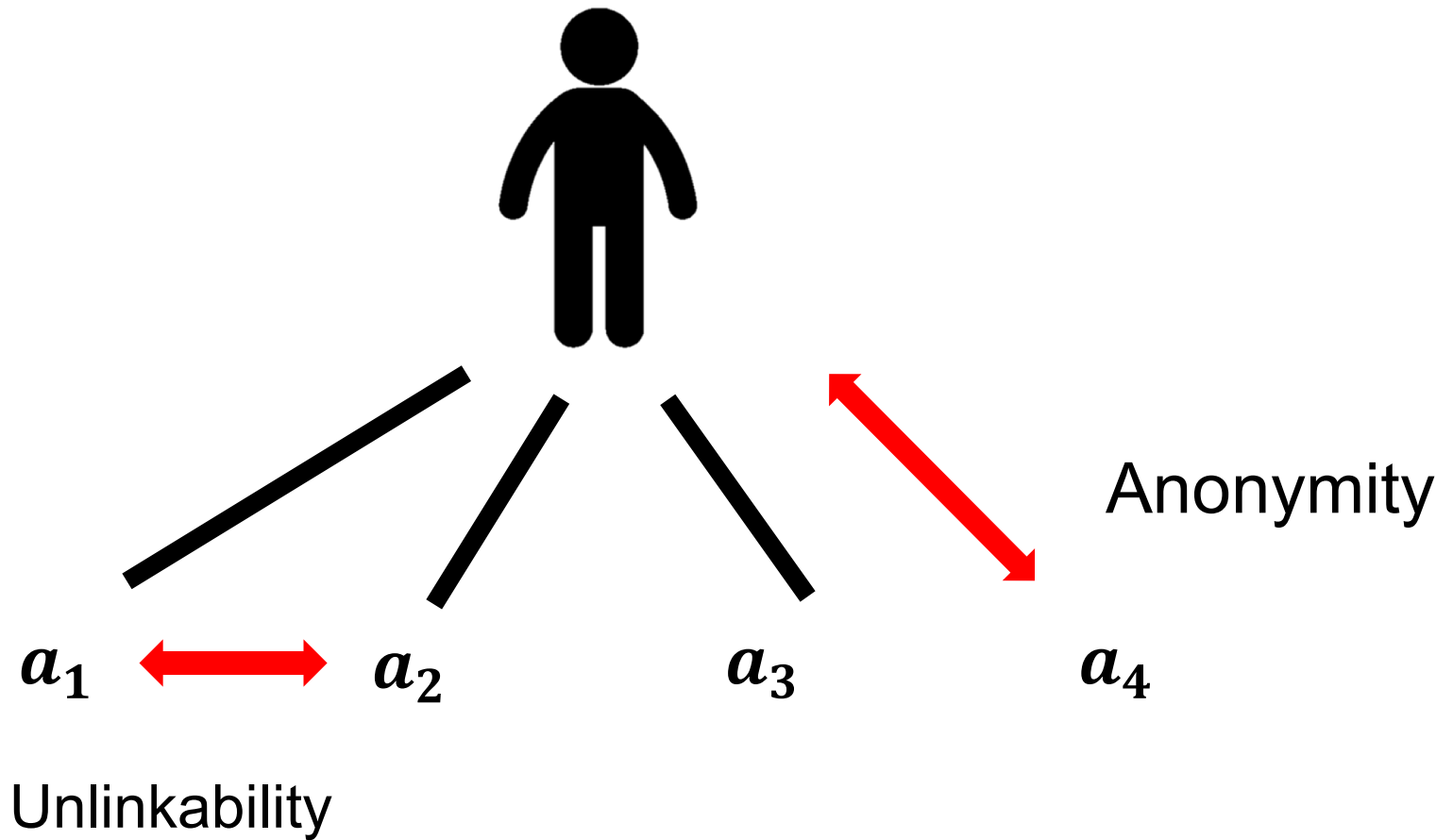


$$\frac{|\{a_1, a_2, a_3\}|}{|\{a_1, a_2, a_3\}|} = 1$$

$$\frac{|\{a_1, a_2\}|}{|\{a_1, a_2, a_3, a_5\}|} = \frac{1}{2}$$

# Anonymity and Unlinkability

---



# Research questions

---

1. Does the number of transactions for an address affect an unlinkability?
2. Which is more unlinkable output address set or transaction time set [Dupont,2015]?
3. How often is address identified?

# Experimental method

---

1. Divide transaction data into training data and test data
2. Predict the answer of the address of the test data by Jaccard re-identification
3. Calculate Recall, Precision and re-identification rate

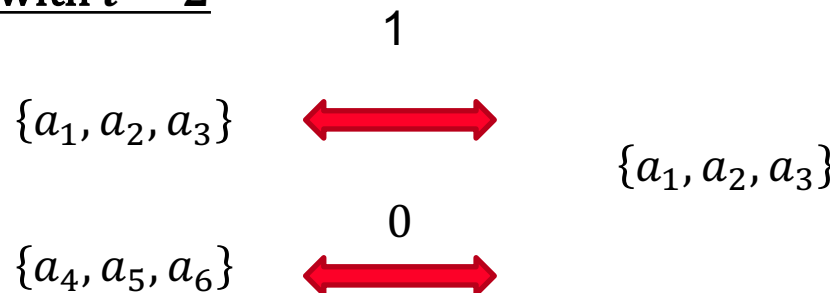
<b>Period</b>	2012.09.22 – 2014.05.10(About 1.5 years)
<b>Address number</b>	559
<b>Block number</b>	200,001 – 300,000(100,000 blocks)

# Jaccard re-identification

Term $i$	1	2	3
	7 months	7 months	7 months
$a_1$	$\{a_1, a_2, a_3\}$	$\{a_1, a_2, a_3\}$	$\{a_1, a_2, a_5\}$
$a_2$	$\{a_4, a_5, a_6\}$	$\{a_2, a_4, a_5\}$	$\{a_2, a_4, a_6\}$
	Training data	Test data	

} Sets of output addresses

**Sample: about  $a_1$  with  $i = 2$**



We predict the first one was sent from the same user to the training set

# Average recall, average precision and re-identification rate

---

- Average recall  $R$

$$R = \frac{1}{n} \cdot \sum_{i=1}^n R_i$$

- Average precision  $P$

$$P = \frac{1}{n} \cdot \sum_{i=1}^n P_i$$

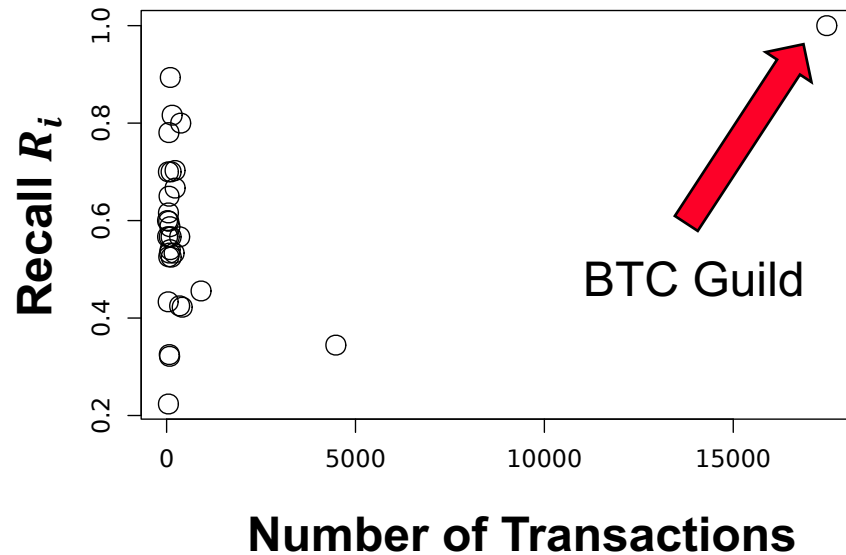
- Re-identification rate

$$F = \frac{2 \cdot R_i \cdot P_i}{R_i + P_i}$$

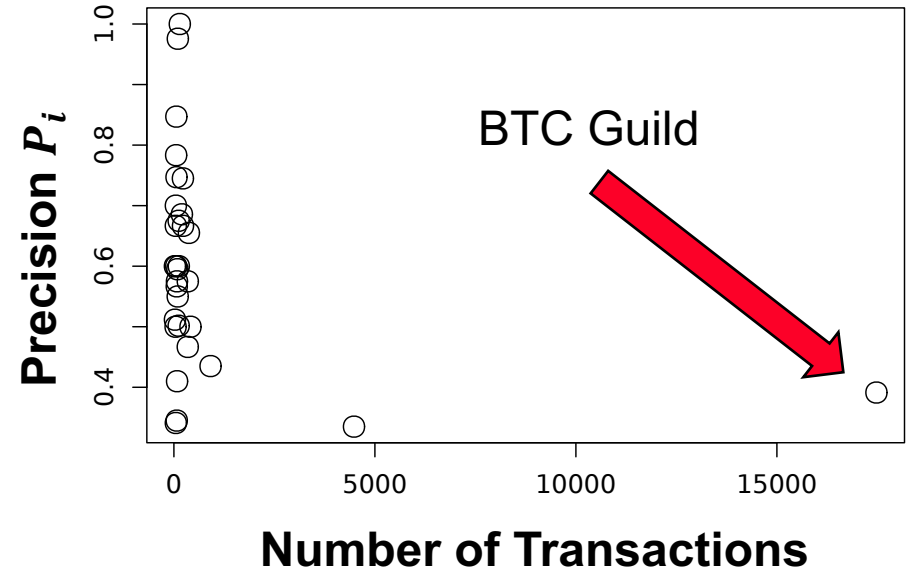
An address can be re-identified if the F-measure is more than **0.5**

# Result1 : Recall and precision by number of transactions in address

Scatterplot of Recall

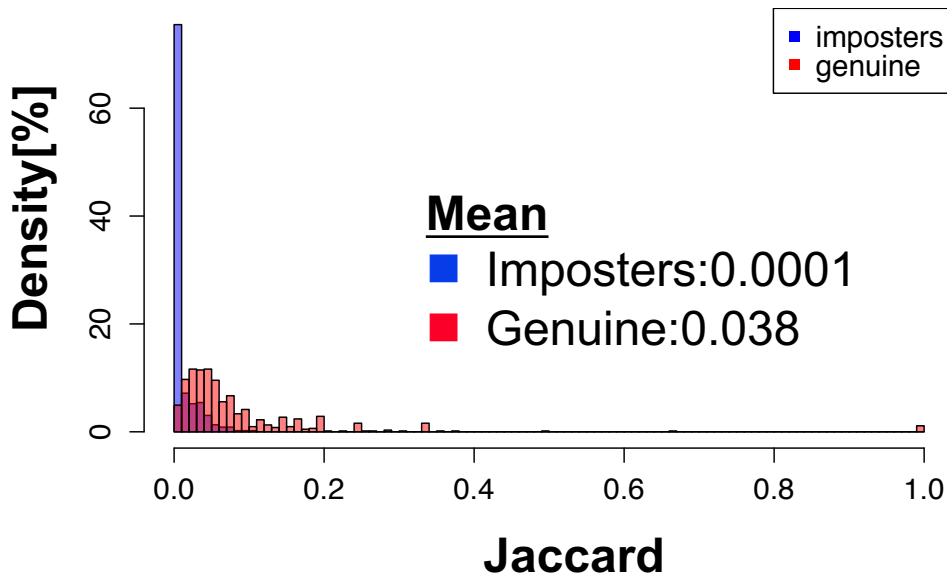


Scatterplot of Precision



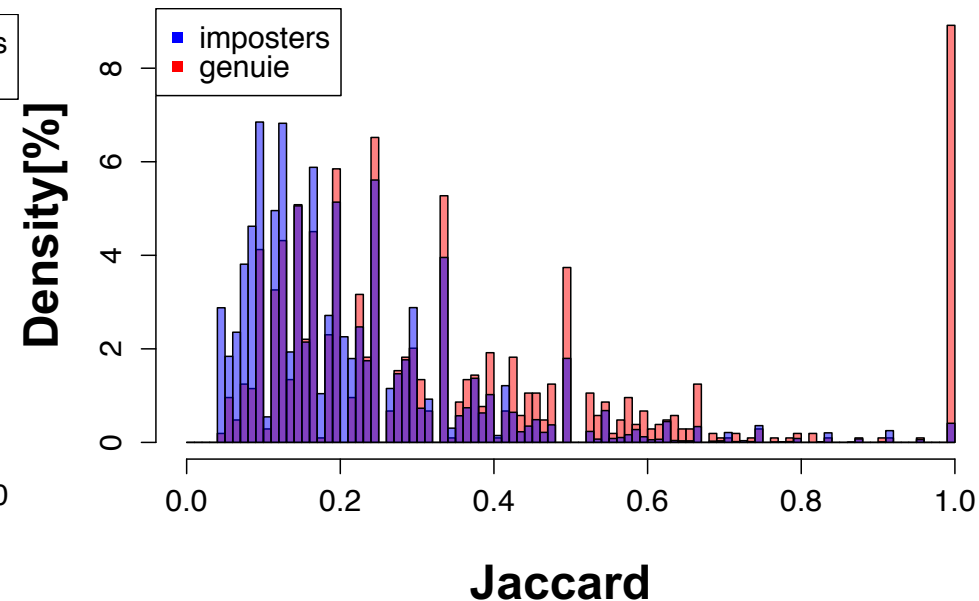
# Result2 : Comparison of output address set and time set

## Jaccard coefficient of output set



- The blue one is distributed within very small value

## Jaccard coefficient of time set

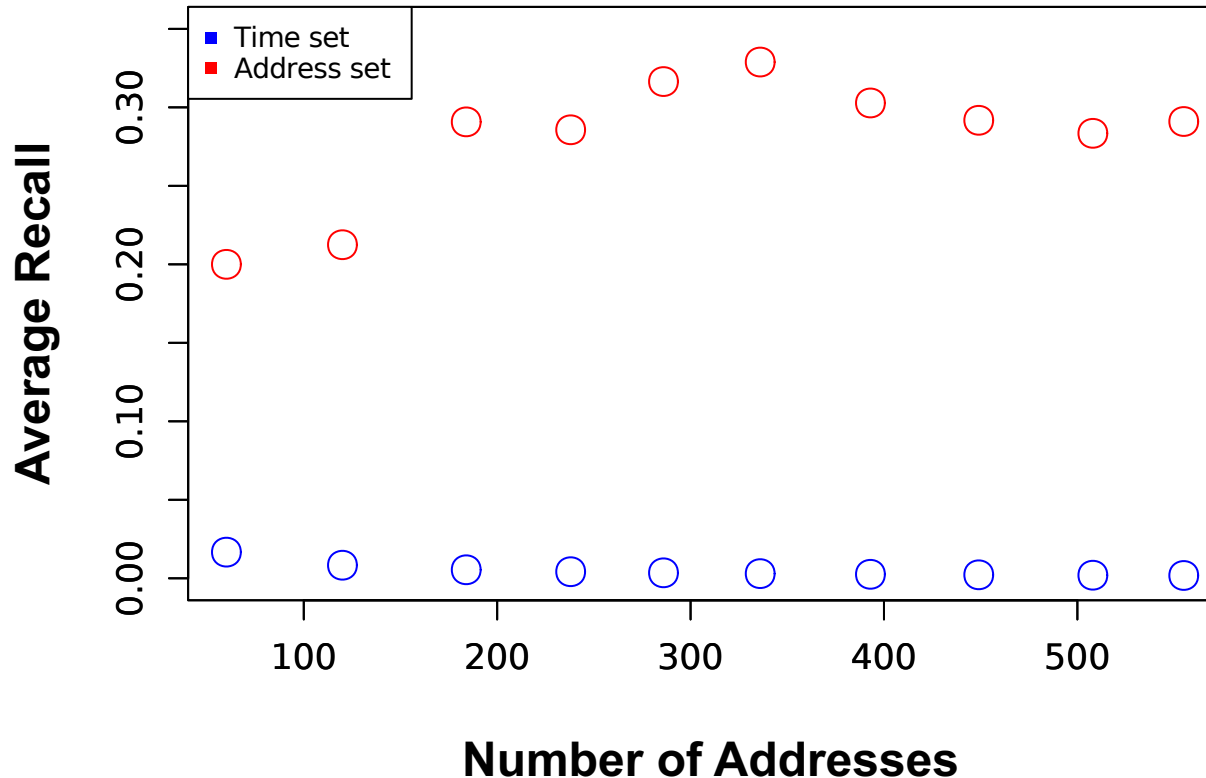


- The sets of transaction time are widely distributed



# Result3 : The average recall with respect to the number of addresses

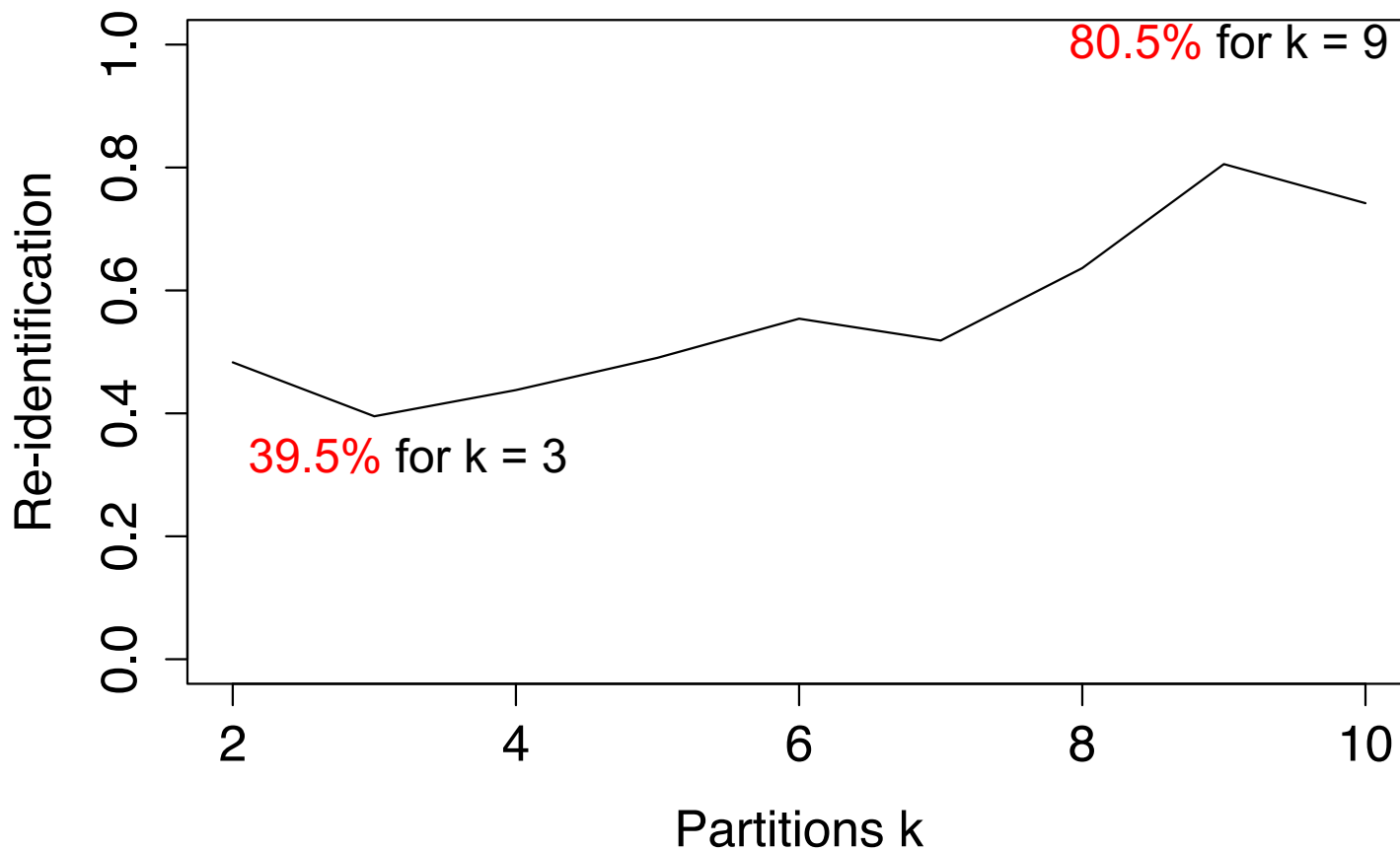
---



- The average recall of the output address set is independent from the number of addresses.

## Result4 : Distribution of re-identification rate with respect to the number of partitions k

---



# Research questions

---

1. **Does the number of transactions** for an address affect a n unlinkability?  
→ No
2. Which is more unlinkable, **output address set or transaction time set** [Dupont,2015]?  
→ output address set
3. **How often** is address identified?  
→80.5%

# Conclusions

---

- An unlinkability **is not affected** by number of transactions
- Output address set affected an unlinkability by **30%** more than time set
- **80.5%** of addresses can be identified from the Jaccard coefficient between subsets of output addresses