2019年2月1日

修士論文発表会

Bitcoinアドレスの送金先集合に基づく匿名性の評価

永田　倖大

菊池研究室

暗号通貨と匿名性

- Bitcoinウォレット数
  - 4年で16倍になっている[1]

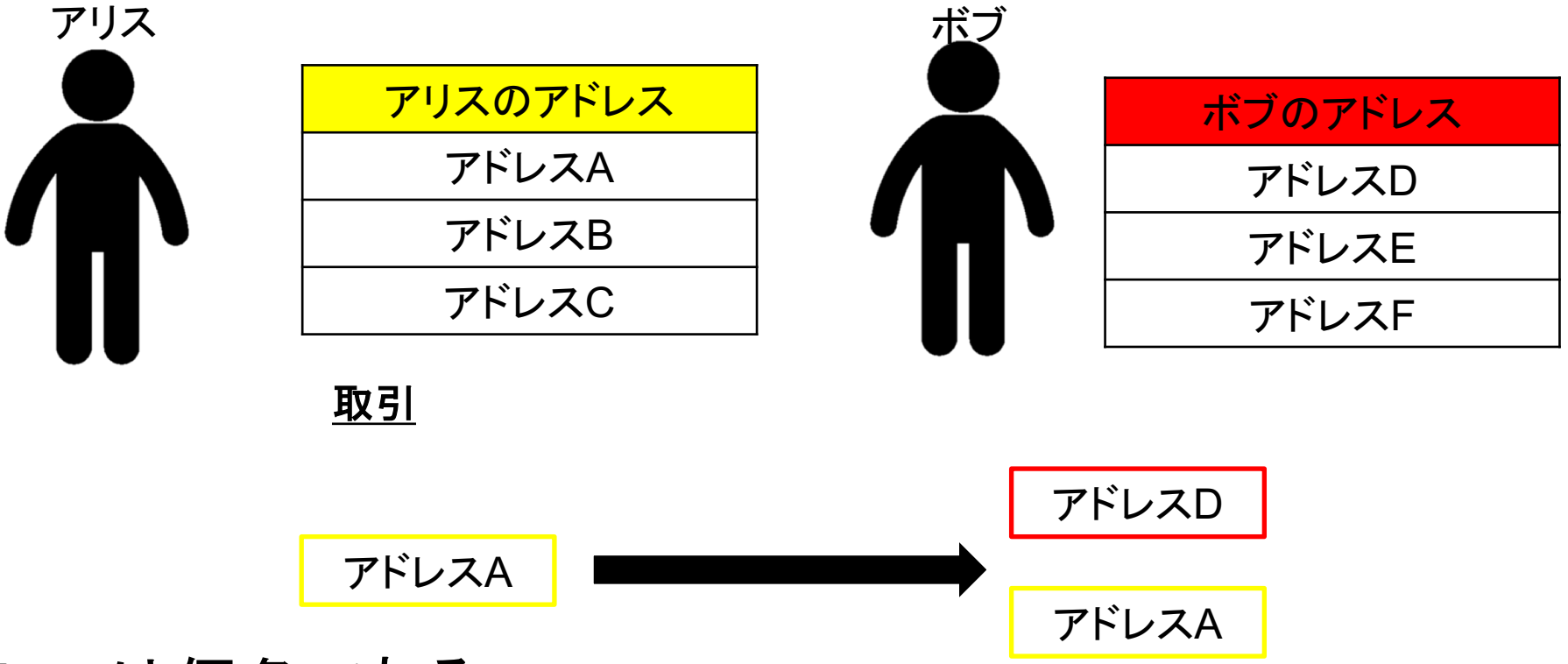
Coincheck: World's biggest ever digital currency 'theft'

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

- コインチェックのNEM流出
  - 誰が盗んだか不明
  - 行方を追うのは難しい

[1]仮想通貨取引についての現状報告https://www.fsa.go.jp/news/30/singi/20180410-3.pdf

Bitcoinのユーザとアドレス



- アドレスは仮名である

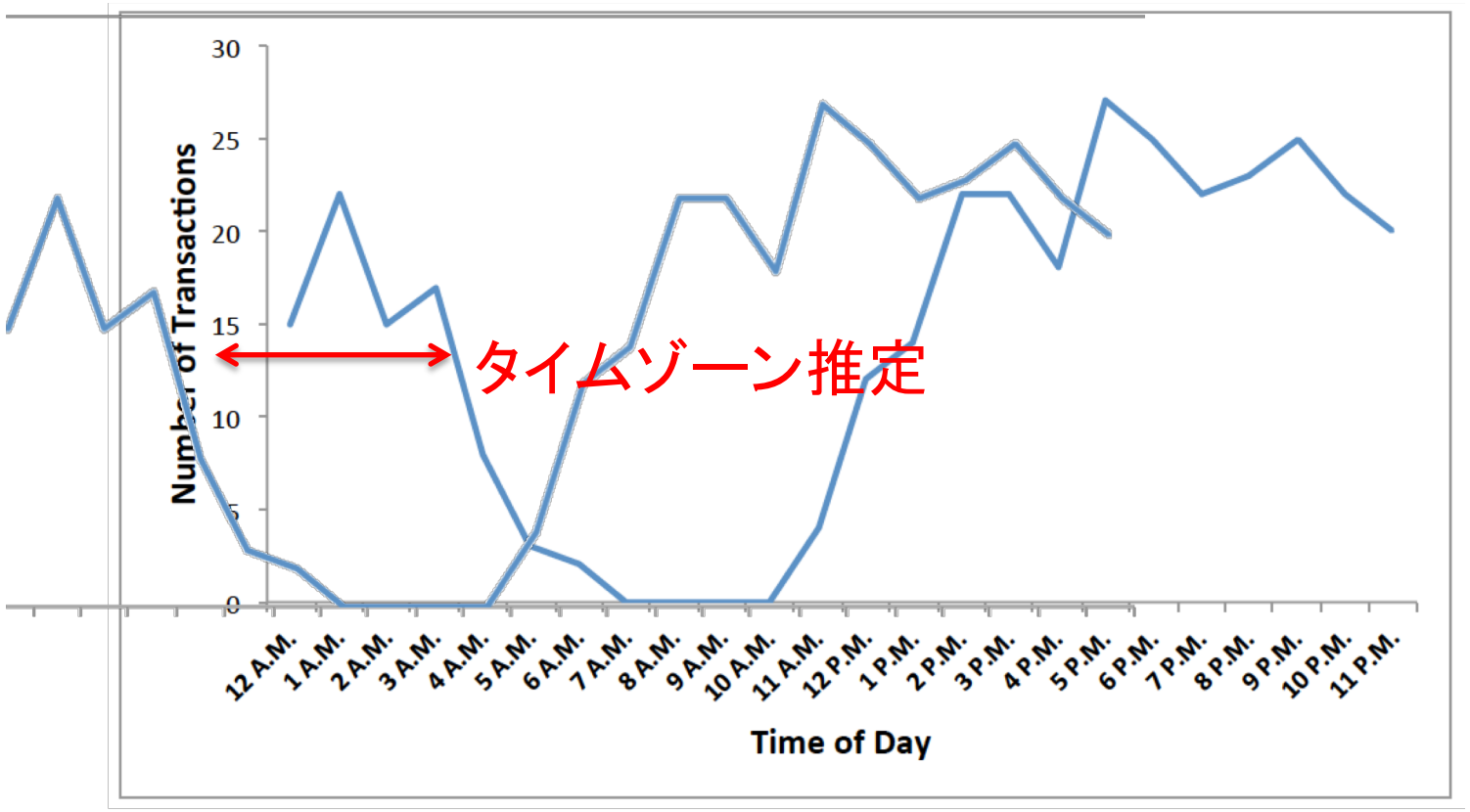
  例：1MBRNCK8HYveyWsZqXpERLQ5H75Uj9CBi4

- アドレスからユーザを識別することはできない（匿名性）

先行研究・研究目的

- 先行研究
  - 同一ユーザが管理するアドレスを識別[Sarah,2013]
  - アドレス管理者のタイムゾーンを特定[Dupont,2015]
- 目的
  - Bitcoinアドレスの匿名性を明らかにする

先行研究[Dupont 2015]



[2] p.2 Figure1より転載

[2]J. Dupont, A. C. Squicciarini. Toward De-Anonymizing Bitcoin by Mapping Users Location. In Proceedings of Conference on Data and Application Security and Privacy (CODASPY'15). ACM, 2015

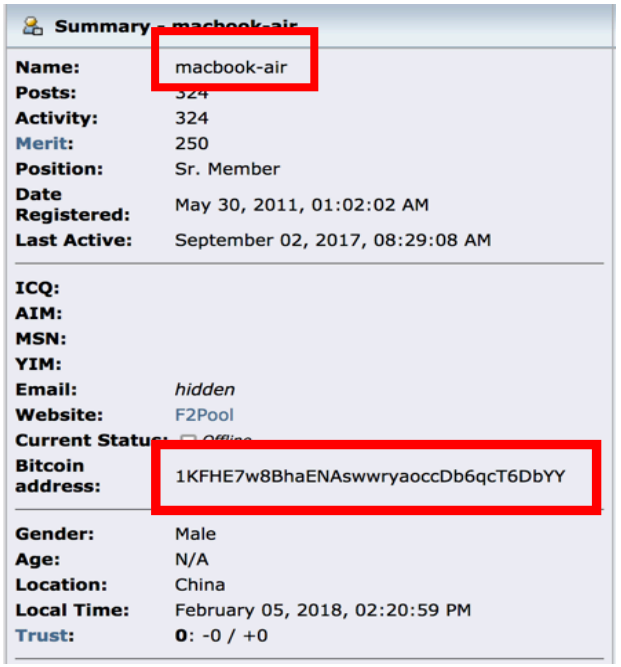
問題点

- 正解データが分からない
  - アドレスとオーナーの正しい関係がわからない
  - アドレスからオーナーを識別できない

アドレスデータ

- 匿名性を評価するために以下の2種類の方法で取得した

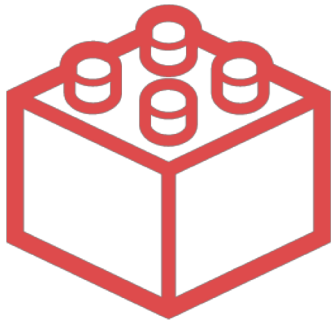
1. Bitcointalkで公開されているアドレス
2. コインベース の出力で指定されたことのあるアドレス



Profile page in Bitcointalk

Addr	Name
1KFHE7w8BhaENAswwryaoccDb6qcT6DbYY	macbook-air
1DNNERMT5MMusfYnCBfcKCBjBKZWBC5Lg2	BitHits
1Anduck6bsXBXH7fPHzePJSXdC9AEsRmt4	Anduck

ブロック



- 取引情報などが格納されている
- ブロックが生成されることで取引承認
- ブロック生成には膨大な計算が必要

コインベース

ブロック内の取引情報

ID	入力	出力	送金額[10^{-8}]
Tx_1	N/A	a_2	1250000000
Tx_2	a_2	a_4	900000
Tx_3	a_3	a_2, a_3	60000000
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

コインベース → Tx_1

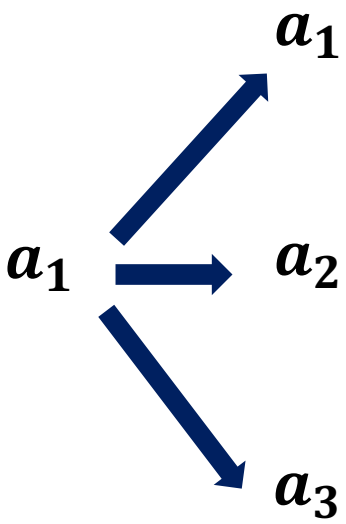
報酬 ← 1250000000

- ブロック作成（マイニング）の報酬

提案識別方式(Jaccard再識別)

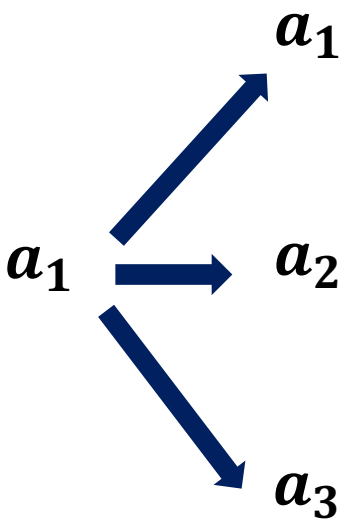
- 取引先アドレスに注目

3月



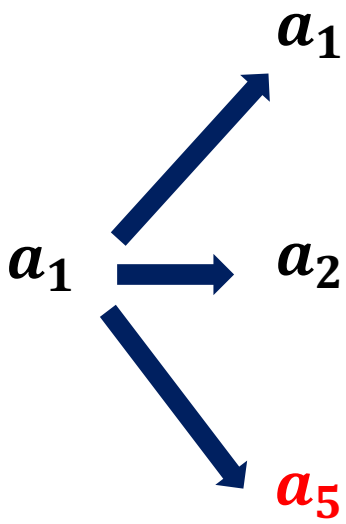
$\{a_1, a_2, a_3\}$

4月



$\{a_1, a_2, a_3\}$

5月



$\{a_1, a_2, a_5\}$

Jaccard係数

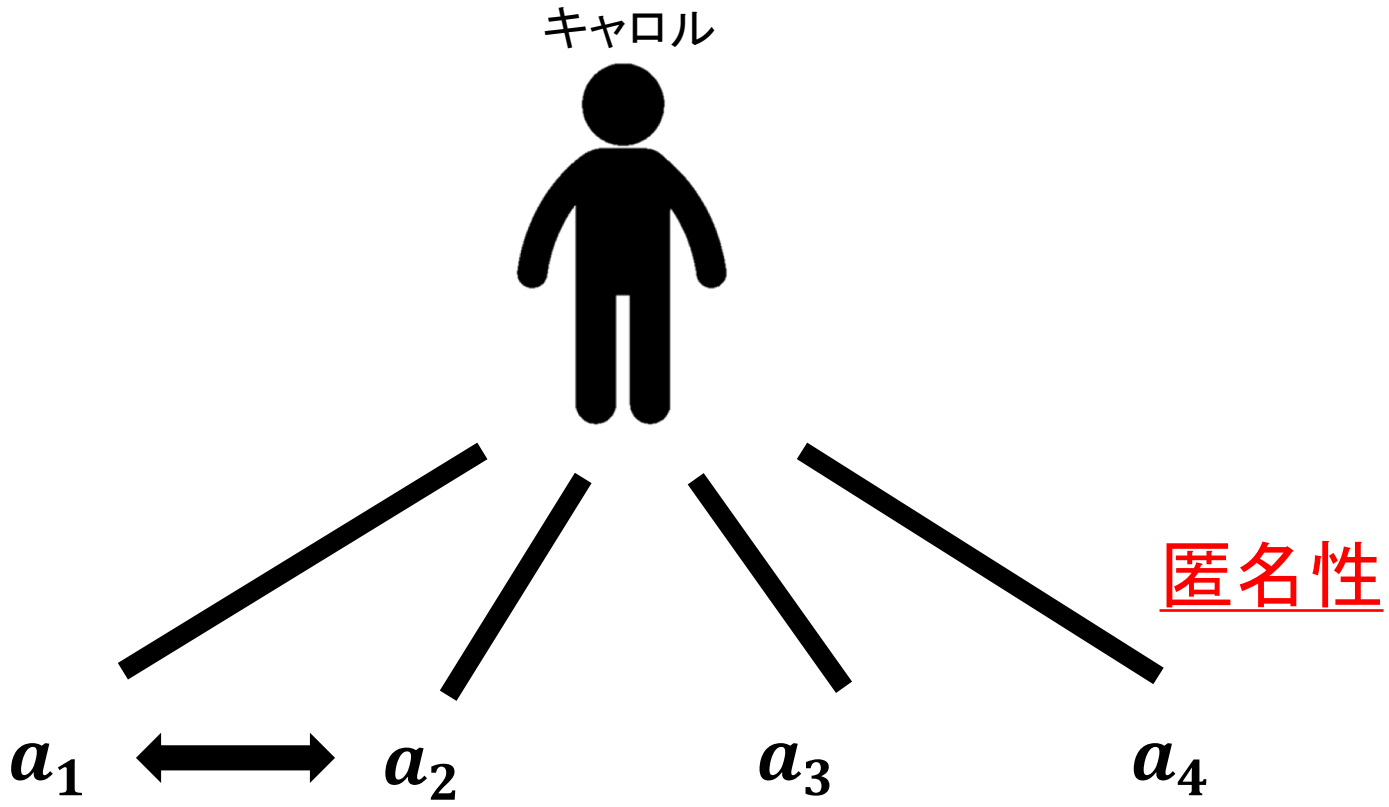
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$: Jaccard係数

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

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

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

匿名性と識別可能性



研究課題

1. 取引数は識別可能性に影響を与えるか？
2. 送金先集合と取引時刻集合[Dupont,2015]で識別可能性に影響を与えるのはどちらか
3. 識別されるリスクの大きさは？

実験方法

1. 取引データを任意の期間に分割し学習データ，評価データを作成
2. Jaccard再識別を用いて評価データのアドレスを予測
3. 再現率，適合率，識別率を求める

期間	2012.09.22 – 2014.05.10(約1.5年間)
アドレス数	559
ブロック	200,001 – 300,000(10万ブロック)

Jaccard再識別

Term i	1	2	3
	7 months	7 months	7 months
a_1	$\{a_1, a_2, a_3\}$ 1	$\{a_1, a_2, a_3\}$	$\{a_1, a_2, a_5\}$
a_2	$\{a_4, a_5, a_6\}$ 0	$\{a_2, a_4, a_5\}$	$\{a_2, a_4, a_6\}$
	Training data	Test data	

送金先集合

評価方法:平均再現率・平均適合率・識別率

- 平均再現率R

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

- 平均適合率P

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

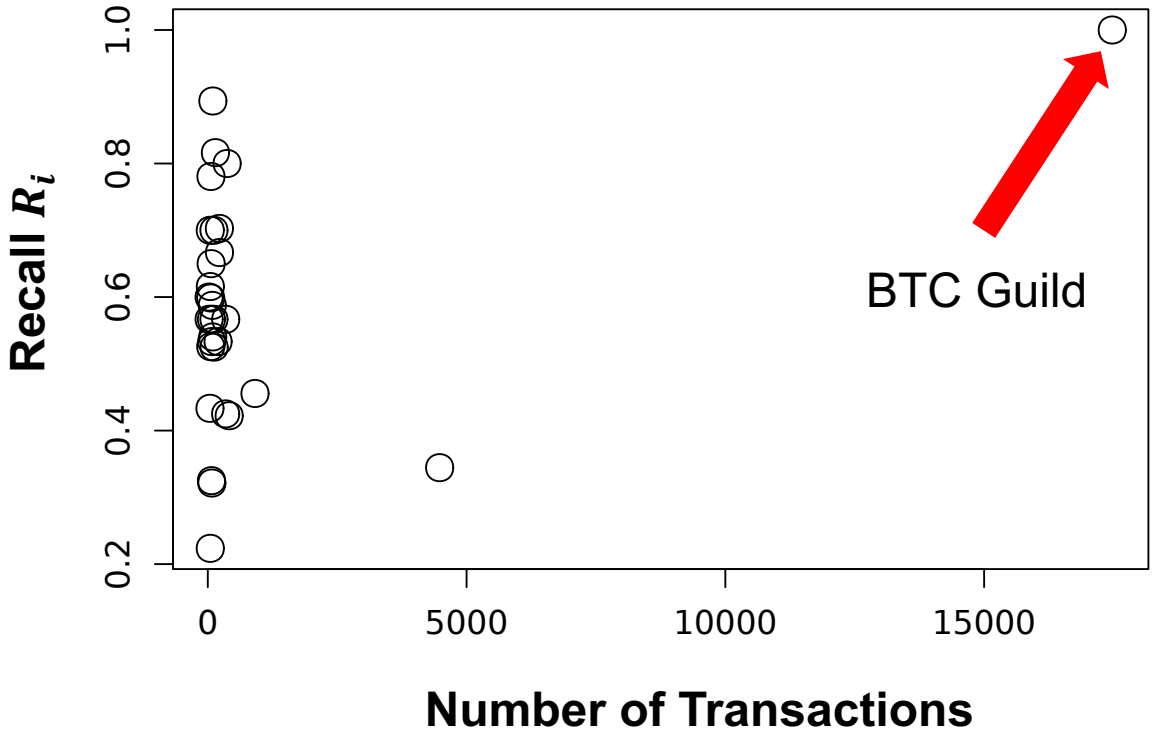
- アドレスにおける識別可能性の定義

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

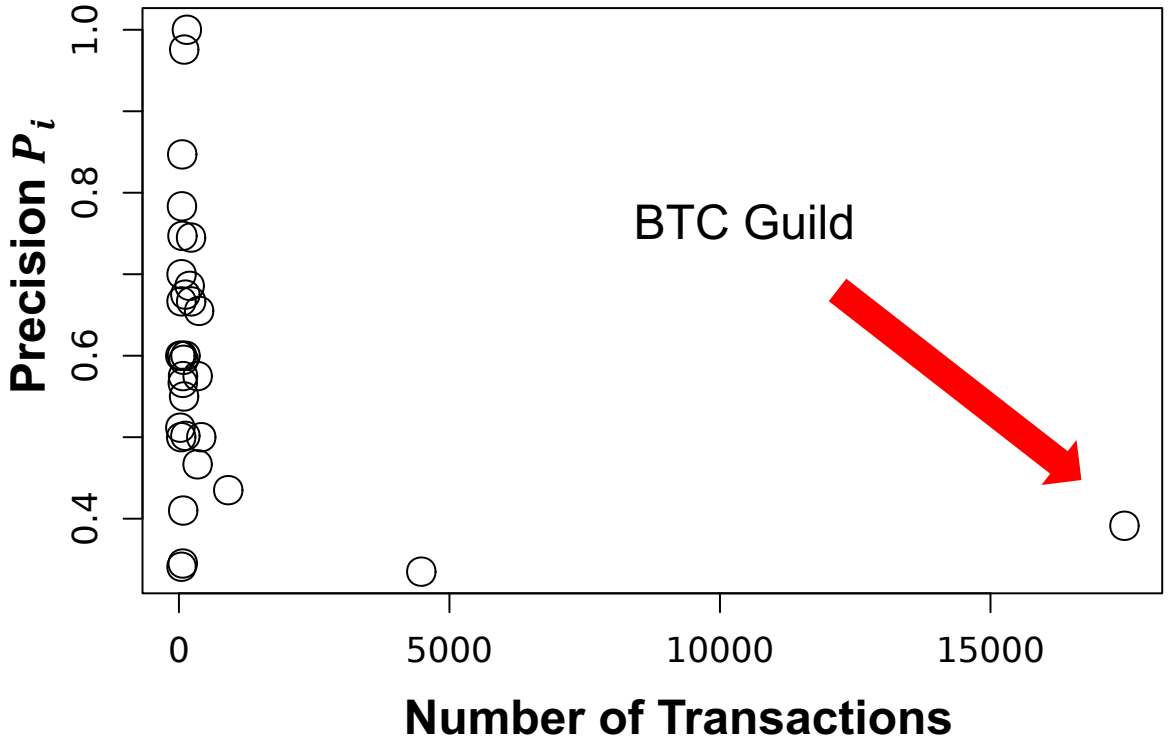
F値が0.5以上のアドレスを識別されたと定義する

実験結果1:取引数による再現率・適合率の変化

再現率の変化

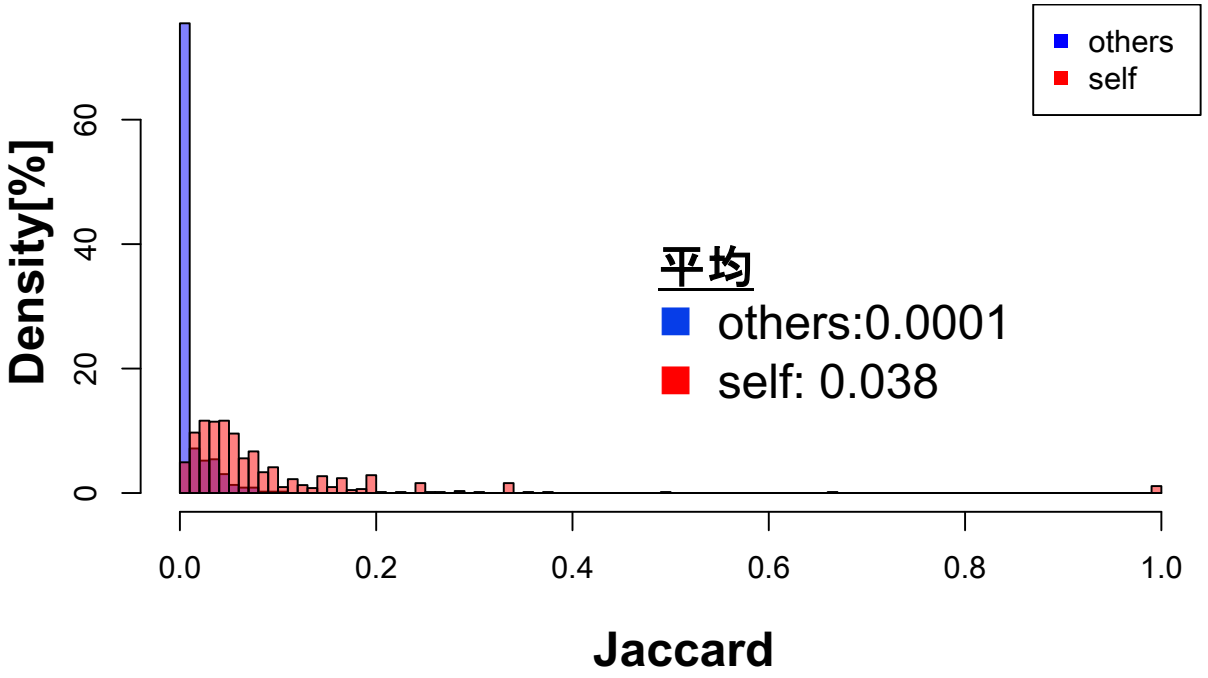


適合率の変化

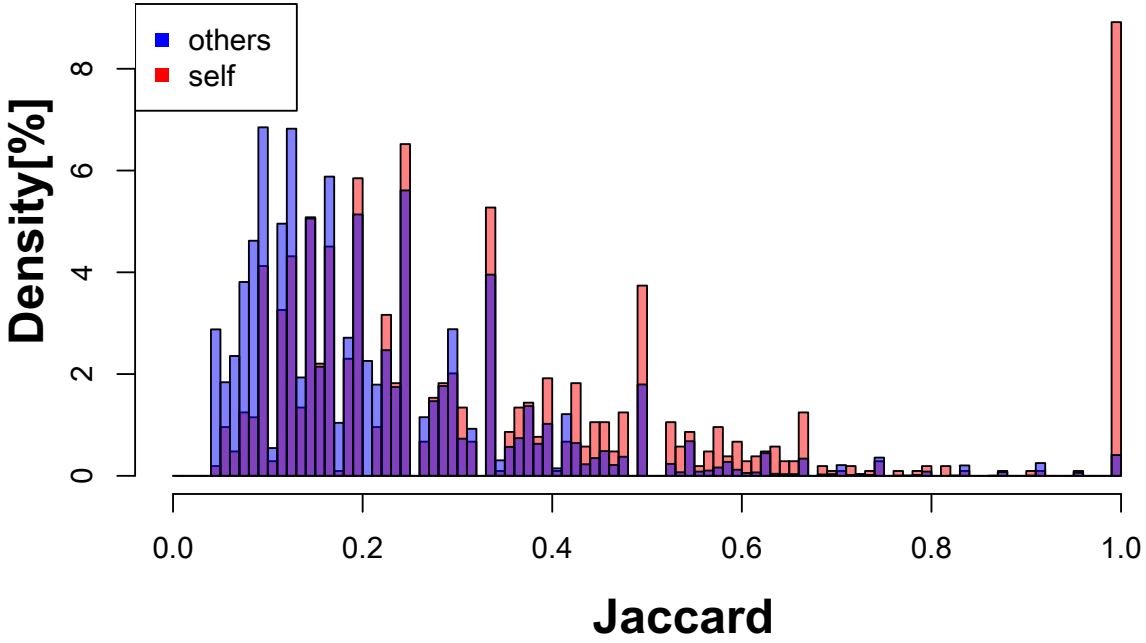


実験結果2:送金先集合と時間集合の比較

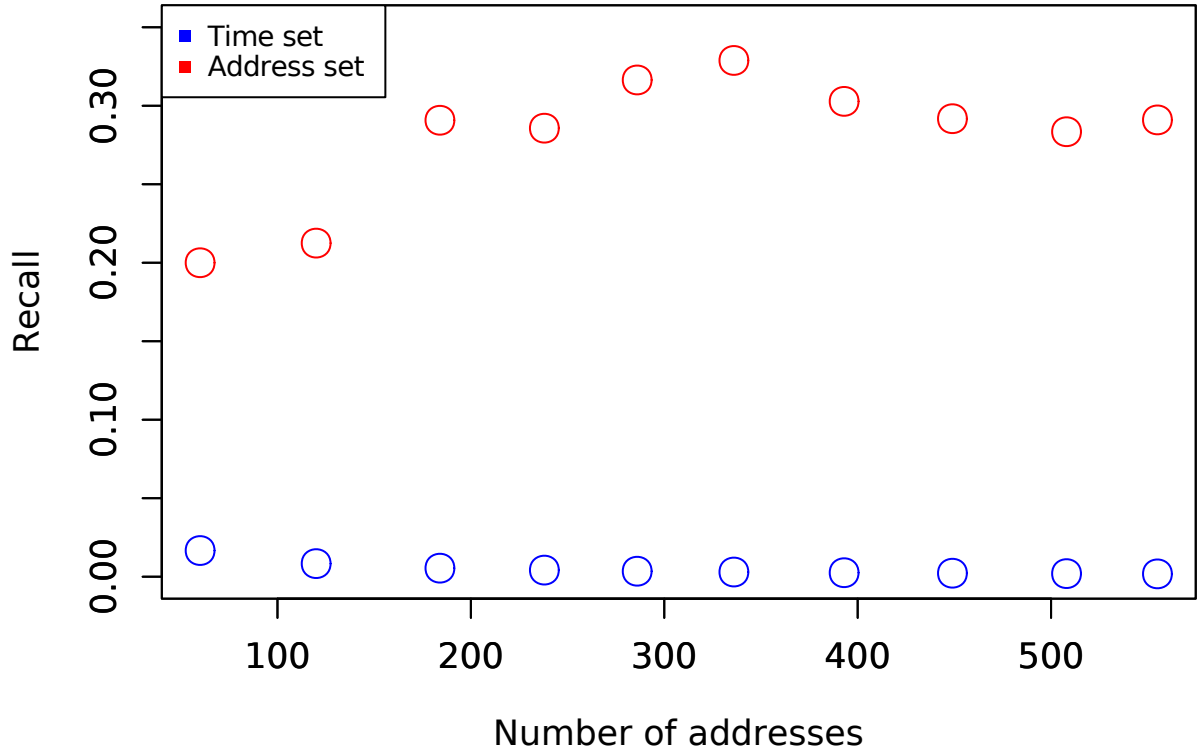
送金先集合jaccard



取引時刻集合jaccard

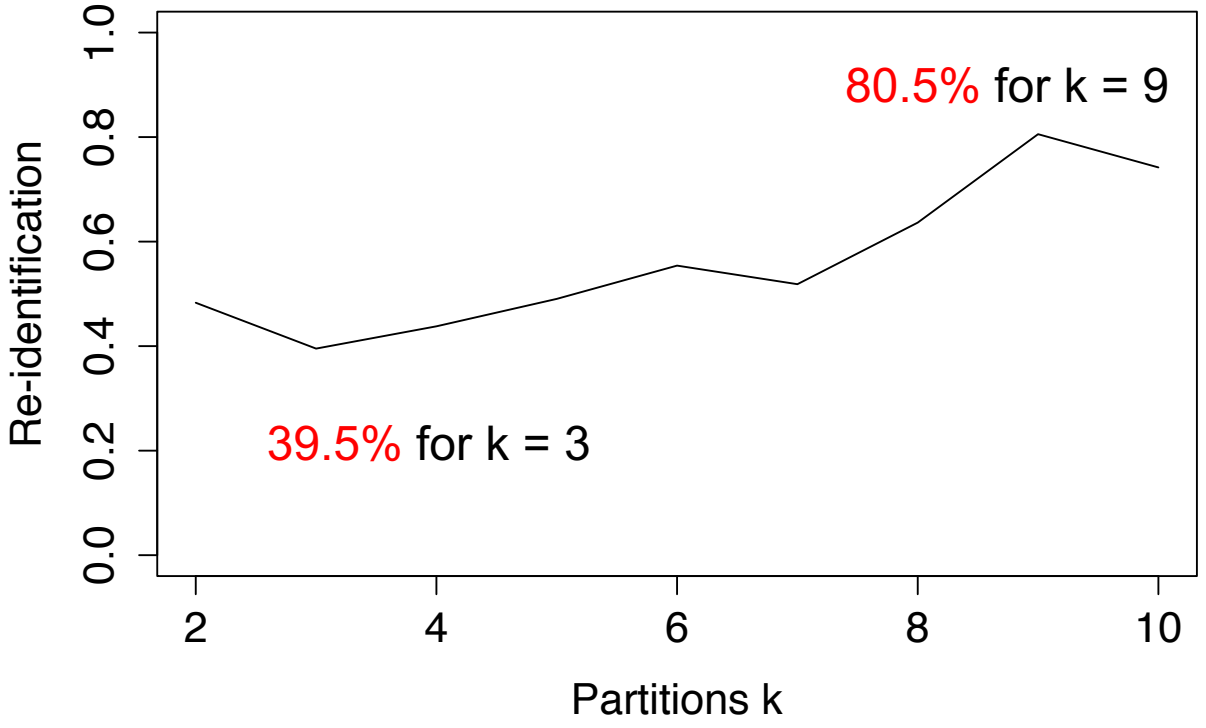


実験結果3:対象アドレス数による平均再現率



- 送金先集合は大きな変化は見られなかった
- 取引時刻集合はnに対して$\frac{1}{n}$の割合で再現率を下げる

実験結果4:分割数による識別率



- 分割数が増えるにつれて増加傾向

研究課題

1. 取引数は識別可能性に影響を与えるか？

   →与えない

2. 送金先集合と取引時刻集合[Dupont,2015]で識別可能性に影響を与えるのはどちらか

   →送金先集合

3. 識別されるリスクの大きさは？

   →80.5%

おわりに

- アドレスの取引数は識別可能性に影響を与えない
- 送金先集合は取引時刻集合に比べ，30%以上ほど大きく識別可能性に影響を与える
- 送金先集合を用いた再識別で，最大80.5%のアドレスが再識別されることを示した