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[1]　菊池浩明[2]

1. 明治大学大学院先端数理科学研究科

2. 明治大学総合数理学部

暗号通貨と匿名性

- 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

迫真 コインチェック騒動（３）「流出ＮＥＭ、追いかける」

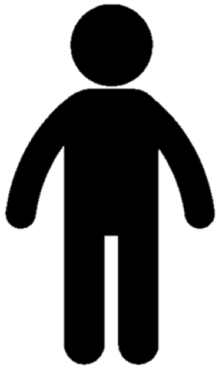
2018/3/1付

日本経済新聞(2018/3/1)

- Bitcoinの匿名性について先行研究
  - 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]

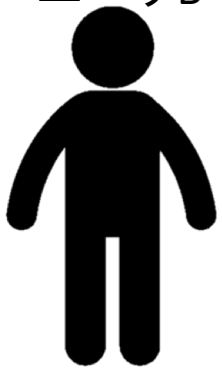
Bitcoinのユーザとアドレス

ユーザA



ユーザAのアドレス
アドレスA
アドレスB
アドレスC

ユーザB



ユーザBのアドレス
アドレスD
アドレスE
アドレスF

取引

- アドレスは仮名である
- アドレスからユーザを識別することはできない（匿名性）

研究目的

1. Bitcoinの匿名性を明らかにする
2. アドレスの仮名による匿名性はあるのか

先行研究[Dupont 2015]

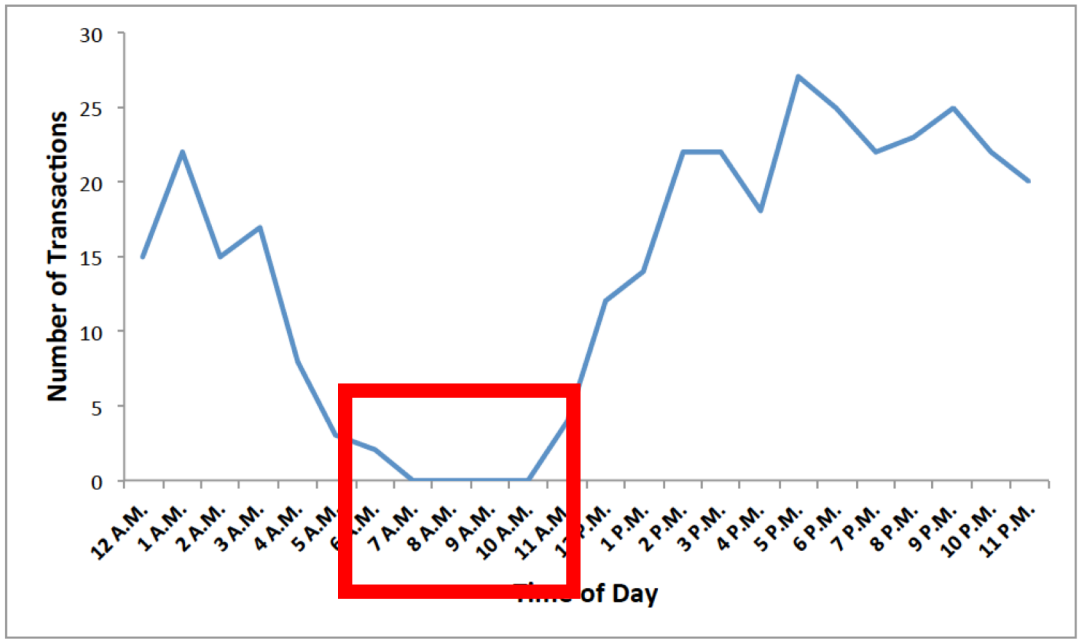


Figure 1: The time of day of a Bitcoin user's purchases, reported with the network's timezone (UTC 0).

{7 A.M., 8 A.M., 9 A.M., 10 A.M.}

↓

{UTC −2, UTC −3, UTC −4, UTC −5}

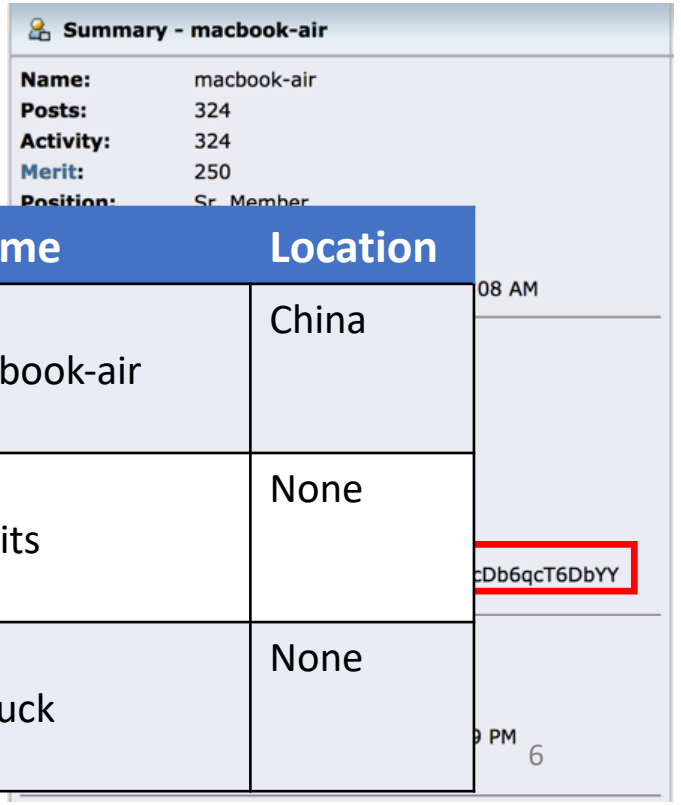
タイムゾーンを推定

問題点

- 正解データが分からない
  - アドレスとユーザの正しい関係がわからない
  - アドレスとユーザに関連がない

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



Addr	Name	Location
1KFHE7w8BhaENAswwryaoccDb6qcT6DbYY	macbook-air	China
1DNNERMT5MMusfYnCBfcKCBjBKZWBC5Lg2	BitHits	None
1Anduck6bsXBXH7fPHzePJSXdC9AEsRmt4	Anduck	None

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	2500000000
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

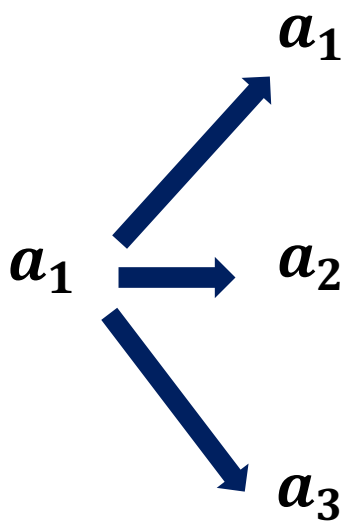
報酬

- ブロック作成（マイニング）の報酬

提案識別方式(jaccard再識別)

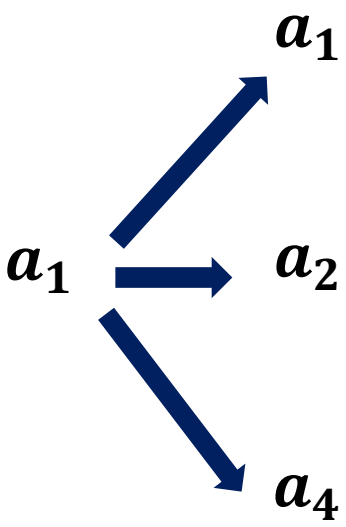
- 取引先アドレスに注目

3月



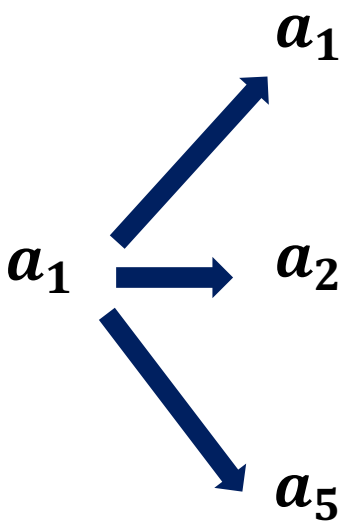
$\{a_1, a_2, a_3\}$

4月



$\{a_1, a_2, a_4\}$

5月



$\{a_1, a_2, a_5\}$

学習データ，評価データ

期間 i	1	2	3
	7 ヶ月	7 ヶ月	7 ヶ月
a_1	$\{a_1, a_2\}$	$\{a_3, a_4\}$	N/A
a_2	$\{a_2, a_5\}$	$\{a_4, a_5\}$	$\{a_5\}$
a_3	$\{a_3, a_6\}$	$\{a_4, a_6\}$	$\{a_5, a_7\}$
	学習データ	評価データ	

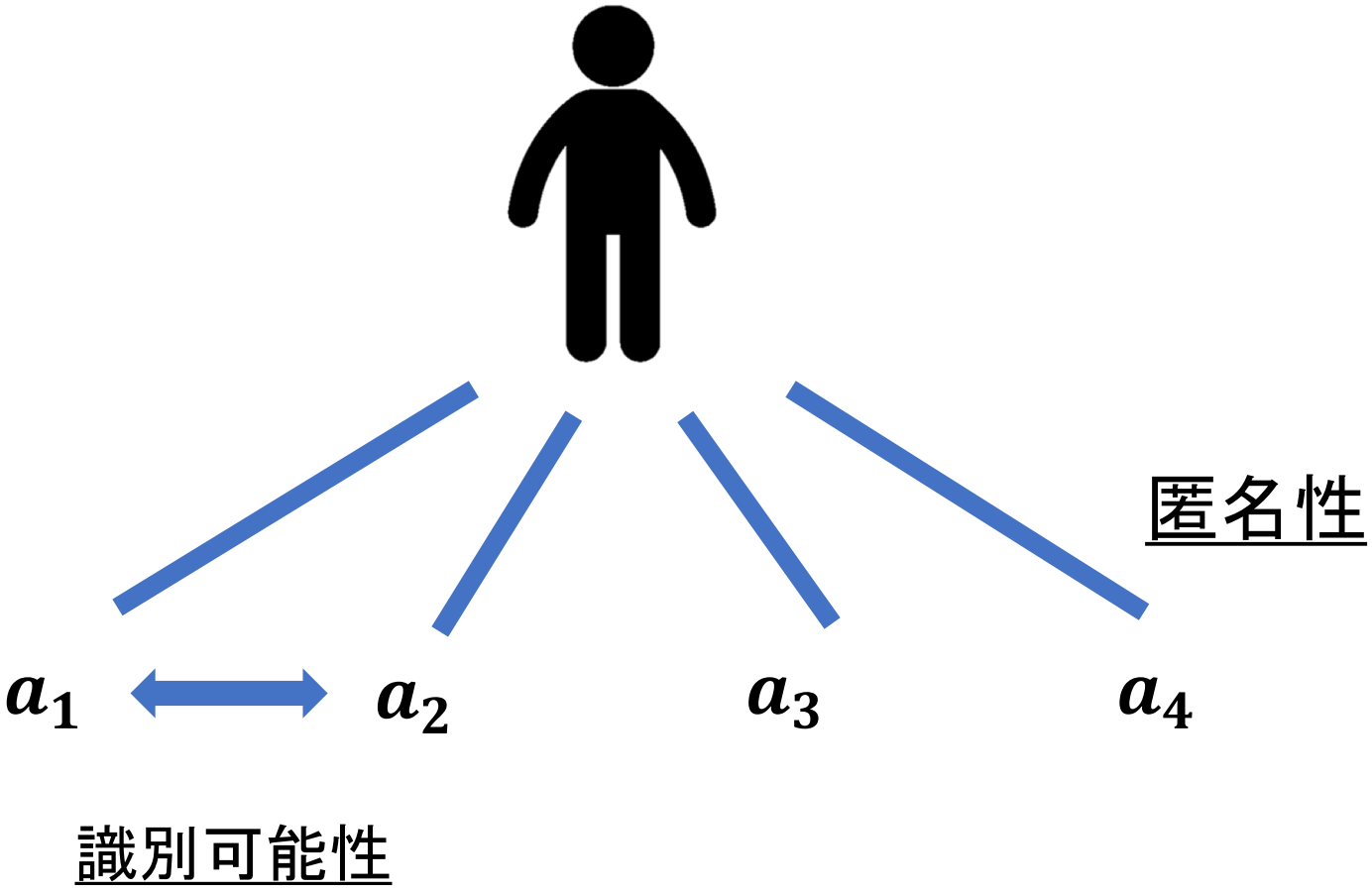
例：$i = 2$でのa_2について

$J(O_1(a_2), O_2(a_2)) = \frac{1}{3}$

$J(O_1(a_3), O_2(a_2)) = 0$

a_2を正解と予測

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

$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$: jaccard係数を用いた集合A,Bの類似度

評価方法：平均再現率・平均適合率・識別率

- 平均再現率R

$$R = \frac{1}{n} \cdot \sum_{a \in A} \frac{\text{正しく}a\text{と識別したデータ数}}{a\text{のデータ数}}$$

- 平均適合率P

$$P = \frac{1}{n'} \cdot \sum_{a \in A'} \frac{\text{正しく}a\text{と識別したデータ数}}{a\text{と識別したデータ数}}$$

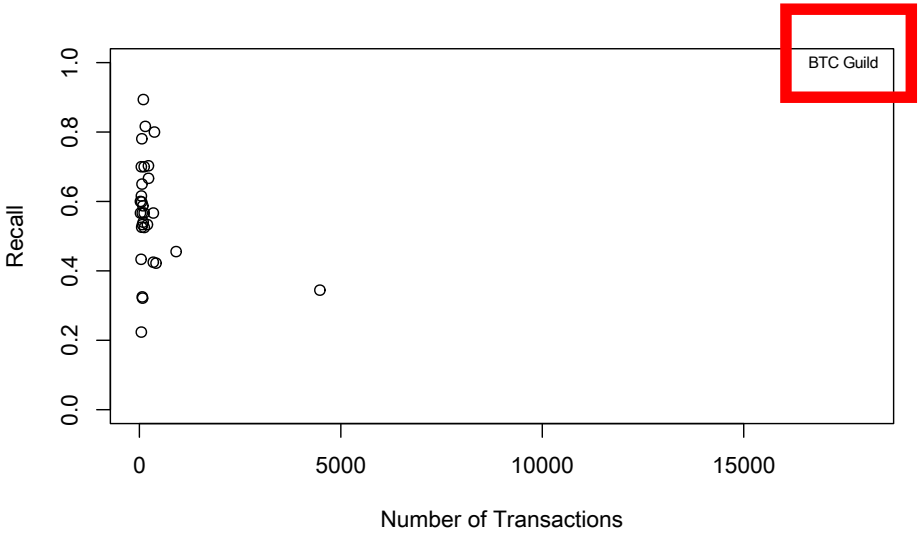
- アドレスにおける識別可能性の定義

$$F = \frac{2 \cdot (\text{再現率} \cdot \text{適合率})}{\text{再現率} + \text{適合率}}$$

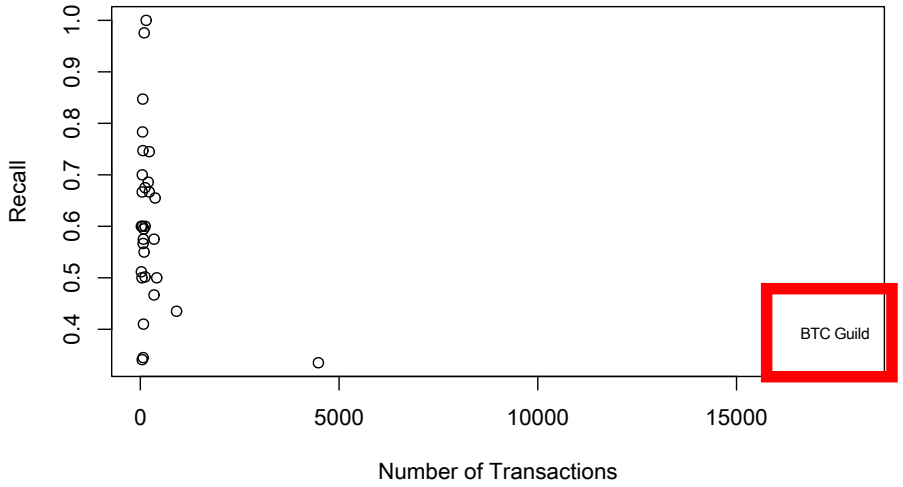
F値が0.5以上のアドレスを識別されたと定義する

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



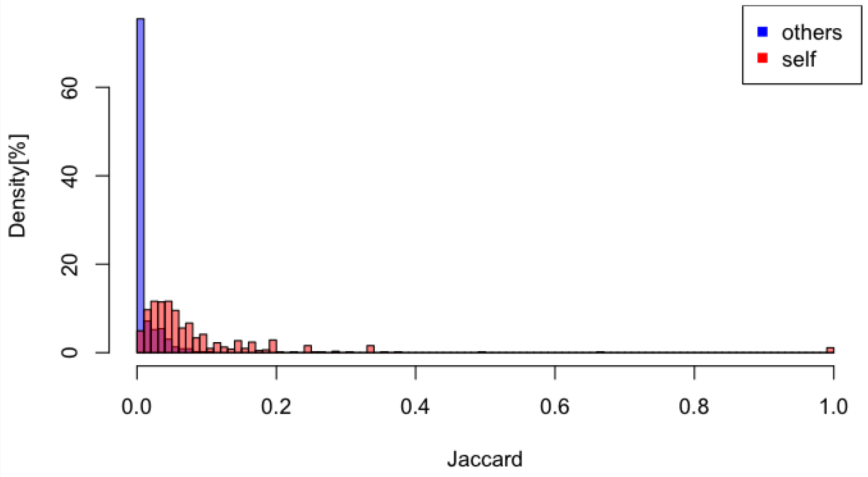
適合率の変化



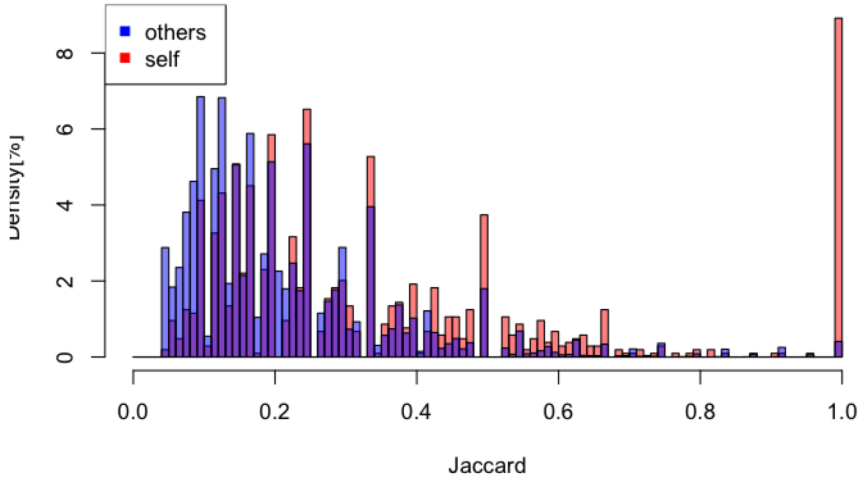
- 取引数と再現率・適合率に有意な相関はない
- BTC Guild
  - マイニングプールのアドレス
  - 多くのアドレスが，このアドレスに再識別されている

実験結果2:送金先集合と時間集合の比較

送金先集合jaccard

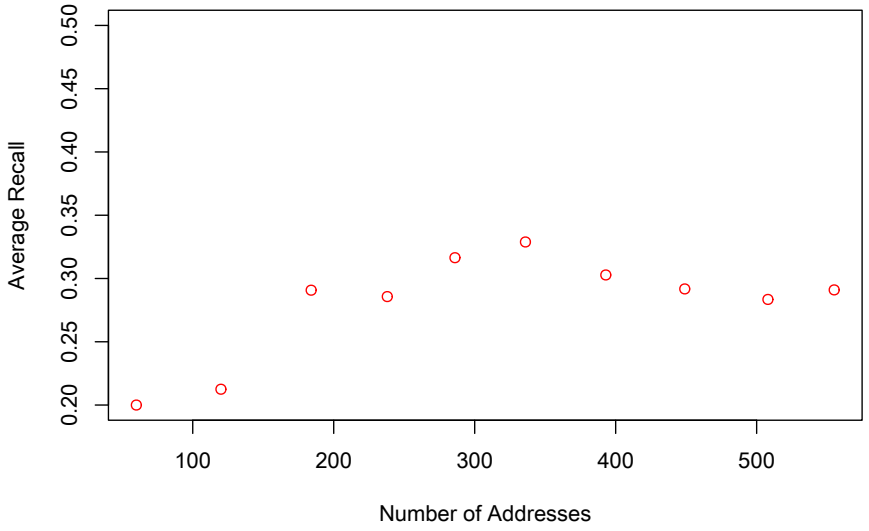


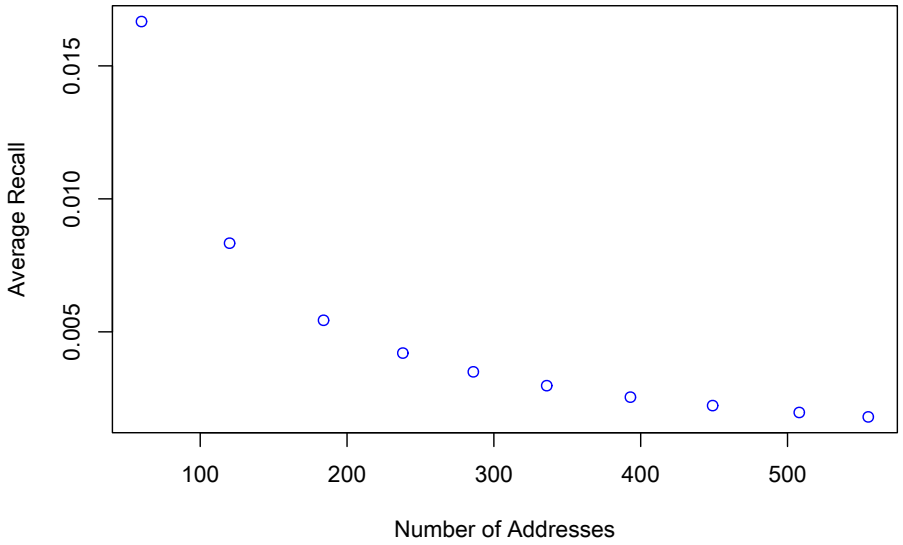
取引時刻集合jaccard



- 送金先集合
  - 異なるアドレス間で同一の送金先が少ない
  - 同一アドレスのjaccard係数(赤)は大きく分布している
- 取引時刻集合
  - 自他の差が小さい

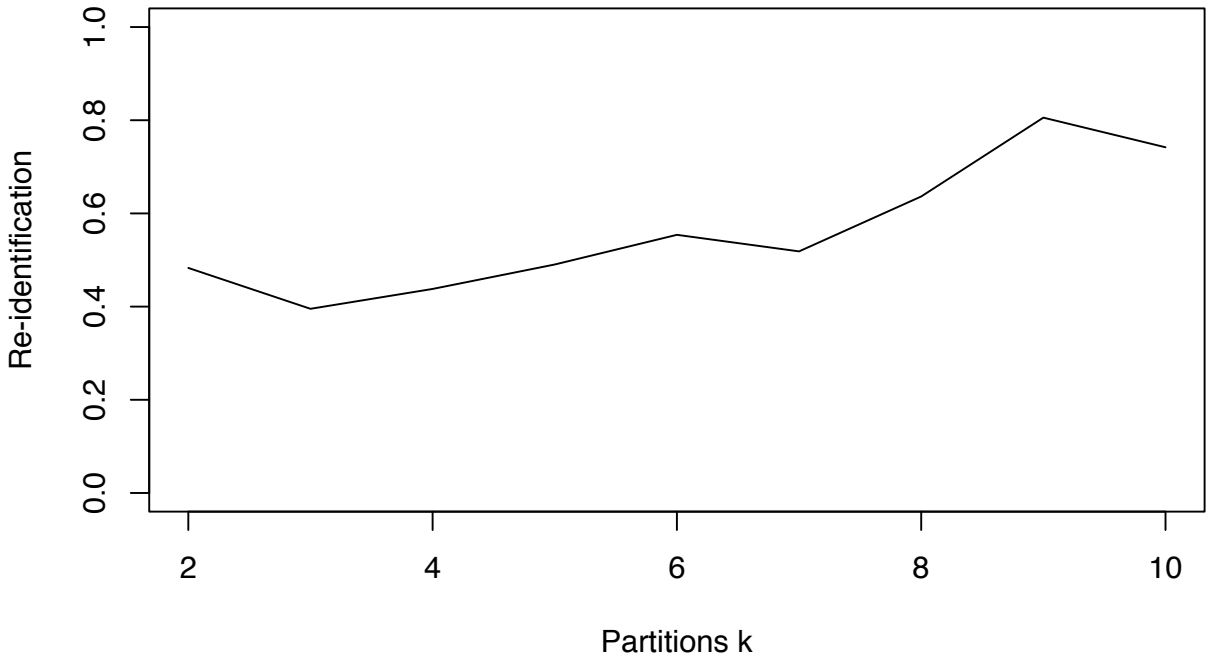
実験結果3:対象アドレス数による平均再現率





- 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



- 最大識別率は9分割時の80.5%（29/36)
- 最小識別率は3分割時の39.5%(117/296)
- 分割数が増えるにつれて増加傾向

考察

- 取引数と再現率・適合率に有意な相関はない
  - 同一のアドレスと取引を続けているアドレスが識別可能性が高い
- 分割数が増えるにつれて識別率が増加
  - 識別対象のアドレスが減少したため

k	n
2	559
3	296
4	153
5	104
6	74
7	54
8	44
9	36
10	31

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