

# Bitcoin アドレスの送金先集合に基づく匿名性の評価

永田 倅大<sup>1</sup> 菊池 浩明<sup>2</sup>

**概要**：近年、匿名性の高い暗号通貨 Bitcoin が注目されている。しかし、その匿名性は Bitcoin アドレスの仮名に基づく弱いものであり、取引履歴や頻度などの統計情報に基づく、アドレスの識別が可能であり、匿名性に影響を与えられ考えられる。そこで、我々は Bitcoin のブロックチェーンの取引に関するデータに基づき、Bitcoin アドレスの送金先集合に基づいたアドレス識別実験を行い、匿名性の評価をする。

**キーワード**：暗号通貨, Bitcoin

## Anonymity evaluation of Bitcoin addresses based on a set of output addresses

KODAI NAGATA<sup>1</sup> HIROAKI KIKUCHI<sup>2</sup>

### 1. はじめに

近年、暗号通貨の注目が増している。その中でも 2009 年から運用が開始された Bitcoin は銀行などの第三者機関を介さず取引できることや、国を超えた送金が容易にできる、匿名性が高いという特徴がある。しかし、匿名性が高いと言われているが 2013 年に Meiklejohn らによって特定の取引パターンから同一ユーザが管理しているアドレスを特定した研究 [1] や、2015 年に Dupont らによってアドレスを管理するユーザの居住地のタイムゾーンが特定された研究 [2] がある。さらに、アドレスの取引頻度や、送金先がどれほど匿名性に影響を与えるかということが明確ではない。

そこで本研究では、アドレスの取引頻度や送金先集合がどれほどアドレスの匿名性に影響を与え、識別されるリスクがあるかを明らかにすることを目的とする。

そのために Bitcoin のブロックチェーンから取引に関するデータを収集し取引データベースを作成し、送金先集合

に基づいたアドレス識別実験を行い、各アドレスに対して匿名性の評価をする。

その結果、最大で 80.5% のアドレスが識別されることが判明した。

### 2. Bitcoin

Bitcoin は Nakamoto 氏の論文 [3] を基に、2009 年より運用が開始された暗号通貨である。取引の検証や承認、新たなビットコインの発行は全てユーザによって行われる。ビットコインの取引に関する情報などはブロックに格納される。ブロックは約 10 分に 1 個生成され、各ブロックが 1 つ前のブロックと繋がっておりブロックチェーンを構成し分散管理されている。ブロックや取引に関する情報は Blockchain[4] で確認することが可能である。

#### 2.1 アドレス

ビットコインの送受金を行うために `1A1zP1eP5QGefi2DMPTfTL5SLmv7DivfNa` といったようなアドレスが使用される。この文字列はユーザが作成した公開鍵に対し、ハッシュ関数を適用した出力結果である。一般的にアドレスはユーザの個人情報を特定することができない仮名である。アドレスはウォレットによって管理され、複数のアドレス所有することも可能である。

<sup>1</sup> 明治大学大学院先端数理科学研究科  
Graduate School of Advanced Mathematical Sciences, The University of Meiji

<sup>2</sup> 明治大学総合数理学部  
School of Interdisciplinary Mathematical Sciences, The University of Meiji

表 1 取引構造

フィールド	説明
Version	取引が従うルールのバージョン
Input Counter	取引の入力数
Inputs	取引の入力データ
Output Coutner	取引の出力数
Outputs	取引の出力データ
LockTime	ブロック高または Unix タイムスタンプ

表 2 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

## 2.2 取引

表 1 に取引の構造を示す。ビットコイン取引は入力と出力の 2 つのフィールドに分けられている。入力には送金者のアドレスを指定し、出力には受取者のアドレスと送金額を指定する。どちらのフィールドも複数のアドレスを指定することが可能である。

表 2 の例では、 $a_2$  の取引数は  $\{Tx_2, Tx_4\}$  の 2 件、 $a_3$  の取引数は  $\{Tx_1\}$  の 1 件、 $a_5$  の取引数は  $\{Tx_4\}$  の 1 件となる。また  $Tx_4$  の入力の  $a_2$  のように、同一アドレスが複数指定されることもある。取引の多くは  $Tx_3$  のように出力アドレスの 1 つに送金者のアドレスを指定する。この理由は、取引で生じるお釣りを受け取るためである。

## 2.3 ブロック

ビットコインの取引はブロックに格納される。ブロックはユーザによって作成され、作成したユーザは報酬としてビットコインを受け取ることができる。報酬を受け取る取引はコインベースと呼ばれており、新しいビットコインを発行している。コインベースの取引は表 2 の  $Tx_1$  に該当する。入力フィールドは空白であり、出力には報酬を受け取るアドレスを指定する。

## 3. データ収集

本節では実験に使用する取引データとアドレスデータの収集方法について説明する。

### 3.1 取引データ

本実験では、*bitcoind* クライアントを用いて Bitcoin のブロックデータはダウンロードし、ブロックデータに対して *bitcoind* クライアントを用いてパースを行い、取引に関するデータを収集した。478,184 ブロックから 242,799,426 取引のデータを収集し、SQLite3 のデータベースに格納し

The screenshot shows a profile page for a user named 'macbook-air'. It lists various statistics and personal information. Key details include: 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. It also shows ICQ, AIM, MSN, and YIM status as hidden. Email is hidden, Website is F2Pool, and Current Status is Offline. The Bitcoin address is 1KFHE7w8BhaENAswrryaoccdB6qcT6DbYY. Gender is Male, Age is N/A, Location is China, Local Time is February 05, 2018, 02:20:59 PM, and Trust is 0: -0 / +0.

図 1 bitcointalk プロフィールページ

た。データベース内には Input Table, Output Table の 2 つのテーブルが含まれる。

### 3.1.1 Input Table

表 3 に Input Table の一部を示す。Input Table には 5 つの属性がある

### 3.1.2 Output Table の例

表 4 に Output Table の一部を示す。Output Table には 5 つの属性がある。

## 3.2 アドレスデータ

本実験で匿名性を評価するアドレスを 2 種類の方法で取得した。1 つ目はコインベースの出力で指定されたことのあるアドレスである。2 つ目は Bitcoin のオンラインフォーラムである BitcoinTalk[5] から収集した。BitcoinTalk にはユーザごとに図 1 で示されるようなプロフィールページが用意されており、そこで公開されている Bitcoin address の項目から取得した。ユーザがアドレスを公開しているのはフォーラムで回答したことへの寄付を受け付けるためなどの理由が考えられる。

BitcoinTalk から集めたアドレスデータの一部を表 5 に示す。本来 Bitcoin アドレスは仮名であり実ユーザとの対応はないが、表 5 の 3 行目のデータでは、アドレスの文字列とユーザ名が一致している。

## 4. 実験

本節では対象アドレスに対しての匿名性評価を行う。取引データにおいて、

- $N$  = 識別対象アドレス数
- $A = \{a_1, \dots, a_n\}$  : 識別対象アドレスの集合
- $K$  = 分割数
- $O_{ki}(a_i) = \{o_1, \dots, o_p\}$  : 分割数  $K$  の時の期間  $i$  におけるアドレス  $a_i$  の送金先アドレス集合
- $T_{ki}(a_i) = \{t_1, \dots, t_q\}$  : 分割数  $K$  の時の期間  $i$  におけるアドレス  $a_i$  の時間アドレス集合

表 3 Input Table の例

属性	説明	値例
Time	取引が格納されたブロックの発掘時刻	2012/09/22 10:47:23
Height	取引が含まれるブロック番号	200001
TxHash	取引 ID	d635410b5408592d54f59a010ae77974726b2a7ccd26bc76f9a68e02babe3ee5
PreTxHsh	入力に使われるビットコインを受け取った取引 ID	2d6dc2475b5ca40a081b857cc2b7e9fa29376bc299bed62c2d72244ec5a05a6a
InputAddr	送金者のアドレス	1EEYSdwDg9Rvu7bj3AjjJ662yyDbUG1fNi

表 4 Output Table の例

属性	説明	値例
Time	取引が格納されたブロックの発掘時刻	2012/09/22 10:47:23
Height	取引が含まれるブロック番号	200001
TxHash	取引 ID	d635410b5408592d54f59a010ae77974726b2a7ccd26bc76f9a68e02babe3ee5
OutputAddr	受取者のアドレス	1ArR7vf17C9ThWi5yt3c74TamCnPuaGb6e
Value	受け取ったビットコインの額 [ $10^{-8}$ BTC]	560000000

を定義する。期間  $i$  はデータセットの分割数によって変化する。表 7 ではデータセットを 3 分割しているため期間  $i$  は  $\{p_1, p_2, p_3\}$  の 3 つである。データセットの分割方法は、ブロック番号を基準に分割している。そのため分割数が増加するにつれて分割データの期間は短くなっていく。時間集合は取引が格納されたブロックの発掘時刻を集合とする。表 3 の場合、Time の 10:47:23 であるため、時間集合の要素に 10 が含まれる。

jaccard 係数は、

$$J(A, B) = \frac{|O_i(A) \cap O_i(B)|}{|O_i(A) \cup O_i(B)|}$$

で定まる集合  $A, B$  間の類似度である。

#### 4.1 実験概要

表 6 に本実験で使用する取引データとアドレスデータの概要を示す。

本実験では、データセットを分割し、送金先集合を学習データと評価データに分類し jaccard 再識別を用いることで、どれくらいのアドレスが識別されるのかを明らかにする。

#### 4.2 実験結果

##### 4.2.1 分割数による平均再現率・平均適合率の変化

図 2 は分割数による平均再現率・平均適合率の変化を表す。分割数が多くなるにつれて、両方の値が増加している。

##### 4.2.2 取引数による平均再現率・平均適合率の変化

図 3 は 10 分割時におけるアドレスの取引数による平均再現率の変化を表す。取引件数と平均再現率には相関がないと考えられる。

図 4 は 10 分割時におけるアドレスの取引数による平均適合率の変化を表す。こちらも取引件数と平均適合率には相関がないと考えられる。

##### 4.2.3 自他の Jaccard 係数の比較

同一アドレスと他者アドレスとの jaccard 係数の比較を

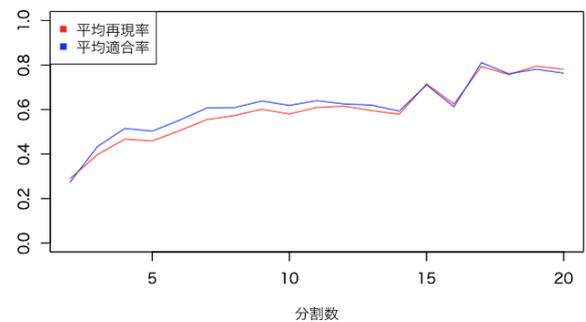


図 2 分割数による平均再現率・平均適合率の変化

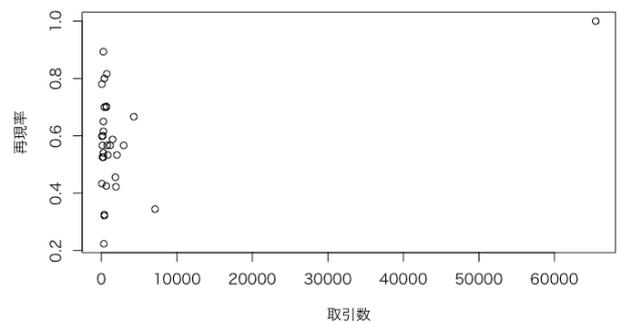


図 3 アドレスの取引数による再現率の変化

行う。

図 5 は送金先集合における 10 分割時の自他アドレス jaccard 係数のヒストグラムを示す。jaccard 係数が 0 のデータは除いている。他者アドレスの jaccard 係数の最大値が 0.012 であることから、異なるアドレス間では同一の送金先が少ない。図 6 は時間集合における 10 分割時の自他アドレスとの時間 jaccard 係数ヒストグラムである。こちらのヒストグラムも jaccard 係数が 0 のデータを除いている。

表 5 Bitcointalk から収集したデータ例

Addr	Name	Location
1KFHE7w8BhaENAswryaocDb6qcT6DbYY	macbook-air	China
1DNNERMT5MMusfYnCBfcKCBjBKZWBC5Lg2	BitHits	None
1Anduck6bsXBXH7fPHzePJSXdC9AEsRmt4	Anduck	None

表 6 データセット概要

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

表 7 3 分割時のデータセット例

	$p_1$ ( 200,001 - )	$p_2$ ( 233,334 - )	$p_3$ ( 266,667 - )
$a_1$	$a_1, a_2$	$a_3, a_4$	$N/A$
$a_2$	$a_2, a_5$	$a_4, a_5$	$a_5$
$a_3$	$a_3, a_6$	$a_4, a_6$	$a_5, a_7$

Algorithm 1 : jaccard 再識別

入力: 取引データ

- Step 1. データセットを任意の期間に分割
- Step 2. 各アドレスの期間ごとに  $O_{ki}$ , または  $T_{ki}$  を作成.  
1 つ以上の期間で集合の大きさが 0 のものは対象から外す
- Step 3. Step 2 で作成したデータを学習データ, 評価データに分け jaccard 係数を求め, 最大のものを正解と予測
- Step 4. 分割数に応じて Step3 を繰り返す

出力: 予測したアドレスを返す

例) 表 7

$a_1$  は  $p_3$  において集合の大きさが 0 なので実験対象から外す.  
まず,  $p_1$  を学習データ  $[\{a_2, a_5\}, \{a_3, a_6\}]$  として  $p_2, p_3$  の各アドレスに対して jaccard 係数を求める.  
 $a_2$  の  $p_2$  時の学習データとの jaccard 係数を求めると  $\{0.667, 0\}$  となるため, 最大値である  $a_2$  を正解アドレスと予測する.

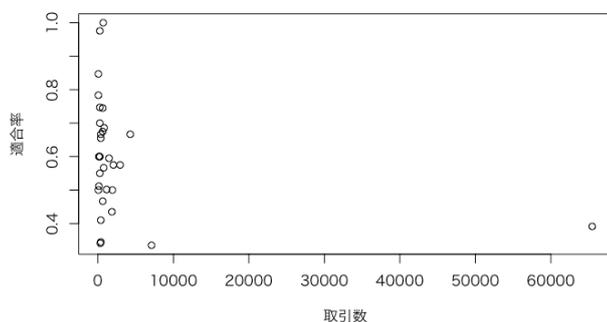


図 4 取引数による適合率の変化

表 8 は図 5, 図 6 の概要を表す. どちらのヒストグラムも, 同一アドレスが最大値と平均値が他者アドレスを上回っている.

#### 4.2.4 アドレス数による平均再現率の変化

本節では対象アドレス数が平均再現率に与える影響を説

表 8 自他アドレスとの jaccard 係数の概要

		最小値	最大値	平均値	0 の割合 [%]
送金先集合	同一アドレス	0	1.0	0.038	55
	他者アドレス	0	0.110	0.0001	99
時間集合	同一アドレス	0	1.0	0.264	25
	他者アドレス	0	1.0	0.155	29

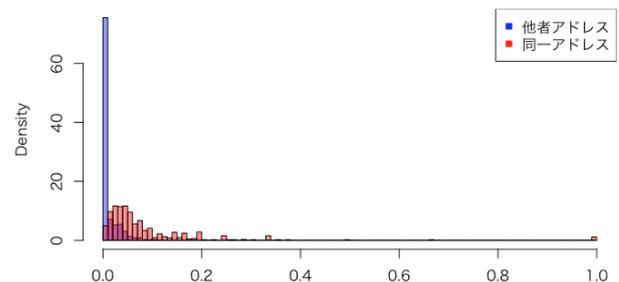


図 5 送金先集合における自他アドレスの jaccard 係数ヒストグラム

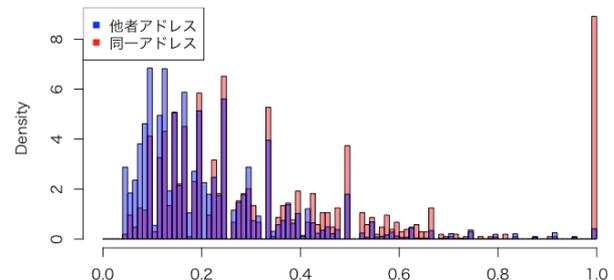


図 6 時間集合における自他アドレスの jaccard 係数ヒストグラム

明する. 図 7 は 2 分割時の送金先集合と時間集合における, 対象アドレス数による平均再現率である. 送金先集合は時間集合に比べて平均再現率が高い. 対象のアドレス数が増加しても, 平均再現率に大きな変化は見られなかった.

## 5. 評価

### 5.1 匿名性の定義

本実験では匿名性を以下のように定義する.

$$F = \frac{2 * (\text{平均再現率} * \text{平均適合率})}{\text{平均再現率} + \text{平均適合率}}$$

F 値が高いほど匿名性が低く, 0.5 以上のアドレスを識別されたと定める.

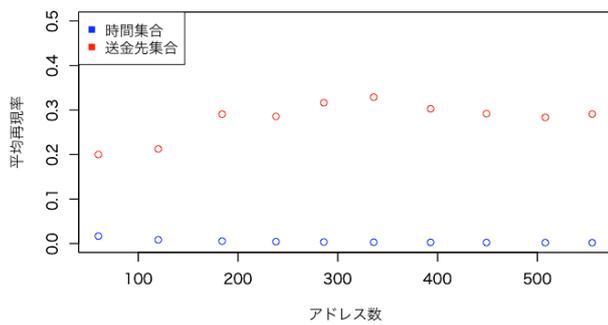


図 7 送金先集合と時間集合の平均再現率の比較

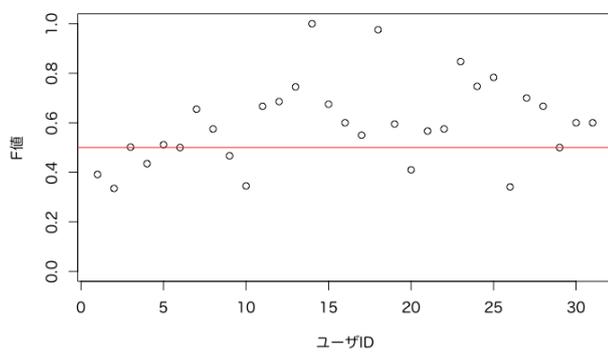


図 8 10 分割時の F 値の分布

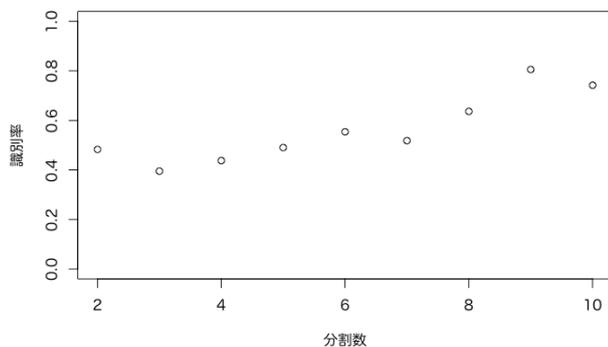


図 9 分割数による識別率

図 8 は 10 分割におけるアドレスの F 値の分布を示す。74.1%(23/31) のアドレスが識別された。

## 5.2 識別率

図 9 は分割数による識別率を示す。最大識別率は 9 分割時の 80.5%，最小識別率は 3 分割時の 39.5%であった。分割数が増加するにつれて識別率も増加する傾向にある。

## 6. 考察

本実験では、図 2、図 9 で示されたように分割数によ

表 9 分割数による対象アドレス数

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

て平均再現率・平均適合率、識別率が増加しているが、これは表 9 で示されるように分割数が増えるに依り実験に用いるアドレスが減少し、対象となるアドレスがデータセットの期間中に定期的に取り行なっているため jaccard 係数において他アドレスとの違いが明確であったからであると考えられる。

また本実験では、アドレスの取引数の多さが平均適合率・平均再現率には影響を与えないという結果が得られた。このことから送金先アドレスは安定していないと考えられる。jaccard 再識別においては同じ送金先アドレスと取引を続けているアドレスほど識別されるリスクが高いと。そのためアドレスの匿名性を保つためには、同じアドレスとの取引を続けないこと、もし取引を行う場合は送金者がアドレスを変更することが必要であると考えられる。さらに、図 5 において送金先集合では自他アドレスの jaccard 係数が重なっている部分が少ないことが平均再現率・平均適合率、識別率が高くなった要因であると考えられる。

## 7. おわりに

本実験では Bitcoin アドレスの送金先集合に基づく匿名性の評価を行なった。その結果、送金先集合を用いて jaccard 再識別を行うことでアドレスが高い確率で識別されるリスクがあることが判明した。さらに、取引数が平均再現率・平均適合率に影響を与えないことが判明した。しかし、対象となるアドレスが少ないため、より多くのアドレスに対し実験を行うことが必要である。

データセットのアドレス数を増やすこと、最近の取引に対しても実験を行うことを今後の課題とする。

## 参考文献

- [1] S. Meiklejohn, M. Pomarole, G. Jordan, K. Levchenko, D. McCoy, G. M. Voeker, S. Savage. A fistful of bitcoins: Characterizing Payments Among Men with No Names. *In Proceedings of Conference on Internet Measurement Conference (IMC'13)*. ACM, 2013.
- [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

- [3] S. Nakamoto, Bitcoin: A Peer-to-Peer Electronic Cash System. <https://bitcoin.org/bitcoin.pdf>, 2008.
- [4] Blockchain. <https://blockchain.info>
- [5] Bitcointalk. <https://bitcointalk.org/>