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]

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と，アドレスの識別が可能であり，匿名性に影響を与えると考えられる．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[1]　Hiroaki Kikuchi[2]

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に取引できることや，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，その匿名性はビットコインアドレス (以下アドレスと呼ぶ) がランダムな仮名であることに基づいており，複数の取引から同一ユーザによるか否かは容易に識別可能である．この匿名性についてはいくつかの先行研究がある．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複数のアドレスが識別できることが示されている [1]．2015 年に Dupont らは取引の時刻に注目し，その時刻分布に基づい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できることを示した [2]．しかしながら異なるユーザでも同様の時間帯に活動することはあり，平均識別精度は 72%にとどまっていた．

そこで本研究では，時間情報よりも，識別に有効な情報としてアドレスの取引頻度やと送金先集合に注目する．なぜならば，ユーザごとに取引する相手は決まっており，アドレス空間は大きいので，そこからユーザを追跡することは十分可能と考えるからである．本研究は，アドレスの取引頻度と送金先の情報が，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 2012 年からの 1.5 年間分のブロックチェーンから取引に関するデータを収集し，10 万ブロックの取引データベースを作成する．送金先集合として既知の 559 アドレスを用い，基づいたアドレス識別実験を行い，アドレスについて匿名性を評価する．その精度を [2] と比較する．

本実験に基づき，最大で 80.5%のアドレスが識別できることを示す．

2. Bitcoin

2.1 概要

Bitcoin は Nakamoto 氏の論文 [3] を基に，特定の中央管理者を持たず，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] 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
[2] 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

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．

2.2 アドレス

ビットコインの送受金は 1A1zP1eP5QGefi2DMPTfTL5SLmv7DivfNa といったようなアドレスについて行われる．アドレスはユーザが作成した公開鍵にハッシュ関数 SHA256，RIPEMD160 を適用し，チェックサムを追加した後に base58 で符号化している．従ってアドレスは特定の個人を特定することが不可能な仮名である．アドレスはウォレットによって管理され，1 ユーザが複数のアドレス所有することも可能である．

2.3 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 つのフィールドから成る．入力には送金者のアドレスを，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 に 5 つの取引 Tx_1,...,Tx_5 を含むブロックの例を示す．表 2 の例では，入力アドレス a_2 の取引は Tx_2,Tx_4 の 2 件，a_3 の取引は Tx_1 の 2 件，a_5 の取引は Tx_4 の 1 件である．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 a_3 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．

ブロックをマイニングに成功したユーザは報酬として一定額の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，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 a_2 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を説明する．

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

3.1 取引データ

本研究では，*bitcoind* クライアントを用いて Bitcoin の全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．表 6 に本研究で使用するデータ概要を示す表 3 に Input Table の一部を示す．Input Table には 5 つの属性がある．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

アドレスは仮名なので，本来，誰が管理しているか分からない．しかし，我々は，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．その多くはマイニングプール業者であり，その位置や国などの情報は公開されていることが多い．2 つ目は Bitcoin のオンラインフォーラムである Bitcointalk[5] にて公開されているアドレスである．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 行目のアドレスの様に，アドレスの文字列とユーザ名が一致しているものも混じっている．

4. 再識別実験

4.1 概要

本節では 3 章の対象アドレスに対しての匿名性評価を行う．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．

取引データにおいて，

- $A = \{a_1, \ldots, a_n\}$: 識別対象アドレスの集合

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

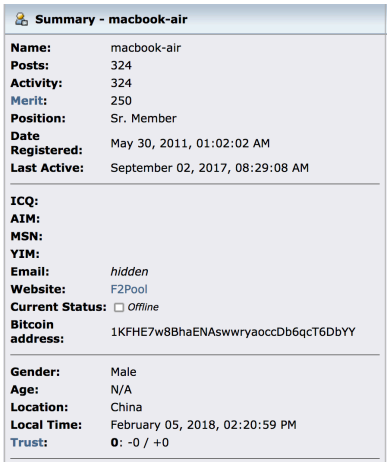

図 1: bitcointalk プロフィールページ

- $O_i(a_j) = \{o_1, \ldots, o_{n_{i,j}}\}$: 期間 i における入力アドレス a_j の送金先アドレス集合
- $T_i(a_j) = \{t_1, \ldots, t_{n_{i,j}}\}$: 期間 i における入力アドレス a_i の取引時刻集合

を定義する．

本実験ではデータセットを k 個に分割して変化を観測する．表 7a にデータを 2 分割した時の O_i の例を．表 7b に 3 分割の例を示す．データセットを 3 分割しているため期間 i は 3 つである．データセットの分割は，ブロック番号を基準に等分割している．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．

表 7b の a_1 は期間 3 では取引がないことに注意せよ．本実験では学習データに取引がないアドレスは識別の対象から外す．

出力アドレス集合の類似度に基づいて，アドレスを識別する．集合 A, B の類似度は，

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

で定まる jaccard 係数を用いる．提案方式を Alogorithm1 に示す．

先行研究 [2] では時間情報に基づいて識別している．ここでは Algorithm 1 に合わせて，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．

4.2 実験結果

4.2.1 分割数 k による平均再現率・平均適合率の変化

図 2 に分割数 k による平均再現率・平均適合率の変化を表す．

ここで平均再現率 R は

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

平均適合率 P はアルゴリズムが予測したアドレスの集合 $A' = \{a_1, \ldots, a_{n'}\}$ について

$$R = \frac{1}{n'} \cdot \sum_{a \in A'} k = \frac{\text{正しい a のデータ数}}{\text{a のデータ数}}$$

分割数が多くなるにつれて，両方の値が増加している．これは，分割が多くなると取引が 0 となるアドレス数が増えて，識別対象が減るためである．表 9 に分割数 k に対する対象入力アドレス数を示す．

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 $k = 10$ 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有意な相関はない．

図 4 は $k = 10$ 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．ここで，15,000 を超える取引をしている BTC Guild に注意せよ．主にコインベースとして多くの取引に関わるために再現率は 1.0 だが，他の多くの

表 5: Bitcointalk から収集したデータ例

Addr	Name	Location
1KFHE7w8BhaENAswwryaoccDb6qcT6DbYY	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: データセット例

(a) k = 2

期間 i	1	2
	10 ヶ月	10 ヶ月
a_1	{ a_1,a_2,a_3 }	{ a_2,a_3,a_4 }
a_2	{ a_2,a_5 }	{ a_4,a_5 }
a_3	{ a_3,a_4,a_6 }	{ a_4,a_5,a_6,a_7 }
	学習データ	評価データ

(b) k = 3

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	

Algorithm 1 : jaccard 再識別

入力: アドレス集合 A，送金先 $o_1, \ldots, o_k$

Step 1. データセットを k の期間に分割した o_i,t_i 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 2. データ $O_1(a_j)$ を学習データ，$O_2(a_j)$,...,$O_k(a_j)$ を評価データに分け，$a_l \in A$ について
jaccard 係数 $J(O_i(a_l)$,$O_1(a_j))$ を最大とする l を
アドレス a_j の識別先とする

出力: 予測したアドレスを返す

アドレスがこのアドレスに再識別されている．

図 5 は $k = 10$ におけるアドレスの取引数による平均再現率と平均適合率の関係を表す．両者には正の相関関係が見られた．

4.2.3 自他の Jaccard 係数の比較

送金先集合に基づく再識別の精度を確かめるため，同一アドレスと他者アドレスとの jaccard 係数を調べる．

図 6 に送金先集合における $k = 10$ の自他アドレス jaccard 係数の分布を示す．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 であり，異なるアドレス間では同一の送金先が少ない．一方，同一アドレスの jaccard 係数 (赤) はより大きく分布してい

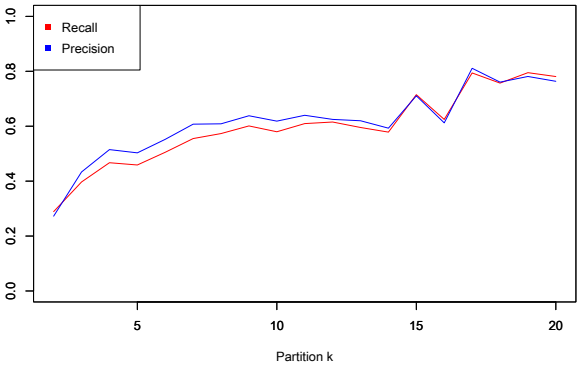

図 2: 分割数による平均再現率・平均適合率の変化

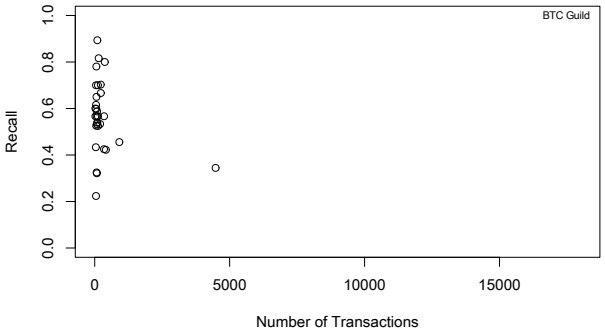

図 3: 取引数による再現率の変化

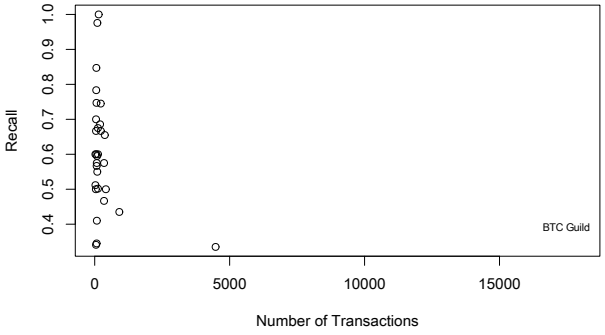

図 4: 取引数による適合率の変化

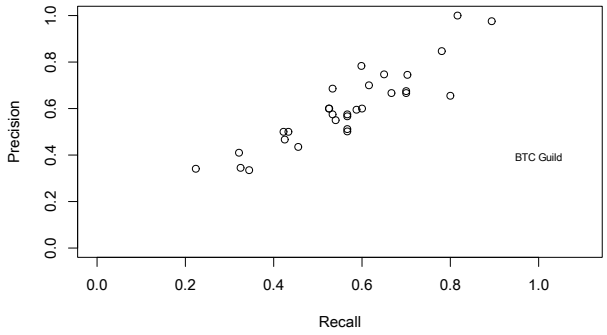

図 5: 取引数による再現率・適合率の関係

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

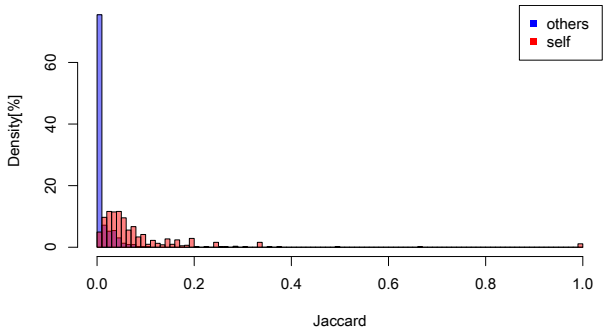

図 6: 送金先集合における自他アドレスの jaccard 係数ヒストグラム

る．この差により再識別が行われる．

図 7 は時間集合における $k = 10$ 分割時の自他アドレスとの時間 jaccard 係数の分布である．ここでも jaccard 係数が 0 のデータを除いている図 addr hist と比べて，自と他の差が小さいことが示されている．.

表 8 に jaccard 係数の統計値を整理する．どちらの分布も，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考える．図 8 に $k = 2$ の送金先集合と時間集合の各々の平均再現率を示す．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．一方，時間集合はアドレス数

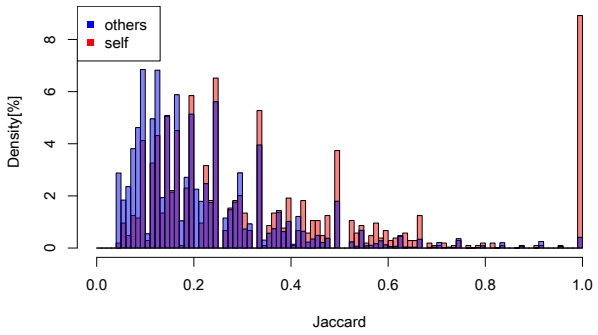

図 7: 時間集合における自他アドレスの jaccard 係数ヒストグラム

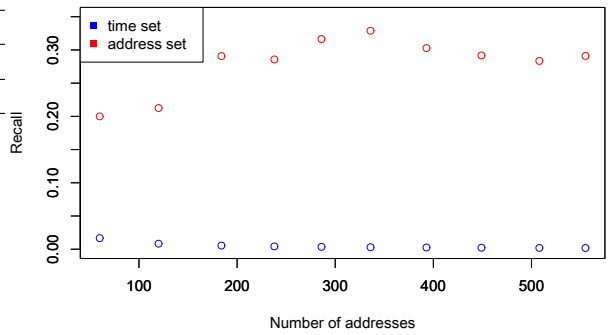

図 8: 送金先集合と時間集合の平均再現率の比較

n に対して，$\frac{1}{n}$ の割合で再現率を下げる．

5. 評価

5.1 匿名性の定義

本実験では匿名性を以下 F 値で評価する．

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

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．

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，平均再現率・平均適合率はアドレスの取引数の多さに依存しなかった．このことから送金先アドレスは安定せず，多くの場合は異なる相手と取引をしていると考えられる．. jaccard 再識別においては同じ送金先アドレ

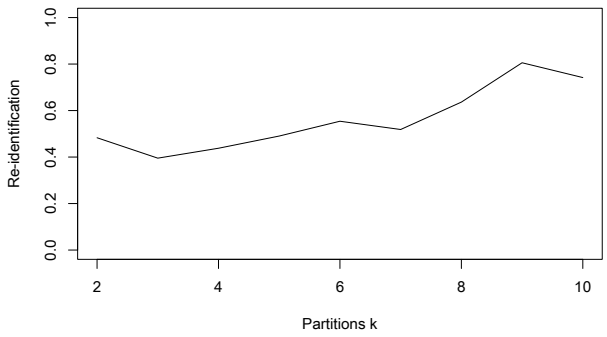

図 9: 分割数による識別率

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

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．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最大 80.5%のアドレスが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を示した．

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/