

明治大学大学院 先端数理科学研究科

2018年度

修士学位請求論文

Bitcoin アドレスの送金先集合に基づく匿名性の評価

学位請求者 先端メディアサイエンス専攻  
永田 倅大

# 目次

<b>第 1 章</b>	<b>序論</b>	<b>1</b>
1.1	本研究の背景	1
1.2	本研究の目的	1
1.3	本稿の構成	2
<b>第 2 章</b>	<b>基礎定義と従来研究</b>	<b>3</b>
2.1	基礎定義	3
2.1.1	Bitcoin とは	3
2.1.2	アドレス	3
2.1.3	取引	3
2.1.4	ブロック	4
2.2	従来研究	5
2.2.1	Sarah	5
2.2.2	Dupont	6
<b>第 3 章</b>	<b>送金先集合に基づく匿名性評価</b>	<b>9</b>
3.1	導入	9
3.2	データ収集	9
3.2.1	取引データ	9
3.2.2	アドレスデータ	9
3.3	再識別実験	10
3.3.1	概要	10
3.3.2	実験結果	12
3.4	評価	15
3.4.1	匿名性の定義	15
3.4.2	識別率	15
3.5	考察	15
3.6	まとめ	16
<b>第 4 章</b>	<b>平均取引時間分布の相関を用いた Bitcoin ユーザのタイムゾーン属性の推定</b>	<b>19</b>
4.1	はじめに	19
4.1.1	方法	19
4.1.2	結果	20

4.2 第3章手法との比較・考察 . . . . .	21
<b>第5章 まとめ</b>	<b>22</b>
謝辞	26
<b>付録A Bitcoin ユーザアンケート調査</b>	<b>27</b>
A.1 概要 . . . . .	27
A.2 アンケート内容 . . . . .	27
A.3 結果 . . . . .	30

# 第1章 序論

## 1.1 本研究の背景

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[1]は、匿名性が高く、国を超えた送金が容易にできるという特徴がある。さらに、全ての取引はブロックチェーンに記録されており、誰でも確認が可能である。Bitcoinのウォレットを保持しているユーザ数は、平成26年では約150万であったが平成30年には約2400万まで増加している[2]。その一方で、仮想通貨交換業者の仮想通貨が流出する事件が発生している。例えば、2014年にはMTGOX(マウントゴックス)社は合計約85万ビットコイン、金額にして470億円程度を消失していたことが判明した[3]。2018年にはコインチェック社から約580億円の仮想通貨NEMが流出した[4]。

しかし、匿名性が高いと言われているが、こ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構成されてい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。さらに、全ての取引はブロックチェーンに記録されているため確認することが可能である。そのため、ユーザがビットコイン取引を行うことで個人を特定される可能性もある。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らは、特定の取引パターンから同一ユーザが管理するアドレスをクラスタリングできることを示した[5]。

2015年にDupontらはアドレスの取引時刻の分布に注目し、アドレスを管理しているユーザのタイムゾーンを識別できることを示した[6]。しかし、異なるユーザで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。さらに、1ユーザに対し平均11個のタイムゾーンを推測していた。

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と送金先集合に注目する。なぜならば、ユーザごとに取引を行う相手は決まっており、アドレス空間は十分に大きいため、そこからユーザを追跡することが可能であると考えられるからである。

## 1.2 本研究の目的

本研究の目的は、アドレスの取引頻度と送金先集合が、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、アドレスの取引時刻から、どれほどタイムゾーンを識別されるリスクがあるかを明らかにすることである。

### 1.3 本稿の構成

本稿は5章で構成される。

- 第1章：本稿の研究背景，目的を述べた。
- 第2章：本稿の基礎定義と，関連研究について述べる。
- 第3章：アドレスの取引頻度と送金先集合を用いた，アドレス識別手法を提案し，精度について評価実験を行う。
- 第4章：アドレスの取引時刻に注目し，相関係数を用いたタイムゾーン識別手法を提案し，精度について評価実験を行う。
- 第5章：本稿のまとめを行う
- 付録A：Bitcoin ユーザのアンケート調査を行う

## 第2章 基礎定義と従来研究

### 2.1 基礎定義

#### 2.1.1 Bitcoin とは

Bitcoin は Nakamoto 氏の論文 [1] 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。国を超えて容易に送金できること、取引手数料が安い、匿名性が高いという利点を持つ。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[7] で確認することが可能である。Bitcoin には 2100 万 BTC が採掘上限として決められている。この上限には 2140 年ごろに到達する見込みである。ユーザはビットコインを管理するためにウォレットと呼ばれるものを使用する。ウォレットには Multibit や Bitcoin Core や、ペーパーウォレットというものがある。

#### 2.1.2 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。

#### 2.1.3 取引

取引は各ノードが P2P 方式で行い、取引完了には時間を要する。取引を行うためにはアドレスが必要になる。表 2.1 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。図 2.4 のようにユーザ X がユーザ Y にビットコインを送金する際は、アドレス A を入力として送金先としてアドレス D を指定することで Bitcoin の送金を行うことができる。取引はブロックの中に格納される。ブロックは約 10 分に 1 個生成されており、ブロックが生成されることで取引は承認を得る。ブロックを生成するためには膨大な計算量が必要となる。



図 2.1: Multibit の画面

表 2.2 に 5 つの取引  $Tx_1, \dots, Tx_5$  を含むブロックの例を示す. 表 2.2 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  $a_2$  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 2.2 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.

#### 2.1.4 ブロック

ブロックはブロックチェーンを形成する. ブロックチェーンは, ビットコインの二重取引や改ざんを防止する役割がある. ブロック生成を行うのはユーザであり, 最初にブロック生成したユーザには報酬としてビットコインを手に入れることができる. ブロックを生成する作業は Proof of work と呼ばれている.

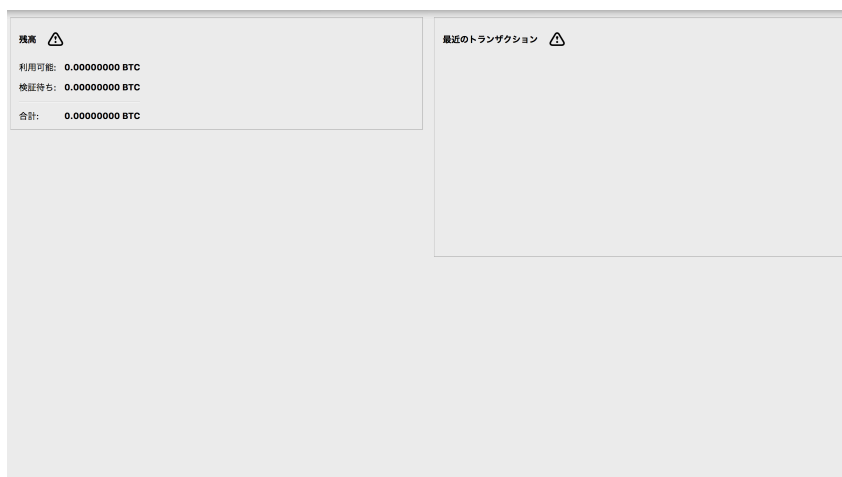


図 2.2: Bitcoin Core の画面



図 2.3: ペーパーウォレット

## 2.2 従来研究

### 2.2.1 Sarah

Meiklejohn らの先行研究 [5] は、ビットコインの取引の際に使用される、アドレスをクラスタリングし再識別を行うことで管理者を明らかにすること、Bitcoin 市場の長期的な変化の分析、その変化による Bitcoin システムへの影響、犯罪や詐欺目的で使用されたビットコインの検知を目的としている。アドレスの管理者を明らかにするために、アドレスにタグ付けを行う手法を用いている。

図 2.6 はアドレスのクラスタリングの手法の一つとして紹介されている。ビットコインの取引では入力アドレスの管理者は同一であるため、図 2.6 のように入力アドレスが 2 つ以上あるような取引 1, 取引 2 があった場合、アドレス A,B,C の管理者は同一であると定めている

アドレス間の変化を見るために、Change Address というクラスタ手法 22.7 も用いている。Change Address とは複数のアドレスで管理しているビットコインを 1 つのアドレス送金する方法である。研究結果として、3,384,179 個のクラスターができ、その中で管理者が分かったものは 2197 個であり、1,800,000 アドレスの管理者を特定したとっている



表 2.1: 取引構造

フィールド	説明
Version	取引が従うルールバージョン
Input Counter	取引の入力数
Inputs	取引の入力データ
Output Counter	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 2.2: 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

### 2.2.2 Dupont

Dupont らの先行研究は、アドレスを管理するユーザのタイムゾーンを特定することを目的としている。ユーザの居住地が判明している 518 アドレスに対して評価を行なっている。タイムゾーンを特定するために、取引時刻の分布に注目している。さらに、人間は AM4 には金融に関する取引を行わないなど、時間帯によって異なる行動をするのではないかという前提に基づいて実験を行なっている。

図 2.8 はあるアドレスが送金者として取引を行なった時刻のヒストグラムである。しかし、このヒストグラムには不自然な点がある。それは、AM7 ~ AM10 では取引が一度も行われていないにもかかわらず、AM4 に 15 回も取引が行われているということである。この理由として、ビットコインのブロックチェーンに記録されるタイムスタンプは全て UTC 0 であるため、実際のユーザのタイムゾーンと差が生まれてしまっている。そこで、

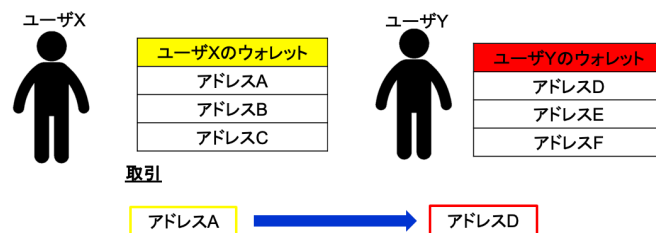


図 2.4: 取引方法

## ブロック #447435


概要		ハッシュ	
取引件数	2210	ハッシュ	000000000000002a5c7b4021058813349f3003cdfbd1398e82df69e32cc6
合計出力	6,399.66996745 BTC	前のブロック	0000000000000002f3b69492d3073fa45a77edb1a034fc6588efabc6376a6
推定取引量	1,715.71072066 BTC	次のブロック (複数可)	00000000000000001c3976488a39f0b85d3bed4dfd7fee615cbf72868d52e12
取引手数料	0.62203919 BTC	マークルルート	9fbadeb8d797182d780ed7cd81624916d942a9bf8e97b205e42bdc0e755b926
ブロック高	447435 (主観)	伝搬ネットワーク	
タイムスタンプ	2017-01-10 05:58:45		
受け取り時刻	2017-01-10 05:58:45		
中継所	F2Pool		
難易度	317,888,400,354.03		
ビット	402879999		
サイズ	999.891 KB		
バージョン	0x20000000		
ノンス	97575038		
ブロック報酬	12.5 BTC		

図 2.5: ブロック情報

正しいタイムゾーンを以下の提案手法で推定している。

1. ヒストグラムを1時間単位で分割し、最も取引が少ない時刻を見つける。
2. 1で見つけた時刻を、ユーザのローカルタイムで AM5 と仮定する。
3. 実際のタイムゾーンの候補である UTC offset のリストを作る。

例えば、図 2.8 では最も取引が少ない時刻は { 7 A.M., 8 A.M., 9 A.M., 10 A.M. }。そして、これらの時刻が AM5 となる UTC は { UTC -2, UTC -3, UTC -4, UTC -5 } である。

提案手法の精度を評価するために、518 アドレスに対して実験行なっている。その結果 72% のアドレスのタイムゾーンが正しく特定することができた。



図 2.6: クラスタリング手法 1

## 取引3(ブロック #800)

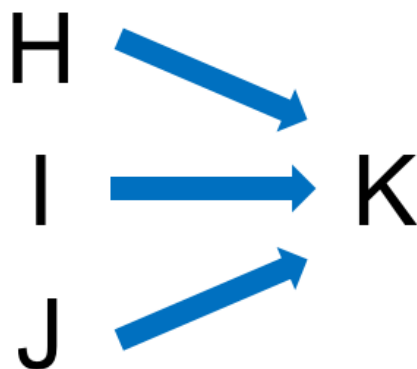


図 2.7: クラスタリング手法 2

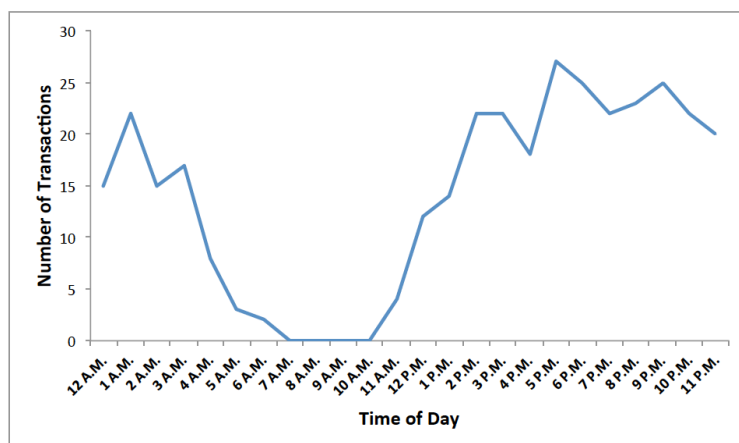


図 2.8: [6], p.2 の Figure 1 より転載

## 第3章 送金先集合に基づく匿名性評価

### 3.1 導入

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[6]。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[6]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。

### 3.2 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。

#### 3.2.1 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3.4に本研究で使用するデータ概要を示す表3.1にInput Tableの一部を示す。Input Tableには5つの属性がある。表3.2にOutput Tableの一部を示す。Output Tableには5つの属性がある。

#### 3.2.2 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

表 3.1: 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 3.2: 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	1ArR7vf17C9ThWi5yt3c74TamCnPUGb6e
Value	受け取ったビットコインの額 [ $10^{-8}$ BTC]	560000000

が多い。2つ目は Bitcoin のオンラインフォーラムである Bitcointalk[8] 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 3.1 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 3.3 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 3.3 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。

### 3.3 再識別実験

#### 3.3.1 概要

本節では 3.2 節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、

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

表 3.3: Bitcointalk から収集したデータ例

Addr	Name	Location
1KFHE7w8BhaENAswwryaoceDb6qcT6DbYY	macbook-air	China
1DNNERMT5MMusfYnCBfcKCBjBKZWBC5Lg2	BitHits	None
1Anduck6bsXBXH7fPHzePJSXdC9AEsRmt4	Anduck	None

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

図 3.1: Bitcointalk プロフィールページ

- $O_i(a_j) = \{o_1, \dots, o_{n_{i,j}}\}$ : 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
- $T_i(a_j) = \{t_1, \dots, t_{n_{i,j}}\}$ : 期間  $i$  における入力アドレス  $a_j$  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 3.5 にデータを 2 分割した時の  $O_i$  の例を. 表 3.6 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 3.6 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 Algorithm1 に示す.

先行研究 [6] 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 3.1 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.

表 3.4: 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 3.5: データセット例 (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

### 3.3.2 実験結果

アドレスの識別精度を評価するために、平均再現率、平均適合率を使用する。ここで平均再現率  $R$ 、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  $A' = \{a_1, \dots, a_{n'}\}$  について

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

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

$R_i, P_i$  はアドレス  $a_i$  の再現率、適合率である。

#### 分割数による平均再現率・平均適合率の変化

図 3.2 に分割数  $k$  による平均再現率・平均適合率の変化を表す。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 3.8 に分割数  $k$  に対する対象入力アドレス数を示す。

表 3.6: データセット例 (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  $o_1, \dots, o_k$

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,  $O_2(a_j), \dots, O_k(a_j)$  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

---

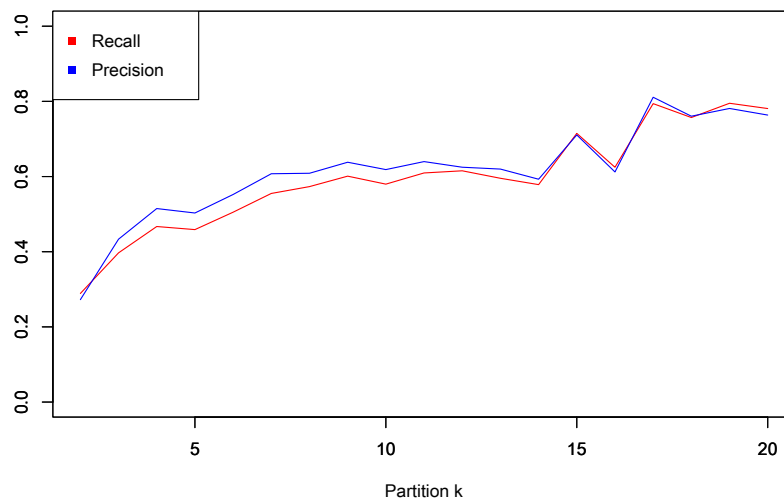


図 3.2: 分割数による平均再現率・平均適合率の変化

### 取引数による平均再現率・平均適合率の変化

図 3.3 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 3.4 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、他の多くのアドレスがこのアドレスに再識別されている。

図 3.5 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。



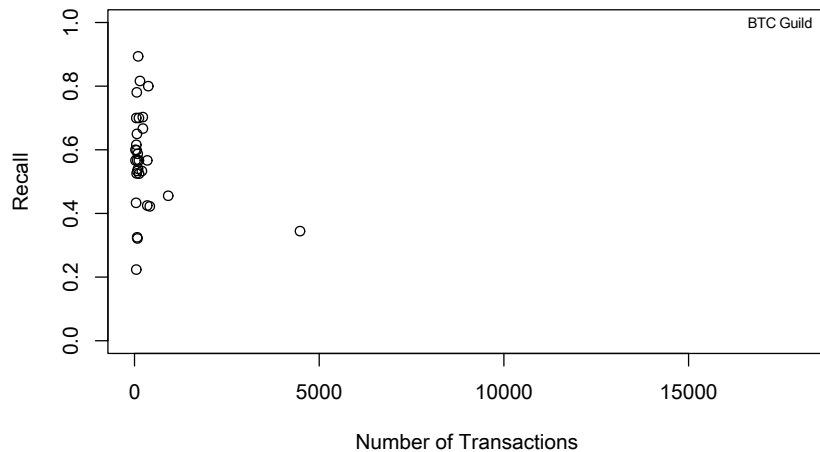


図 3.3: 取引数による再現率の変化

表 3.7: 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

### 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 3.6 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る。この差により再識別が行われる。

図 3.7 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??と比べて、自と他の差が小さいことが示されている..

表 3.7 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。

### 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 3.8 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

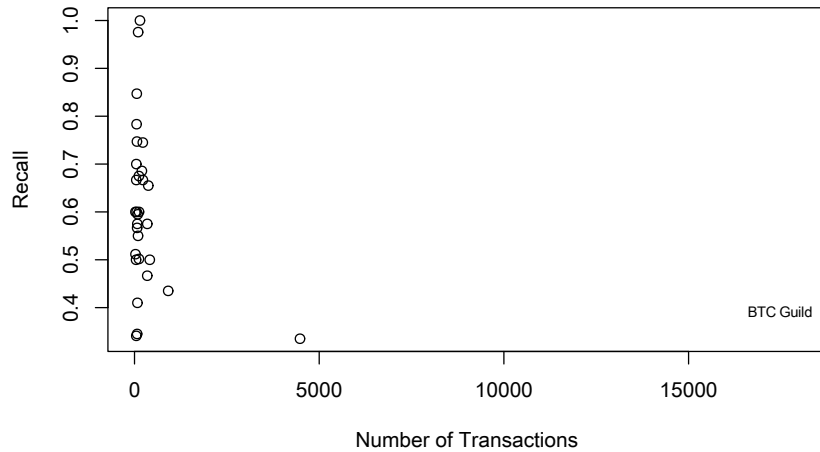


図 3.4: 取引数による適合率の変化

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  $n$  に対して、 $\frac{1}{n}$  の割合で再現率を下げる。

### 3.4 評価

#### 3.4.1 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。

#### 3.4.2 識別率

図 3.9 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。

### 3.5 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られる。

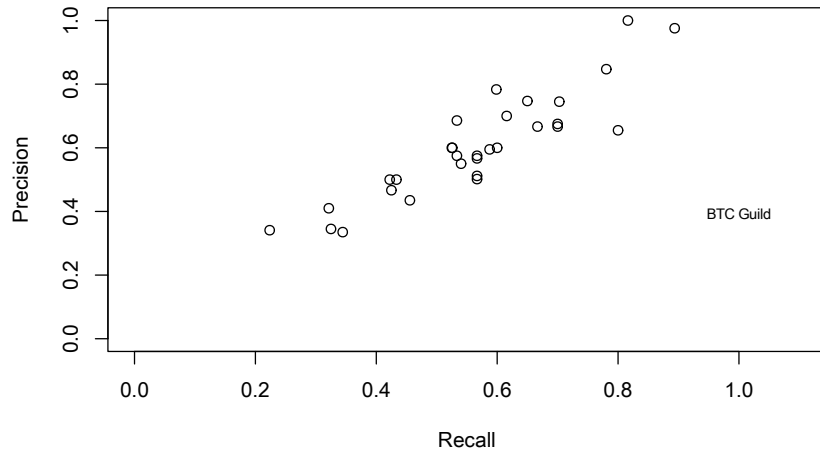


図 3.5: 取引数による再現率・適合率の関係

表 3.8: 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

### 3.6 まとめ

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。

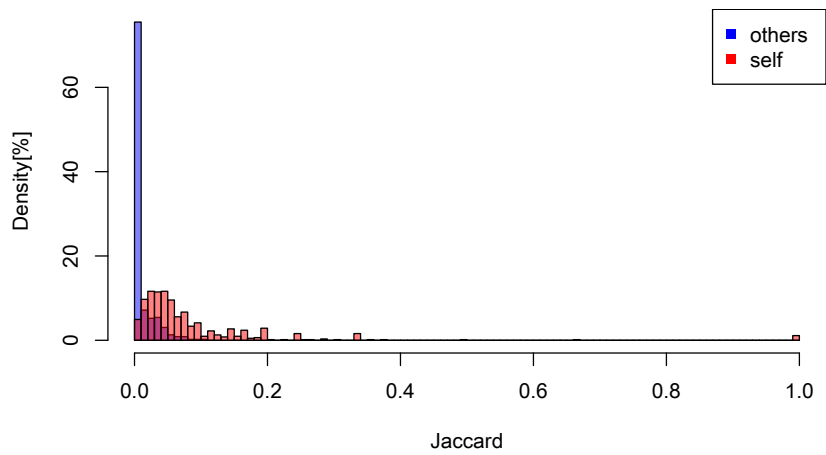


図 3.6: 送金先集合における自他アドレスの jaccard 係数ヒストグラム

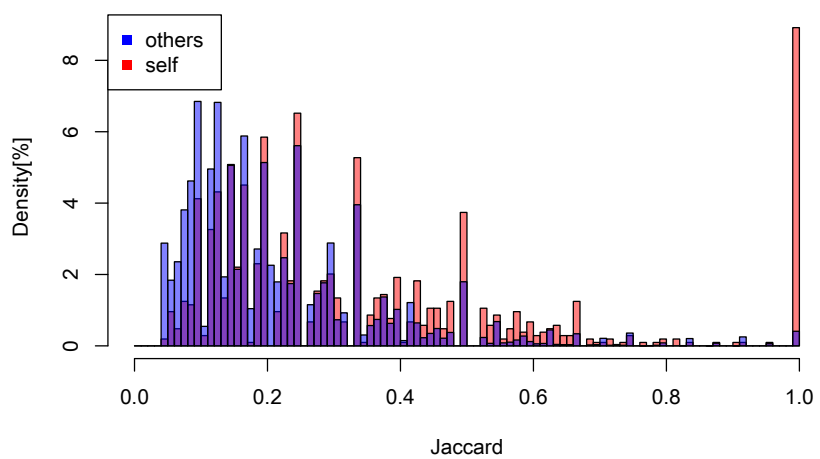


図 3.7: 時間集合における自他アドレスの jaccard 係数ヒストグラム

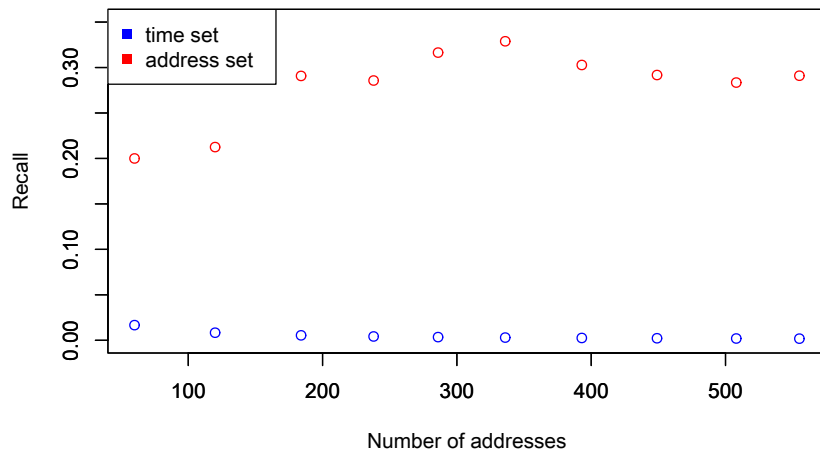


図 3.8: 送金先集合と時間集合の平均再現率の比較

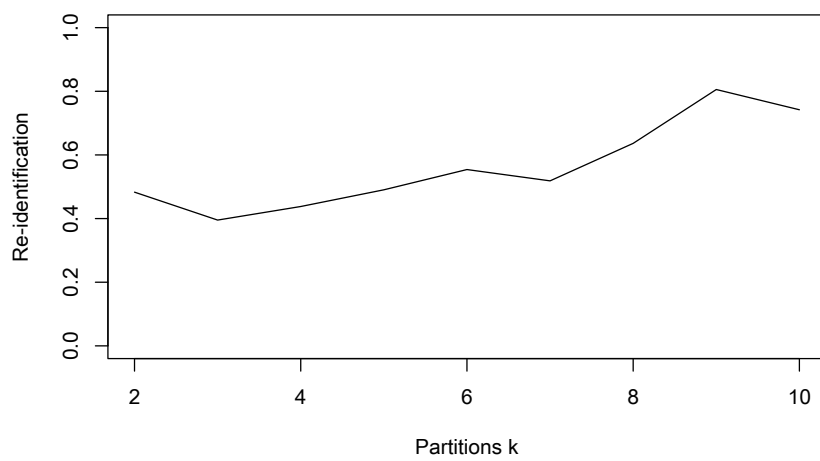


図 3.9: 分割数による識別率

# 第4章 平均取引時間分布の相関を用いた Bitcoin ユーザのタイムゾーン属性の推定

## 4.1 はじめに

本章では, Dupont らと同様に, アドレスを管理するユーザのタイムゾーン推定を行う. 本研究では取引時間分析の相関係数に基づくノイズに対する頑強性の高い推定方法を提案する. さらに, 第3章で提案した Jaccard 再識別との比較を行う.

### 4.1.1 方法

本研究の概要を図 4.1 に示す. 手順を以下で詳しく説明する.

アドレスデータセットでは実験に用いるアドレスを格納している. 取引データセットは各アドレスに関する取引を BLOCKCHAIN[7] から取得している. タイムゾーンデータセットには全タイムゾーンのデータを time zone テーブルに格納する.

平均取引時間分布データは, アドレスデータセットより 25% を一様分布でランダム抽出しこれらを学習用データとする. この学習用データを含む Address の取引データを取引データセットから抽出する. 出力された各 Address の取引データの Timestamp を UTC にし, 全ての Timestamp データをまとめ一つの平均取引時間分布データ  $f_*$  とする. 学習用データ以外の 75% のデータをテスト用データとして使用する. テスト用データの未知の Address を  $i$ , 取引時間分布  $f_i$  とする. 平均取引時間分布 24 個分ずらし 24 個の平均取引時間分布との相関係数  $c(f_i, f_*), c(f_i, f_* + 1), \dots, c(f_i, f_* + 24)$  を求め, 最大化するタイムゾーン  $i_*$  をユーザ  $i$  のタイムゾーンと推定する. すなわち,

$$j_* = \arg \max_{j \in \{0, \dots, 24\}} c(f_i, f_* + j)$$

とする.  $i$  の正しいタイムゾーン  $i_*$  と推定  $j_*$  との差が閾値  $\theta$ [時] 以内を推定成功とする. これをユーザごとに求める.

### 推定成功率の出力

本実験では, 平均取引時間分布データ作成以降の手順を 1000 回実施し, 推定成功回数の平均を求める.

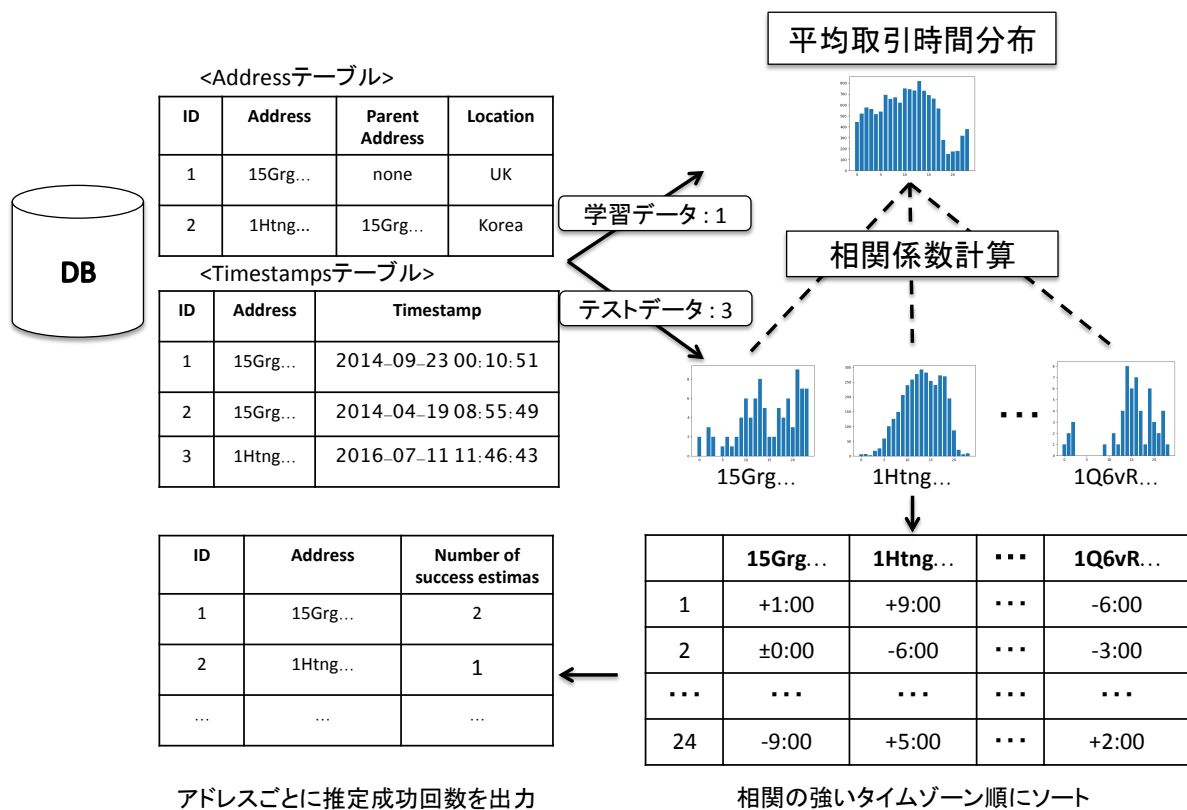


図 4.1: システム構成図

#### 4.1.2 結果

##### 推定成功回数

ユーザごとの取引回数に対し,  $j_*$  と  $i_*$  との差分

$$d_i = |j_* - i_*|$$

を定める.

##### 推定成功率の評価

推定成功回数閾値  $\theta$  回以上の推定成功回数のユーザ数の割合を推定成功率

表 4.1: データセット概要

期間	2009-1-3 2018-9-23 (9.5 年)
アドレス数	1,233
ユーザ数	1,086
タイムスタンプ数	327,310

$$s_\theta = \frac{|\{i \in \mathbb{U} \mid t_i \geq m, |j_* - i_*| \leq \theta\}|}{|\{i \in \mathbb{U} \mid t_i \geq m\}|}$$

と定める。ここで、取引回数が閾値  $m$  回以上のユーザのみを評価対象とする。

推定成功率  $m = 1$  回の条件における推定成功率  $s_1$  は 9% であった。しかし, Dupont らと同じ条件である  $m = 6$  回以上の取引回数, かつ  $\theta = 11$  回の推定成功の条件においては 77% となった。

## 4.2 第3章手法との比較・考察

Jaccard 再識別では、最大で 80.5% のアドレスが再識別が可能であった。本章のタイムゾーン識別と比較すると、少し上回った。しかし、対象アドレス数は本章における実験の方が多く、平等な条件で比較するのは困難であった。

本章では、取引時刻分布からタイムゾーンを推定するものであり、77% という推定成功率が得られた。よって、この手法をアドレス識別に応用することで、よりアドレス識別の精度を上げられると考える。



## 第5章 まとめ

本稿では、ビットコインアドレスの取引頻度と送金先集合が、どれほどアドレスの匿名性に影響を与え、識別されるリスクを評価した。さらにアドレスの取引時刻から、どれほどタイムゾーンを識別されるリスクがあるかを評価した。

送金先集合を用いた匿名性評価では、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。

タイムゾーン識別実験では、相関係数を用いることによって 77% のアドレスのタイムゾーンを識別できた。

## 参考文献

- [1] S. Nakamoto, Bitcoin: A Peer-to-Peer Electronic Cash System. <https://bitcoin.org/bitcoin.pdf>, 2008.
- [2] 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> 2018年5月参照)
- [3] マウントゴックス破綻 ビットコイン 114 億円消失 ([https://www.nikkei.com/article/DGXNASGC2802C\\_Y4A220C1MM8000/](https://www.nikkei.com/article/DGXNASGC2802C_Y4A220C1MM8000/) 2018年1月参照)
- [4] コインチェックの仮想通貨不正流出、過去最大 580 億円 (<https://www.nikkei.com/article/DGXMZO26231090X20C18A1MM8000/> 2018年1月参照)
- [5] 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.
- [6] 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
- [7] Blockchain. <https://blockchain.info>
- [8] Bitcointalk. <https://bitcointalk.org/>
- [9] A. Biryukov, D. Khovratovich and I.Pustogarov, Deanonimisation of Clients in Bitcoin P2P Network, *In Proceedings of Conference on Computer and Communications Security(CCS14)*, 2014
- [10] S. Delagdo-Segura, C. Perez-Sola, G. Navaro-Arribasv and J. Herrea-Joancomarti, “Analysis of Bitcoin UTXO set”, *In Proceedings of Conference on Financial Cryptography and Data Security (FC'18)*, 2018.
- [11] P. Koshy, D.Koshy and P. McDaniel, “ An analysis of anonymity in bitcoin using P2P network traffic” *In Proceedings of Conference on Financial Cryptography and Data Security (FC'14)*, 2014.
- [12] Serge Egelman, Eyal Peer: Scaling the Security Wall Developing a Security Behavior Intentions Scale (SeBIS), *SIGCHI Conference on Human Factors in Computing Systems (CHI' 15)*, 2015.

- [13] 井垣秀星, 永田倅大, 菊池浩明, 平均取引時間分布の相関を用いた Bitcoin ユーザのタイムゾーンの推定', 情報処理学会第 81 回全国大会
- [14] 長谷川彩子, 秋山満昭, 八木毅, 森達哉, オンラインオークションにおけるプライバシーリスクとユーザ認識の調査. コンピュータセキュリティシンポジウム (CSS2017), pp.435-442, 2017.
- [15] 川口雄己, 山田彰, 小澤誠一, 匿名ネットワーク Tor におけるマーケット商品とセキュリティ事件との関連性に関する考察. コンピュータセキュリティシンポジウム (CSS2017), pp.405-441, 2017.
- [16] 中川紗菜美, 佐古和恵, 小出俊夫, 梶ヶ谷圭祐, 不正転売問題を配慮したブロックチェーンベースのチケット管理システムの提案. 暗号と情報セキュリティシンポジウム 2018(SCIS2018), pp.1-8, 2019.

## 業績

- 永田倅大, 菊池浩明, “Bitcoin アドレスの送金先集合に基づく匿名性の評価”, 情報処理学会, CSEC-80, pp.1-6, 2018.
- Kodai Nagata, Hiroaki Kikuchi, Chun-I Fan, “Risk of Bitcoin Addresses to be Identified from Features of Output Addresses”, *The 2018 IEEE Conference on Dependable and Secure Computing(DSC 2018) Workshop #4*, pp.349-354, 2018.
- 井垣秀星, 永田倅大, 菊池浩明, “平均取引時間分布の相関を用いた Bitcoin ユーザのタイムゾーンの推定”, 情報処理学会第 81 回全国大会 (採録済み)

# 謝辞

本稿をまとめるにあたって多くの方のご指導・ご協力を賜りました。

指導教員である明治大学総合数理学部の菊池浩明教授には，研究に対する指導や，多くの成長する機会を与えていただき感謝いたします。国立中山大学 Chun-I Fan 教授には，ワークショップを通じて様々な助言をいただき感謝いたします。合同発表会を通じて助言をいただいた，静岡大学創造科学技術大学院 西垣正勝教授，静岡大学情報学部情報科学科 大木哲史先生，東京電機大学工学部理工学科情報システムデザイン学系 稲村勝樹先生に心から感謝致します。菊池研究室の皆様は，著者の研究に対してのディスカッションや，学生生活を思い出深いものにしてくれた。明治大学の井垣秀星氏には，同じ研究分野の研究者として支援していただき深く感謝します。本研究に様々な助言をしていただいた全ての皆様に感謝いたします。

最後に著者の学生生活を支えてくれた家族に感謝の意を表すると共に，謝辞にかえさせていただきます。

# 付録A Bitcoin ユーザアンケート調査

## A.1 概要

ビットコインユーザにビットコインや他の暗号通貨に関するアンケートを行った。さらにユーザのセキュリティ意識を調べるために SeBIS(Security Behavior Intentions Scale)[12] を用いた。SeBIS はセキュリティ指向度の指標である。回答者は質問に対し 5 段階で回答を行い、得点が高いほどセキュリティ意識が高いと判断する。本調査では 11 人の被験者から回答を得た。

## A.2 アンケート内容

1. Sex
2. Age
3. Your Country
4. What is your time zone?
5. How many bitcoin addresses do you have?
6. How often do you change the address?
  - Below 1 years
  - 1-2 years
  - 2-3 years
  - 3-4 years
  - Never
  - Others
7. Which environment do you use Bitcoin?
8. What kind of wallets do you use? (multiple choice)
9. How long have you used Bitcoin?
  - Every transaction
  - By daily

- By monthly
- By yearly
- Above 5 years

10. How often do you use Bitcoin?

- Almost every day
- 2-3 times per week
- 1 time per week
- 2-3 times per month
- 1 time in 2-3 month
- Less than the above

11. Do you know the terms related to Bitcoin?(multiple answers allowed)

- Blockchain
- Mining
- Node
- SPV Node
- Full Node
- PoW(Proof-of-Work)
- Electronic signature
- Segwit
- Satoshi Nakamoto
- Do not know any

12. Have you ever used Bitcoin for the following uses? (multiple answers allowed)

- Purchasing at retail stores
- Overseas remittance
- Transfer between users
- Payment for any service
- Payment with Ransomware
- Exchange of cryptocurrencies

13. How much anonymous Bitcoin is?

14. What do you think is anonymity in Bitcoin?

15. Which information do you hate to be identified from the Bitcoin address?(multiple answers allowed)

- Sex
- Country
- Time zone
- Another addresses managed by you
- Not particulary
- Others

16. Which of the following cryptocurrencies do you have? (multiple answers allowed)

- XRP
- Ethereum
- Bitcoin Cash
- NEM
- Monero
- Zcash
- Monacoin
- Do not have any cryptocurrencies
- Others

17. Do you think cryptocurrencies will widely spread in the future?

18. What is the reason for the above question?

19. SeBIS(Security Behavior Intentions Scale)

- (a) When I ' m prompted about a software update, I install it right away.
- (b) When my computer wants me to reboot after applying an update or installing software, I put it off.
- (c) I try to make sure that the programs I use are up-to-date.
- (d) I manually lock my computer screen when I step away from it.
- (e) I set my computer screen to automatically lock if I don ' t use it for a prolonged period of time.
- (f) I log out of my computer, turn it off, put it to sleep, or lock the screen when I ' m done using it.
- (g) I use a PIN or passcode to unlock my mobile phone



- (h) I use a password/passcode to unlock my laptop or tablet.
- (i) If I discover a security problem, I continue what I was doing because I assume someone else will fix it.
- (j) When someone sends me a link, I open it without first verifying where it goes.
- (k) I verify that my anti-virus software has been regularly updating itself.
- (l) When browsing websites, I mouseover links to see where they go, before clicking them.
- (m) I know what website I ' m visiting based on its look and feel, rather than by looking at the URL bar.
- (n) I backup my computer.
- (o) When browsing websites, I mouseover links to see where they go, before clicking them.
- (p) I use some kind of encryption software to secure sensitive files or personal information
- (q) I do not change my passwords, unless I have to.
- (r) I use different passwords for different accounts that I have.
- (s) I do not include special characters in my password if it ' s not required.
- (t) When I create a new online account, I try to use a password that goes beyond the site ' s minimum requirements.
- (u) When I ' m done using a website that I ' m logged into, I manually log out of it.
- (v) I submit information to websites without first verifying that it will be sent securely (e.g., SSL, “ https:// ” , a lock icon).
- (w) I use privacy software, “ private browsing, ” or “ incognito ” mode when I ' m browsing online.
- (x) I clear my web browsing history.

### A.3 結果

表 A.1 に質問 11 の結果を示す。ビットコインに関する特有の用語，例えば Blockchain, Node, Segwit や Satoshi Nakamoto を知っている被験者が多かった。一方でビットコイン取引で用いられる，一般的な技術 Electronic signature を知っている被験者は少なかった。

図 A.1 に，質問 13 と各被験者の SeBIS スコアの関係を示す。被験者は質問 13 に対して 7 段階で回答しており，数字が高いほど Bitcoin の匿名が高いと認識している。匿名性が高いと感じている被験者の方が，そうでない被験者より SeBIS のスコアが低いという結果が得られた。

表 A.1: 質問 11 結果

Blockchain	10
Mining	7
Node	7
SPV Node	7
Full Node	7
PoW(Proof-of-Work)	8
Electronic signature	4
Segwit	10
Satoshi Nakamoto	11
Do not know any	0

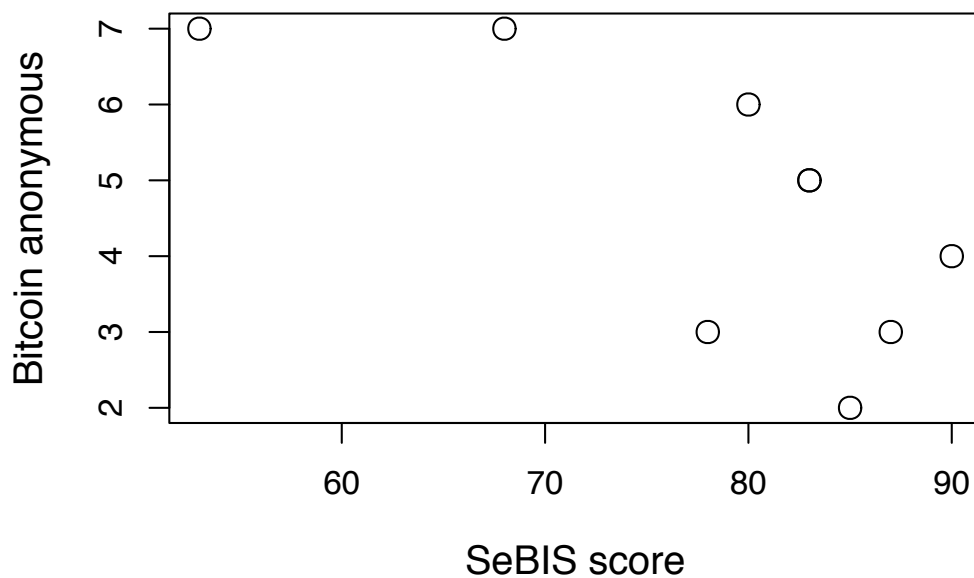


図 A.1: 質問 13 と SeBIS スコアの関係