明治大学大学院　先端数理科学研究科

2020 年度

修士学位請求論文

取引履歴の特徴量に基づく Bitcoin アドレスの識別リスクの評価

学位請求者　先端メディアサイエンス専攻

松本 寛輝

あらまし

Bitcoin では, プライバシー保護の観点からアドレスの使用を一度限りとし, ユーザは自分のアドレスを明かさないようにすることが推奨されている. しかしながら, 取引に利用されているアドレスの中には再利用されているものも多く, 送金を行った取引情報の特徴量に基づいて同一ユーザが所有するアドレスの識別リスクを分析した研究成果が報告されている. 本研究は,Bitcoin の受け取り取引に着目した新たな 2 つのアドレス識別手法を提案し, 識別精度を明らかにすることを目的とする. Bitcoin の取引情報を用いて識別実験を実施し, アドレスの取引回数と利用目的の観点による精度の変化を評価する. 実験結果より, 提案手法のアドレス識別精度は既存手法を利用した識別精度よりも統計的に有意な水準で高いことが示された.

目 次

第1章　はじめに

1.1　研究背景

Bitcoin[1] を代表とする暗号資産は高い匿名性を持つとされ, 国を超えた送金や投資目的など様々な用途で利用されている. しかしながら,Bitcoin が持つ匿名性は Bitcoin アドレスが持つ仮名のランダム性に基づくものであり, 取引履歴の統計情報からのアドレスの識別やユーザ居住地等の属性情報が推定されるリスクが知られている. Bitcoin の追跡可能性についてはオープンソースプロジェクトの Bitcoin.org で次のように述べられている [17].

> ビットコインは, 殆どの人が今まで経験した事のないほどの透明性をもって動作します. ビットコインのトランザクションの全ては, 公共性があると共に追跡可能であり, 永久的にビットコイン・ネットワークに保管されます. ビットコイン・アドレスは, ビットコインがどこに割り当てられ, どこに送付されたかを定義するのに使用される唯一の情報です. これらのアドレスは, 各ユーザーのウォレットで個人的に生成されています. しかし, アドレスが使用されると, それが自分の全トランザクション履歴と関連付けられ, 誰もが, これらのアドレスの残高と全トランザクション履歴を見る事ができます. ユーザーがサービスや商品を受け取るには, 身元を明かさなければならないため, ビットコイン・アドレスは完全に匿名であり続ける事はできません. また, ブロックチェーンは永久に存在するため, 現在は追跡不能であっても, 将来的には追跡可能性が生じうる点は, 留意すべき事でしょう. そのため, 各ビットコイン・アドレスの使用は一度限りとし, ユーザーは自分のアドレスを明かさない様に注意しなければなりません.

Bitcoin.org では, プライバシー保護の観点からアドレスの使用を一度限りとし, ユーザは自分のアドレスを明かさないようにすることを推奨している. ユーザは取引ごとに新たなアドレスを作成することで匿名性を高めることが期待できる.

その一方で, 用途によってはアドレスが長期間に渡り繰り返し利用されることがある. 代表的な例として, 自身のアドレスへ寄付を受け付ける目的で掲示板や SNS などに公開する場合や Mining pool 事業者などが営利用アドレスを意図的に使い続ける場合等がある. 同じアドレスを一定回数繰り返し使用することで同一ユーザの所有するアドレスが識別されるリスクは否定できない.

1.2　研究目的

Bitcoin アドレスの識別リスクを評価する先行研究として,Meiklejohn らは, 取引の *Input* に並列に指定された複数アドレスが同一ユーザによることを指摘し, そのリスクを分析している [2]. また, 永

表 1.1: 先行研究との比較

	Meiklejohn らの研究 [2]	永田らの研究 [5]	本研究の提案手法 1	本研究の提案手法 2
目的	Bitcoin 市場の調査	アドレス識別手法の提案		
手法	取引の *Input* アドレスを利用	取引の宛先アドレスを利用	取引の受け取りアドレスを利用	取引の *Output* アドレスを利用
対象期間	2009 — 2013/4/13	2012/9/22 — 2014/5/10	2009/1/4 — 2019/11/18	
アドレス数	12,056,684	559	45,329	
アドレスの種類	6	2	5	

田らは, 取引の宛先を学習することで, 同一ユーザの所有するアドレスであるかを識別できることを示している [5].

しかしながら, 取引履歴から学習されるユーザの特徴はそれらに限らない. 本研究では, 取引の *Input* でなく *Output* に指定されたアドレスにも, ユーザを特定する重要な特徴があることを新たに主張する. 加えて, 取引の宛先だけでなく, 送信元の情報の特異性にも着目し, 新たな識別方式を 2 つ提案する.

一方, 識別精度を正確に評価するのは困難である. なぜならば, そのアドレスに関わる取引の頻度や, 交換所などのサービス事業者によるものか, 単なるエンドユーザのものかといった利用目的などに応じて識別精度が大きく左右するためである. そこで, 掲示板,ATM, 交換所, マイニングプール,Dark web サービスの 5 つの代表的なアドレスを取り上げ, それらによる識別率の変化を明らかにする.

先行研究の Meiklejohn ら [2] と永田ら [5] の 2 つの手法と, 本研究で新たに提案する 2 つの識別手法の違いを表 1.1 に示す.

1.3 研究の貢献

本稿の貢献は次の通りである.

- 取引の特徴量を用いて, 新たに 2 つの識別方法を提案した.
- 10 年間のアドレスデータセットを用いて, 長期間継続して利用されているアドレスの有する識別リスクを定量化した.
- 5 種類の利用目的を考慮したアドレスの識別精度を明らかにした.

本研究は,Bitcoin アドレスの識別実験において大規模な取引期間と複数の利用目的を考慮した最初の報告である.

先行研究の 2 つの識別手法と本研究で新たに提案する 2 つの識別手法は,Bitcoin と同種の取引方式を採用している Ethereum[1]や Zcash[2]等の暗号資産においても有効な手法である. 本稿で明らかになっ

[1]Ethereum (`https://ethereum.org/ja/`)
[2]Zcash (`https://z.cash/`)

表 1.2: 本稿の構成とこれまでの業績の対応

本稿の構成	学会名 (業績)
1 章	CSS2020 (国内研究会 1.)
2 章	
3 章	
4 章	
5 章	CSEC88,NBiS2020 (国内研究会 2. 国際会議論文 1.)
6 章	CSS2020,CSEC88,NBiS2020
7 章	

たアドレスの識別リスクは Bitcoin 以外の暗号資産にも共通する課題であり,Bitcoin とは異なる暗号資産にも同様の結果が期待される.

1.4 本稿の構成

本稿の構成は以下のようになっている.

1 章では, 研究背景とその目的を示し, 本稿における貢献をまとめた. 2 章では,Bitcoin の基礎知識と本稿で扱う Bitcoin アドレスの識別と取引構造を定義し, 既存手法を説明する. 3 章は本稿で提案する新たな識別手法とアドレスの利用目的について述べる. 4 章で,4 つの手法を用いたアドレス識別実験を実施し, 実験結果に関する考察と既存手法と提案手法の精度を比較した検定などの評価を行う. 5 章では, アドレスの利用目的の特徴を分析するため,7 種類の利用目的のアドレス推定実験を実施し, 考察を行う. 6 章は本稿で取り扱ったアドレスや提案手法に関する制限事項を整理する. 7 章で本稿の結論をまとめる.

本稿で報告する研究結果と本稿巻末の筆者の研究業績と対応した構成を表 1.2 に示す.

第2章　基本定義

2.1　暗号資産とBitcoin(ビットコイン)

暗号資産 Bitcoin[1] は 2008 年に仮名の著者 Satoshi Nakamoto によって提案された．その後, ウォレットソフトウェアが広く実装され,2021 年 1 月 7 日には 1BTC が約 400 万円以上で取引が行われている．

多くの暗号資産はブロックチェーン (blockchain) と呼ばれる技術を利用している．ブロックチェーンとは, ブロックと呼ばれる単位でデータを格納・検証し, 鎖のようにつなげて蓄積する方式である．ブロックの中には, 送金元や宛先, 金額, 時刻などが記載された取引情報やナンス (nonce: number used once) と呼ばれる乱数が記録されている．また, ブロックに記録された取引情報は全て一般に公開されているため, 誰でも閲覧可能である．

ブロックチェーンの仕組みを図 2.1 に示す．ブロックが正しく検証 (承認) されるためには, ナンスの値を変更し, ハッシュ値が条件 (先頭が 000 . . . となる) を満たすまで繰り返し計算される．図 2.1 では, ブロック 100 を作成するために,1 つ前のブロック 99 のハッシュ値が条件を満たすように計算が行われている．この作業はマイニング (mining) と呼ばれている．マイニングを行う際には膨大な計算量が必要なため, 過去の取引を改竄するためには, 現在までに作成された全てのブロックのハッシュ値を再度計算する必要がある．そのため, ブロックチェーンに一度取り込まれた情報は改竄することが難しいと考えられている．

ブロックには, 取引の送金元, 宛先にアドレスと呼ばれる公開鍵が記録されている．図 2.2 に Bitcoin アドレス “1BGgHJDv2Z8WyjCHfecjdSVUs6cfdjqBAJ” を例として示す．ユーザはアドレスを複数所有することが可能であり, ウォレット [18] と呼ばれるソフトウェア上で管理する．ウォレットの代表的な種類と利用例を表 2.1 に示す．

[1]BLOCKCHAIN ONLINE ブロックチェーンの基本的な仕組み (`https://blockchain-jp.com/guides/4`)
[2]bitpay (`https://bitpay.com/`)

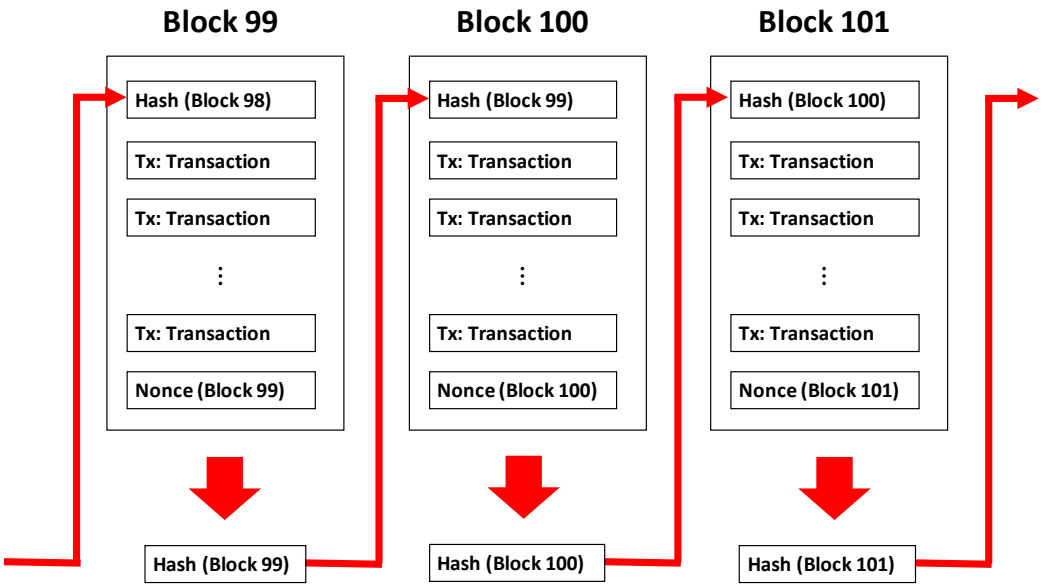


図 2.1: ブロックチェーンの仕組み[1]

図 2.2: Bitcoin アドレスの例

表 2.1: ウォレットの種類と例

種類	管理方法	利用例
オンラインウォレット	web 上でアカウントを作成	交換所 (Coincheck[2]) など
モバイルウォレット	スマホなどのアプリケーション	bitpay(図 2.3) など
デスクトップウォレット	PC 上のソフトウェア	Bitcoin Core[3]など
ハードウェアウォレット	USB などの記録媒体	Ledger[4]など
ペーパーウォレット	QR コードなどでアドレス 秘密鍵を紙に印刷	図 2.2 など

[2] Coincheck (`https://coincheck.com/ja/`)

[3] Bitcoin Core (`https://bitcoincore.org/ja/`)

[4] Ledger (`https://hardwarewallet-japan.com/`)

図 2.3: モバイルウォレット (bitpay[2]) の例

図 2.3 にモバイルウォレットの例を示す．ここで, 表 2.2 を用いて, アドレスとウォレットの関係を銀行口座の例に説明する．送金者 (明治太郎) の銀行口座には 30,000 円の預金があり, 送金先 (明治花子) には 10,000 円送金を行うと仮定する．送金者が送金する際には, 送金先や口座番号を指定し, 送金したい金額と手数料を銀行に振り込む.

Bitcoin の送金は, 送金先のアドレスと送金したい金額, 手数料を指定する．このとき, 手数料はブロックの承認作業を行うマイナー (miner) に支払われる．取引を確定するためには, 送金元アドレスの秘密鍵を用いて取引に署名を行い, 署名された取引情報を Bitcoin ネットワーク上に伝搬させる．Bitcoin ネットワーク上の取引情報は, マイナーが自身のブロックに取り込む．このブロックがマイニングに成功することで, 取引が承認される.

図 2.3 では, 送金先の名義人 (bfly) に対して,0.025124BTC を送金した記録が残っている．口座の管理者名 (MMWallets) や送金先の名義 (bfly) はウォレットの管理者が設定した情報であり, 一般的な情報ではないことに注意せよ.

2.2 Bitcoin アドレスの匿名性の定義と問題

Bitcoin はブロックチェーン技術を用いることで匿名性に優れていると考えられている．ここで,Bitcoin の匿名性の定義はアドレスと管理者 (ユーザ) の情報が結びつかないこととする．Bitcoin アドレスの管理とは, ユーザがアドレスに対応する秘密鍵を保有し, 所有者のみが対応する資産を自由に移動する権限を有することと定める.

例として, 図 2.2 に示した Bitcoin アドレス “1BGgHJDv2Z8WyjCHfecjdSVUs6cfdjqBAJ” を考える．このアドレスの管理者について, アドレスの情報のみを用いて管理者を特定することは困難であ

表 2.2: ウォレットとアドレスの関係

取引情報	銀行口座の例	ウォレット (図 2.3) の例
サービス名	明治銀行	bitpay
口座の管理者	明治太郎	MMWallets
送金先の名義	明治花子	bfly
送金先	送金先の口座番号	アドレス (図 2.3 中には記載なし)
送金額	10,000 円	0.025124 BTC
送金手数料	410 円	0.00002 BTC(図 2.3 中には記載なし)
送金日	2021/1/7	1 月 7,2021
残高	19,590 円	0.00 BTC

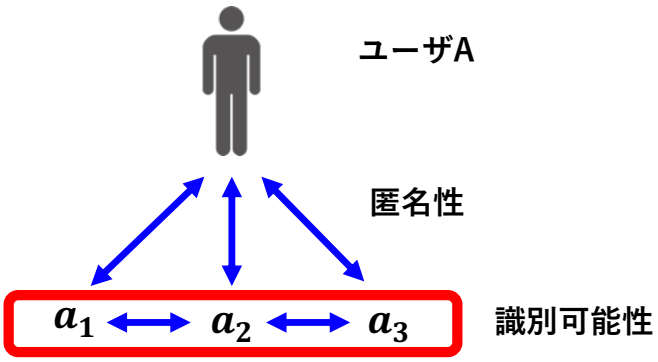


図 2.4: Bitcoin アドレスの識別と匿名性

る. 同様に, ブロックチェーンに記録されているアドレスの情報を管理者と結びつけることは難しい. 例外として, 交換所で管理されているアドレスは利用契約時に本人確認が必要な場合がある. しかし, 管理者とアドレスを紐付ける情報は公開されていないため, アドレスの情報から個人を特定することはできない. 以上が Bitcoin は匿名性に優れていると考えられている要因である.

匿名性の一方で, ユーザのアドレスを管理する交換所が不正アクセスの被害に遭うケースが発生している. Bitcoin の流出事件では 2014 年当時最大規模の交換所であった Mt.Gox が約 390 億円分の盗難被害に遭う事件が発生した [19]. 国内で発生した流出事件としては 2018 年 1 月に Coincheck から暗号資産 NEM が約 580 億円分の盗難被害に遭い, 大きな社会問題となった [20]. 同事件の後も 2018 年 9 月に Zaif より総額 70 億円相当 [21] の暗号資産が流出し,2019 年 7 月にはビットポイントより総額 30 億円相当 [22] の暗号資産が流出している.

これらの事件は,Bitcoin やブロックチェーンの脆弱性ではなく, 交換所システムのセキュリティーホールを狙われた被害である. ブロックチェーンの理論が正しいとしても, 運用する交換所やユーザの利用方法によって, 個人が特定される危険性は残っていると考える.

2.3 Bitcoin アドレスの識別の定義

Bitcoin アドレスの高い匿名性に対して, アドレスが識別される (識別可能性) リスクが知られている.

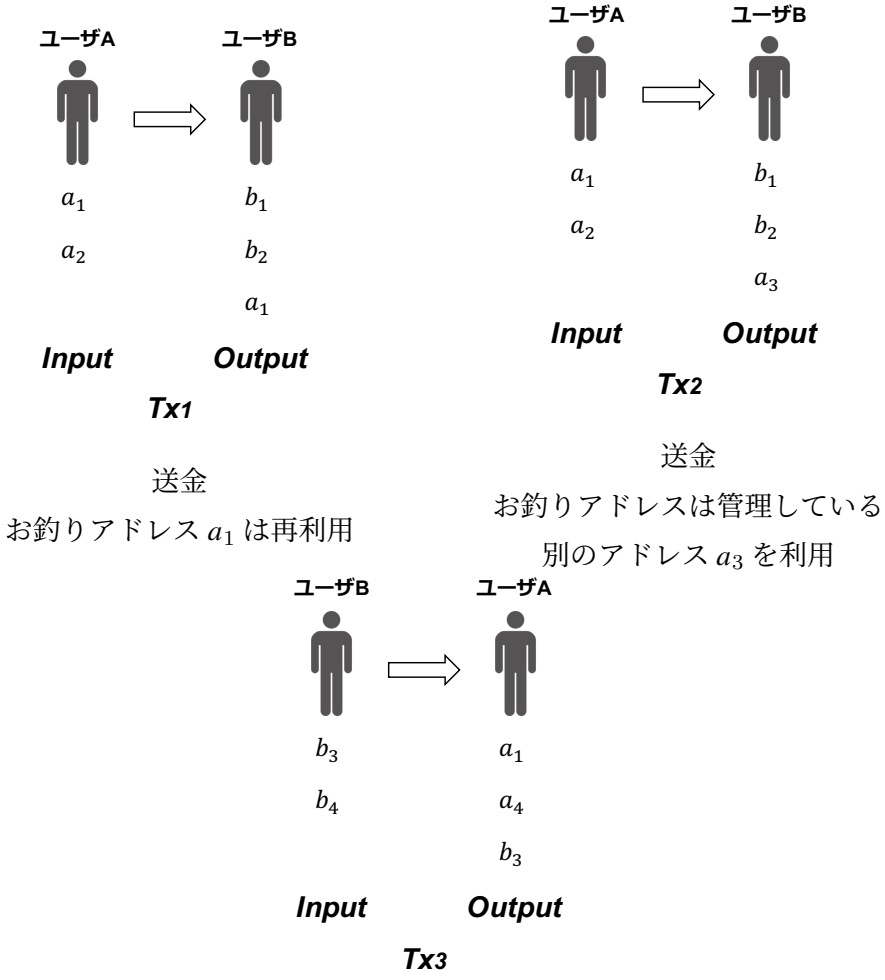


図 2.5: Bitcoin の送金, 受け取りを行う取引構造

Bitcoin アドレスの識別問題は, あるユーザが管理している複数のアドレスを与えて, そのユーザの管理する他の未知のアドレスを他人のアドレスから識別する問題である. 例えば, 図 2.4 のアドレスの識別例を考えよう. ここではユーザ A が a_1,a_2,a_3 のアドレスを管理している. a_1 が与えられた時, 対象全アドレスの集合から A の管理する a_2 や a_3 を正しく選べれば識別が成功したと考える. Bitcoin アドレスが識別されても, 必ずしもアドレスの所有者が特定されるわけではないことに注意せよ. アドレス a_1,a_2,a_3 の情報から管理しているユーザの名前 A などが特定されることではない.

2.4 Bitcoin アドレスの取引

Bitcoin アドレスに関する取引の例を図 2.5 に示す. Bitcoin の取引 Tx 中に, 送金を行うアドレスは *Input*, Bitcoin を受け取るアドレスは *Output* に指定されている. 図 2.5 の Tx_1 では, ユーザ A が管理しているアドレス a_1,a_2 を用いてユーザ B が管理しているアドレス b_1,b_2 へ送金を行っている. このとき, ユーザ A は送金を行った際のお釣りを受け取るため *Output* に自身のアドレス a_1 を指定している. 図 2.5 の Tx_1 では, 送金時に使用したアドレス a_1 を再度用いてお釣りを受け取っているが,Tx_2 の

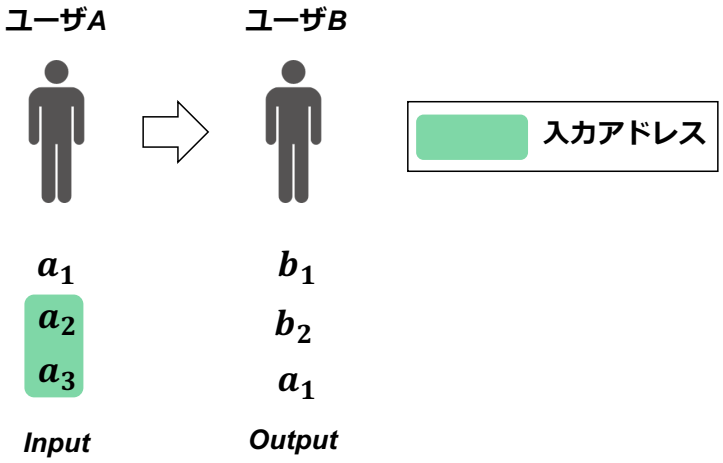


図 2.6: 入力アドレスを特徴量としたアドレス識別 [Meiklejohn,2013]

様に, ユーザ A が自分が管理する別のアドレス a_3 でお釣りを受け取ることもある. Tx_3 では, ユーザ A が管理しているアドレス a_1,a_4 に対してユーザ B が管理しているアドレス b_3,b_4 から送金を行っている. このとき, ユーザ B は送金を行った際のお釣りを受け取るため $Output$ に自身のアドレス b_3 を指定している.

2.5 関連研究

Bitcoin アドレスのプライバシーに関しては, アドレスの匿名性 (識別可能性) に関するものと, タイムゾーンなどの属性を推定するもの, 大きく 2 つの流れがある. 前者の匿名性に関して次の研究 [2],[5] がある. 加えて, 後者の Bitcoin アドレスや管理ユーザのプライバシーに焦点を当てた研究報告についても述べる.

2.5.1 Meiklejohn らの入力アドレスを用いた識別

Meiklejohn らは取引の入力アドレスを用いた識別手法を提案している [2].

Bitcoin の 1 つのトランザクションに送金を行うアドレス (入力アドレス) が複数指定されている場合, 全ての入力アドレスについての署名が必要なために入力アドレス a の秘密鍵が同一のユーザ (管理者) によって管理されていることを利用して, それらの入力アドレスを識別, すなわち, 単一のユーザに管理されていることを明している.

Kappos らは匿名性が Bitcoin よりも高いとされる暗号資産 Zcash においても同様の手法を用いた識別実験を実施している [4]. Bitcoin 以外の暗号資産にも有効である一般的な識別手法である.

図 2.6 に入力アドレスの例を示す. この例では, ユーザ A が管理するアドレス a_1 の入力アドレスは, ユーザ A のアドレス a_2, a_3 となる.

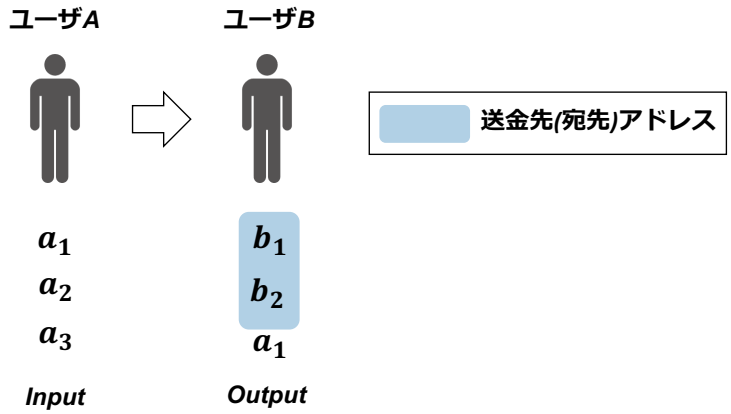


図 2.7: 送金先 (宛先) アドレスを特徴量としたアドレス識別 [永田,2018]

2.5.2 永田らの送金先アドレス集合を用いた識別

永田らは取引の送金先アドレス集合を用いた識別手法を提案している [5]. この手法は, ユーザごとに取引を行う相手のユーザ群が決まっていることを仮定してユーザを追跡する. アドレスの取引頻度と送金先履歴の宛先アドレスを用いて, 過去に行った取引履歴からアドレスを識別する. 永田らは実験に基づき, アドレスの取引数はアドレスの識別に影響を与えない, と主張している.

図 2.7 に送金先アドレスの例を示す. この例では, ユーザ A が管理するアドレス a_1 の送金先アドレスは, ユーザ B のアドレス b_1, b_2 となる.

永田らによる宛先アドレスを用いたアドレス識別手法では, 送金時に利用するアドレスはユーザによって任意に変更することが可能である. 従って, 通常とは異なる複数のアドレスを意図的に混ぜることでアドレス識別を撹乱可能である課題がある.

2.5.3 ユーザが居住している地域のタイムゾーン等の属性推定

Dupont らはアドレスの取引時刻に着目し, 取引の時刻分布からユーザが居住している地域のタイムゾーンを推定する方式を提案している [6]. この手法は, 取引が行われていない時間帯を夜間 (ユーザが眠っている時間帯) と定めることで, タイムゾーンを予測する. また, アドレスとユーザの居住地の情報を示す正解データは, 掲示板サイトでユーザが公開しているプロフィールページの情報を利用している. Dupont らはタイムゾーンの平均推定精度は最大で 72%と報告している.

井垣らはアドレスの平均取引時間分布を用いることで, 最大で 77%のアドレスのタイムゾーンを推定可能と報告している [7]. この研究では,Dupont らの手法では小さなノイズへの耐性が低いことを指摘し, ユーザの取引時間分布を特徴量とした平均取引時間分布との相関係数を利用することで耐ノイズ性が高いことを主張している. 最大の推定精度は,Dupont らは 72%に対して, 井垣らが提案した手法では, 最大で 77%であることを報告している.

山崎, 草野らは取引所を利用しているユーザのタイムゾーンに着目し, 取引所を運営している企業が属する国とユーザの居住地域に関する分析を行っている [8]. この手法は, 世界 80 か所の取引所が管理するアドレスデータを利用し, 取引が行われている時間帯と日本におけるインターネット利用時間帯

のデータを比較することで推定を行っている．山崎らは,80 か所の取引所のうち,65 の取引所が推定値と正解の差が 2 時間以内でである結果を報告している．

2.5.4　アドレスが利用されたサービスの推定

Harlev らは, 著名なサービス事業者のアドレスを用いてアドレスのサービスを推定している [9]．利用されたサービスは交換所やギャンブルなど 10 サービスで, ランダムフォレスト等の機械学習手法を用いて推定している．サービスの推定は 77%の正解率で識別可能であることを報告している．

2.5.5　Bitcoin の取引構造を分析した研究

ブロックチェーン上に公開されている Bitcoin の取引構造を分析し, 資金の流れを追跡する研究が行われている．Ron と Shamir はブロックチェーン上に記録された全ての取引を分析し, 複数のアドレスに共通した特異な取引構造があることを報告している [3]．Ron らが行った取引構造の分析に対して, 取引の特徴量を学習させないためにミキシングサービスが利用されることがある．廣澤らは実際にミキシングサービスを利用した追跡の困難性について報告している [10].

特定のアドレスが行った取引を分析し, 追跡を行った研究が行われている．Huang らはランサムウェアの支払いに使用された Bitcoin アドレスに注目し, ランサムウェアの被害者フォーラムで報告されたアドレスを収集し, 資金の流れを追跡している [11].

井垣らはカナダに設置された Bitcoin ATM と呼ばれる預貯金サービスに着目し, サービスを利用しているユーザの分析を行っている [13]．井垣らは実際にサービスを利用することで ATM に登録されたアドレスを収集し,ATM のアドレスと取引を行ったユーザのアドレスを収集している．

2.5.6　Bitcoin の脆弱性に着目した研究

Bitcoin や関連するシステムの脆弱性を指摘した研究が報告されている．坂間らは Bitcoin の取引時に使用するデジタル署名を分析し, 利用者が使用しているウォレットが原因でアドレスの秘密鍵が漏洩する危険性について考察している [12]．先行研究で報告された乱数の再利用が秘密鍵の漏洩に繋がる脆弱性に対して, 新たに調査実験を実施し, 先行研究と同様の被害が現在も発生していることを報告した．また, 坂間らはこの脆弱性を検知するためのシステムを作成し, その評価実験も実施している．

2.5.7　取引情報が伝搬される通信路に着目した研究

Bitcoin のアドレス空間ではなく, 送金, 受け取りを行うネットワークのレイヤーに着目した研究が行われている．

Biryukov らは Bitcoin の送金が行われたウォレットの IP アドレスに着目し, 通信元からユーザを識別する方式を提案している [14]．また, ユーザが通信元を秘匿して取引を行うため,Tor と呼ばれる匿名通信路の利用を回避する方式を提案している．Garba らは web サイト上に公開されているアドレスを収集し,Bitcoin の支払い時における中間者攻撃のリスクについて考察している [15]．Garba らは独

自に web サイトをクローリングし, 掲載されている Bitcoin のアドレスの種類や web サイトの暗号化通信への対応状況などを調査した. 収集した 10,045 個のアドレスのうち,5,393 個 (48%) のアドレスは通信の暗号化がされていない web サイト上に掲載されており, 中間者攻撃によってサイト訪問者が誤ったアドレスに対して送金してしまうリスクを報告している.

Bitcoin 送金時の通信路に存在するリスクに対して Shaileshh らは Dandelion と呼ばれる通信の伝搬方式を提案している [16]. この手法はユーザが Bitcoin の送金時に複数の中継ノード経由したのち, 取引情報を伝搬することで自身のウォレットの IP アドレスを秘匿する仕組みである.

第3章　アドレス識別手法の提案

3.1　アドレス集合の定義

本稿で取り扱う 4 つのアドレス集合を定義する.

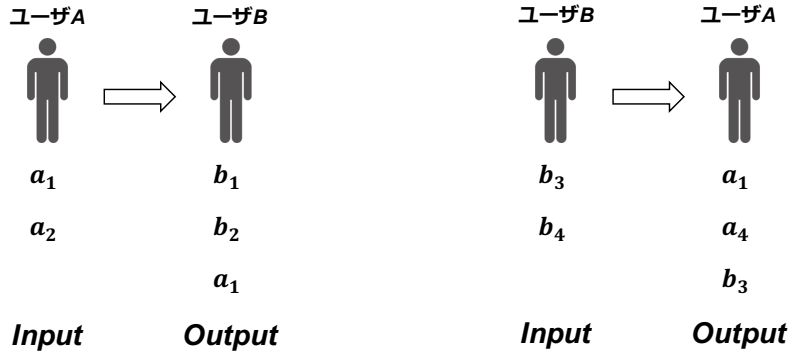
アドレス a_1 の**宛先アドレス集合** $S(a_1)$ は,a_1 から期間内に一度でも送金を行ったアドレスの集合である. 宛先アドレス集合は先行研究で永田らが定義した送金先アドレス集合と同一である. アドレス a_1 の**入力アドレス集合** $I(a_1)$ は,a_1 から送金を行った際に同時に *Input* フィールドに指定されたアドレスの集合である. 入力アドレス集合は Meiklejohn らが定義した入力アドレスの集合と同一である. アドレス a_1 の**送金元アドレス集合** $R(a_1)$ は,a_1 に対して期間内に一度でも送金を行ったアドレスの集合とする. アドレス a_1 の**出力アドレス集合** $O(a_1)$ は,a_1 に対して送金が行われた際に, 取引の *Output* フィールドに指定されたアドレスの集合とする.

これらのアドレス集合を図 3.1 と表 3.1 に整理する. 4 つのアドレス集合には識別を行うアドレス a_1 を含めないことに注意せよ.

3.2　Jaccard 係数を用いた識別

アドレス識別の評価には Jaccard 係数を用いた集合の類似度を利用する. Jaccard 係数とは, 集合 A と集合 B について,$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$ で定められる A,B 間の類似度である [23].

永田らの宛先アドレス集合 S を用いたアドレス識別手法の例を図 3.2 に示す. 識別対象のアドレス a_1, a_2, a_3 から送金された 3 つの取引 Tx_1, Tx_2, Tx_3 の宛先アドレス集合を S_1, S_2, S_3 とする. 学習アドレスを $L = S_1 \cup S_2$ と評価アドレスを S_3 とする.



アドレス a_1 が送金　　アドレス a_1 が受け取る

図 3.1: 4 つのアドレス集合を示す取引例

表 3.1: a_1 についての 4 つのアドレス集合

アドレス集合	定義	図 3.1 中での例	識別方式
宛先アドレス集合 $S(a_1)$	a_1 から送金するアドレスの集合	$\{b_1, b_2\}$	永田ら [5]
送金元アドレス集合 $R(a_1)$	a_1 に送金を行うアドレスの集合	$\{b_3, b_4, a_2\}$	本研究
入力アドレス集合 $I(a_1)$	a_1 の送金時に同時に *Input* に指定されるアドレスの集合	$\{a_2\}$	Meiklejohn ら [2]
出力アドレス集合 $O(a_1)$	a_1 の受け取り時に同時に *Output* に指定されるアドレスの集合	$\{a_4, b_1, b_2, b_3\}$	本研究

識別対象のアドレス	識別対象のアドレスから送金された宛先アドレス S		
	Tx_1 S_1	Tx_2 S_2	Tx_3 S_3
a_1	$\{a_9\}$	$\{a_3, a_6, a_9\}$	$\{a_3, a_6\}$
a_2	$\{a_5, a_6\}$	$\{a_3, a_4\}$	$\{a_6, a_9\}$
a_3	$\{a_{11}, a_{13}, a_{15}\}$	$\{a_{11}, a_{15}\}$	$\{a_9\}$

学習　　テスト

識別対象のアドレス	宛先アドレス集合 S	
	学習アドレス集合 $L = S_1 \cup S_2$	評価アドレス集合 S_3
a_1	$\{a_3, a_6, a_9\}$	$\{a_3, a_6\}$
a_2	$\{a_3, a_4, a_5, a_6\}$	$\{a_6, a_9\}$
a_3	$\{a_{11}, a_{13}, a_{15}\}$	$\{a_9\}$

図 3.2: 宛先アドレス集合 S と Jaccard 係数を用いた識別の例

ここで, ある未知のアドレス a_i の宛先集合 $S_3(a_i) = \{a_3, a_6\}$ が,a_1, a_2, a_3 のどのアドレスから送信されたかを識別したい. ここで図 3.2 の例では,$S_3(a_i)$ と 3 つの学習データ La_1, La_2, La_3 の Jaccard 係数の値は以下のようになる.

$$
\begin{aligned}
J(La_1, S_3(a_i)) &= \frac{|\{a_3, a_6\}|}{|\{a_3, a_6, a_9\}|} = 0.66 \\
> J(La_2, S_3(a_i)) &= \frac{|\{a_3, a_6\}|}{|\{a_3, a_4, a_5, a_6\}|} = 0.50 \\
> J(La_3, S_3(a_i)) &= \frac{\phi}{|\{a_3, a_6, a_{11}, a_{13}, a_{15}\}|} = 0
\end{aligned}
$$

従って, 他のアドレスと比較し $J(La_1, S_3(a_i))$ が Jaccard 係数の値が最も高いため,$S_3(a_i)$ は $a_i = a_1$ から送信されたと推測する. この例ではアドレス識別に成功している. Jaccard 係数が最も高い値を持つアドレスが複数存在する場合は, アドレス識別が失敗したとみなす.

3.3 提案方式 1:取引の送金元アドレスを用いた識別

本手法では, 識別対象のアドレスが Bitcoin を受け取る際の取引に着目する. これは自身のアドレスに対して送金を行うアドレスは指定することができない, という仮説に基づいている.

本手法では, 対象アドレス a に向けて送金をした取引があるアドレス, すなわち, 送金元アドレス集合 $R(a)$ を用いて Jaccard 係数に基づいて識別する. 提案方式を図 3.3 に示す.

入力: $a_1, \ldots, a_m$ の取引 Tx の集合
　　未知アドレス x の送金元アドレス集合 $R(x)$

出力: x の推定アドレス $a_x \in \{a_1, \ldots, a_m\}$

1: $a_1, \ldots, a_m$ の送金元アドレス集合 $R(a_1), \ldots, R(a_m)$ を求める.

2: 未知のアドレス x について, 類似度最大のアドレス a_{x^*} を求める.

$$a_{x^*} = \underset{a \in \{a_1, \ldots, a_m\}}{\arg\max} \; J(R(x), R(a))$$

3: a_{x^*} を出力する.

図 3.3: 提案手法 1 送金元アドレス集合 $R(a)$ を用いた識別手法アルゴリズム

入力: $a_1, \ldots, a_m$ の取引 Tx の集合
　　未知アドレス x の出力アドレス集合 $O(x)$

出力: x の推定アドレス $a_x \in \{a_1, \ldots, a_m\}$

1: $a_1, \ldots, a_m$ の出力アドレス集合 $O(a_1), \ldots, O(a_m)$ を求める.

2: 未知のアドレス x について, 類似度最大のアドレス a_{x^*} を求める.

$$a_{x^*} = \underset{a \in \{a_1, \ldots, a_m\}}{\arg\max} \; J(O(x), O(a))$$

3: a_{x^*} を出力する.

図 3.4: 提案手法 2 出力アドレス集合 $O(a)$ を用いた識別手法アルゴリズム

3.4　提案方式 2:取引の出力アドレスを用いた識別

本手法では, 識別対象のアドレスに対して送金が行われる際に, 同時に Bitcoin を受け取るアドレスに着目する. 提案方式 1 における $R(a)$ の代わりに, 出力アドレス $O(a)$ を用いる方式を提案方式 2 とする. 提案方式を図 3.4 に示す.

3.5　4 つのアドレス集合の識別精度

4 つのアドレス集合と識別精度が高くなると予測される取引の例を表 3.2 に示す.

永田ら [5] によって提案された宛先アドレス集合 S は, ユーザごとに取引を行う相手のユーザ郡が決まっていることを仮定している. 従って, 特定の取引相手 (宛先アドレス) に対して, 頻繁に送金を行ったアドレスは識別精度が高くなると考えられる. Meiklejohn ら [2] の入力アドレス集合 I を用いた識別は, ユーザが複数のアドレスを利用して送金を行うことで識別精度が高くなる. アドレスの識別精度はユーザが管理するアドレス数に依存すると考えられる. 本稿の提案手法 1. の送金元アドレス集合 R は, 送金元のアドレスは送金相手が指定するため, 同じアドレスと取引を行なってしまう, という

表 3.2: 4 つのアドレス集合と識別精度の予測

識別手法	集合	アドレスの管理者による識別回避	推定精度が高くなる取引例
永田ら [5]	S	∘	特定の取引相手に対して頻繁に送金を行う
提案手法 1.	R	×	特定の取引相手から頻繁に送金される
Meiklejohn ら [2]	I	∘	ユーザが複数のアドレスを管理し, 送金を行う
提案手法 2.	O	×	交換所で管理されたアドレスへの送金と受け取り

仮説に基づいた識別手法である. そのため, 取引相手 (送金元アドレス) が同じアドレスを再利用している場合は, 識別の精度が高くなると考えられる. 提案手法 2. の出力アドレス集合 O を用いた識別精度は, 交換所で管理されているユーザのアドレスへの取引回数に依存することがある. 交換所が管理しているユーザのアドレスは, 一定の時間が経過するごとに,1 つのアドレスへ送金が行われる. これは, ユーザが所有する Bitcoin を 1 つのアドレスに集約することで, 交換所が保管する Bitcoin の管理を行いやすくすることが目的である.

永田らと Meiklejohn らのアドレス集合 S, I は, 識別されるアドレスが送金を行う取引のみ着目している. 送金時に利用するアドレスはユーザが指定するため, 識別を回避する目的で別のアドレスを使用することが可能となる. 一方で, 本稿で新たに提案した 2 つのアドレス集合 R, O は, 識別されるアドレスが受け取りを行う取引に着目している. 送金元のアドレスは取引相手のユーザが選択するため, 意図的に識別を回避することができない. 従って, 提案手法 R, O は先行研究の識別手法 S, I よりもノイズ (識別を意図的に回避する取引) が小さく, 識別精度が高いことが予測される.

第4章　アドレス識別実験

4.1　実験目的

2つの提案方式の識別精度を明らかにすることを目的とする. 精度は次の条件に大きく依存すると考えられる.

(1) アドレスあたりの取引回数 n

(2) Bitcoin の利用目的

ここで,n は識別対象のアドレスが観測期間内に送金, 受け取りを行った取引回数とする. 例えば,(1) については, アドレス a_1 が 21 回の送金, 受け取りを伴う取引を行った時, アドレス a_1 の取引回数は $n = 21$ となる. ただし, 多くのアドレスの取引回数をまとめて評価する時は,10 刻みなどに量子化して用いる. (2) については, 交換所やマイニングプール業者の様に同じアドレスで繰り返し送金, 受け取りを必要とする場合と投資目的のエンドユーザとでは, 取引の振る舞いが大きく変わることを想定している. 本研究では, 代表的なユースケースとして, 掲示板,ATM, 交換所, マイニングプール,Dark web サービスの 5 種類を用いる. 5 種類のサービスは実験により独自に収集できるアドレスかつ Bitcoin の代表的な利用方法である.

これらの条件を調整して, 先行研究の 2 方式 (入力アドレス集合 I[2] と宛先アドレス集合 S[5]) と提案方式を比較するために, 次の実験を行う.

実験 1 取引回数による識別精度 ((1) の評価)

実験 2 利用目的による識別精度 ((2) の評価)

実験 3 取引回数と利用目的を考慮した識別精度 ((1),(2) の評価)

4.2　アドレスの利用目的

識別に利用するアドレスはオンライン上に掲載されているアドレスや ATM 実機を用いて取引を行っているアドレスについて, 次の 5 種類の利用目的別にアドレスを収集した.

4.2.1　BBS Bitcointalk

[1]bitcointalk profile (`https://bitcointalk.org/index.php?action=profile;u=907855`)

図 4.1: アドレスを公開している Bitcointalk プロフィールページの例[1]

Bitcointalk[2]は Bitcoin を主にした暗号資産に関する情報を交換する代表的な掲示板サイト (BBS:bulletin board system) である．Bitcointalk では, アカウントを登録しているユーザが図 4.1 で示すようなプロフィールページを作成し, 自身のアドレスを公開していることがある．アドレスをプロフィールページに記載する理由として自身への寄付を受け取ることが考えられる．Bitcointalk のプロフィールページに記載されているアドレスはユーザ自身が管理を行っているアドレスと考え,Bitcointalk ユーザのアドレスとみなす．

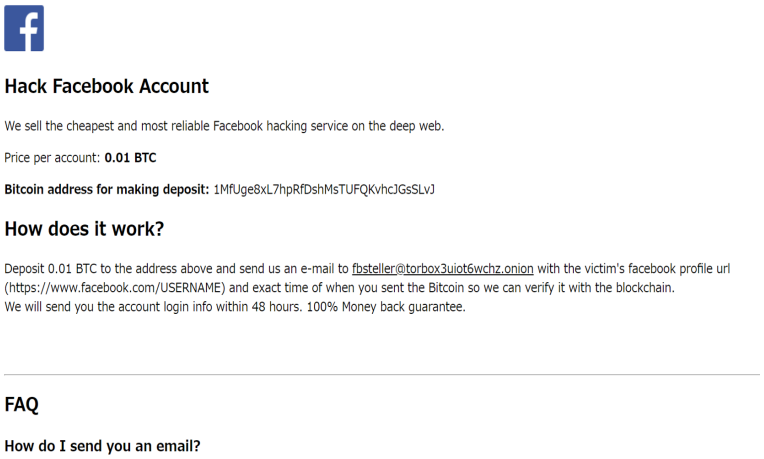
4.2.2 Bitcoin ATM

Bitcoin ATM[4]は Bitcoin を預貯金することができるオフラインのサービスである．Bitcoin ATM の実機の写真を図 4.2 に示す．ユーザは Bitcoin アドレスの公開鍵情報 (QR コード) を ATM に入力し, 任意の額の現金で入金すること, で Bitcoin ATM 事業者のアドレスからユーザのアドレスへ等価の Bitcoin が送金される．

世界中に設置されている Bitcoin ATM のうち, 設置台数が 600 台を超えたカナダにある事業者の実機 (General Bytes[5]社 BATMTwo モデル,pluto[6]社) のアドレスを収集する．ATM の実機に割り当てられた固定のアドレスは ATM 業者のアドレス,ATM のアドレスから送金されているアドレスを ATM ユーザのアドレスと定義する．

[2]bitcointalk (`https://bitcointalk.org/`)
[3]撮影者:井垣 秀星, 2019 年 8 月撮影
[4]Coin ATM Radar Bitcoin ATM Map (`https://coinatmradar.com/`)
[5]General Bytes (`https://www.generalbytes.com/en/`)
[6]pluto (`https://plutobtm.com/`)

図 4.2: カナダ トロントに設置されている Bitcoin ATM 実機の例[3]



Hack Facebook Account

We sell the cheapest and most reliable Facebook hacking service on the deep web.

Price per account: **0.01 BTC**

Bitcoin address for making deposit: 1MfUge8xL7hpRfDshMsTUFQKvhcJGsSLvJ

How does it work?

Deposit 0.01 BTC to the address above and send us an e-mail to fbsteller@torbox3uiot6wchz.onion with the victim's facebook profile url (https://www.facebook.com/USERNAME) and exact time of when you sent the Bitcoin so we can verify it with the blockchain.
We will send you the account login info within 48 hours. 100% Money back guarantee.

FAQ

How do I send you an email?

図 4.3: Dark web 上のアドレスを収集したサイト例[7]

4.2.3 Dark web

Dark web は匿名通信ネットワーク Tor 対応で配信されているウェブページの総称である．Dark web を利用する際には特殊なブラウザを用いることで送信元を匿名のままでアクセスする．

Bitcoin アドレスを収集した Dark web サイトの例を図 4.3 に示す．このサイトでは,Bitcoin アドレス “1MfUge8xL7hpRfDshMsTUFQKvhcJGsSLvJ” 宛に送金することで Facebook のアカウントのハッキングを依頼する不正サービスの例である．アドレスは Tor ブラウザを利用してアクセス可能な.onion ドメインの web サイトより収集する．

クレジットカードの売買や SNS ハッキングサービスなど違法性の高いサイトよりサービス利用時に支払い先のアドレスとして指定された Bitcoin アドレスを Dark web 業者と定義する．また,Dark web のサイト運営者が Dark web のウェブページ上で掲載しているプロモーション用のアドレスを Dark

[7]Hack Facebook Account (`http://r3cnefrmwctd6gb2.onion`)

表 4.1: アドレスを観測した交換所一覧

交換所名	アドレス数	URL
ANXPRO	4	https://anxpro.com/
BitBay	13	https://bitbay.net/en/exchange
Bitstamp	40	https://www.bitstamp.net/
Bittrex	116	https://global.bittrex.com/
CoinHako	2	https://www.coinhako.com/
happyCOINS	1	https://www.happycoins.com/en
Hashnest	199	https://www.hashnest.com/
HitBTC	89	https://hitbtc.com/
Kraken	26	https://www.kraken.com/ja-jp/
Mercado Bitcoin	130	https://www.mercadobitcoin.com.br/
OKCoin	1	https://www.okcoin.com/
Poloniex	110	https://poloniex.com/
YoBit	281	https://yobit.net/en/

web 利用者と定義する.

4.2.4 Exchange

Exchange(交換所) はユーザの所有する Bitcoin を他の暗号資産や現金と交換するサービスである. Bitcoin ATM とは異なり, ユーザは交換所へ登録を行うことでオンライン上で Bitcoin の売買が可能となる. WalletExploer[8]に記載されている Exchanges 一覧リストより取得した主要な交換所の名称と取得したアドレス数を表 4.1 に示す. 交換所と取引しているアドレスを交換所ユーザのアドレスと定義する.

4.2.5 Mining Pool

Mining Pool は多数のマイナーが協力し Bitcoin の取引情報をまとめたブロックに対して取引の検証を行い, 報酬を得るための仕組みである. 報酬を得るためには膨大な計算量が必要とされており, 表 4.2 に示した複数の Mining Pool が報酬を受け取るために競争を行っている. Mining Pool のように多数のマイナーの計算資源を利用したマイニングが主流になっているため, 個人ではマイニング競争に勝ち, 対応する報酬を受け取ることは難しいと考えられている. マイニング報酬を受け取ったことがあるアドレスを Mining Pool が管理しているアドレスと定義する.

[8]WalletExplorer.com (https://www.walletexplorer.com/)

表 4.2: 著名な Mining Pool 一覧

マイニングプール名	マイニングレート [%] (2020/12)
F2Pool	20.32
Poolin	12.24
Binance Pool	11.50
BTC.com	10.67
AntPool	10.11
Huobi.pool	9.00
ViaBTC	8.07
Other	5.48
58COIN&1THash	5.10
Lubian.com	3.80
SlushPool	3.71

Pool Stats - BTC.com (`https://btc.com/stats/pool`)

4.3 実験 1. 取引回数に基づくアドレス識別

実験 1. では, アドレスあたりの取引回数 n に着目し,4 つのアドレス集合の識別精度を評価する. $D = 10$[年] となる長期間継続して利用されているアドレスの識別手法を以下に述べる.

(i) 10 年間 ($D = 10$) の Bitcointalk アドレスを対象とする.

(ii) Bitcointalk アドレスのうち取引回数が 2 回以上,100 回以下となるアドレスを識別対象のアドレスとして使用する.

(iii) 識別に使用するアドレスを取引回数ごとに 100 個のアドレスを 100 回層別サンプリングする.

(iv) 提案手法のアドレス集合 R, O と従来手法のアドレス集合 S, I を用いてアドレスを識別する.

4.4 実験 2. 利用目的に基づくアドレス識別

実験 2. では,Bitcoin の利用目的に着目し,4 つのアドレス集合の識別精度を評価する. $D = 0.5$ となる利用目的を考慮したアドレスの識別手法を以下に述べる.

(i) 半年間 ($D = 0.5$) の 5 種類の利用目的 (Bitcointalk,Bitcoin ATM, Dark web, Exchange, Mining Pool) に分類されたアドレスを対象とする.

(ii) 5 種類のアドレスのうち引回数が 2 回以上となるアドレスを識別対象のアドレスとして使用する.

(iii) 識別に使用する 5 種類の利用目的ごとに 30 個のアドレスを 100 回層別サンプリングする.

(iv) 提案手法と従来手法を用いてアドレスを識別する.

表 4.3: 収集したアドレスデータ

	利用目的	総アドレス数	総取引数	収集期間 D
実験 (1)	Bitcointalk BBS	44,067	3,139,677	2009/1/4 — 2019/11/18 (10 年間)
実験 (2)	Bitcointalk BBS	1,968	28,832	2019/4/1 — 9/30 (半年)
	Bitcoin ATM	404	26,843	
	Dark web	82	35,048	
	Exchange	680	33,252	
	Mining Pool	96	24,449	

4.5 実験 3. 取引回数と利用目的を考慮した識別精度

実験 3. では, 取引回数と利用目的を考慮した識別精度に着目し,4 つのアドレス集合の識別精度を評価する. $D = 10$[年] となる長期間継続して利用されているアドレスを用いた識別手法を以下に述べる.

(i) 10 年間 ($D = 10$) の Bitcointalk アドレスを対象とする.

(ii) Bitcointalk アドレスのうち, 交換所で利用されているアドレスと Bitcointalk のみ使用されているアドレスに分ける.

(iii) Bitcointalk と Exchange アドレスのうち取引回数が 2 回以上となるアドレスを識別対象のアドレスとして使用する.

(iv) 識別に使用するアドレスを取引回数ごとに 100 個のアドレスを 100 回層別サンプリングする.

(v) 提案手法のアドレス集合 R, O と従来手法のアドレス集合 S, I を用いてアドレスを識別する.

4.6 データ収集

本研究で取得したアドレスと取引の数を表 4.3 に示す. 5 種類の利用目的に基づく Bitcoin アドレスの取引記録は Blockchain Explorer[9]の API を用いて収集した. 本実験では, 収集したアドレスのうち 2 回以上取引を行っていたアドレスを利用する. 実験目的に応じて次の 2 つのデータに分割している.

実験 (1) 収集期間は 10 年間 (観測期間 $D = 10$)

実験 (2) 収集期間は半年間 ($D = 0.5$)

実験 (1) で収集したアドレス数と取引回数を表 4.4, 図 4.4 に示す. $n = 100$ までの取引回数では, 識別可能なアドレスが 100 個以上あり,$n = 100$ を超えるアドレスがほとんど存在しない. 従って, 実験 1. では $n = 10$ から $n = 100$ となる 23,541 個のアドレスを対象とする.

実験 (2) で収集したアドレス数を表 4.5 に示す. 5 種類の利用目的ごとに収集した 1,358 個のアドレスを対象とする.

[9]Blockchain Explorer (`https://www.blockchain.com/ja/explorer`)

表 4.4: 長期間継続して利用された Bitcointalk アドレスと取引回数

取引回数 n	アドレス数	サンプリング数
10	12,493	100
20	4,948	100
30	2,535	100
40	1,408	100
50	842	100
60	499	100
70	335	100
80	211	100
90	153	100
100	117	100
合計	23,541	1,000

表 4.5: 5 種類の利用目的の識別アドレス数

利用目的	アドレス数	サンプリング数
Bitcointalk BBS	844	30
Bitcoin ATM	106	30
Dark web	49	30
Exchange	274	30
Mining Pool	85	30
合計	1,358	150

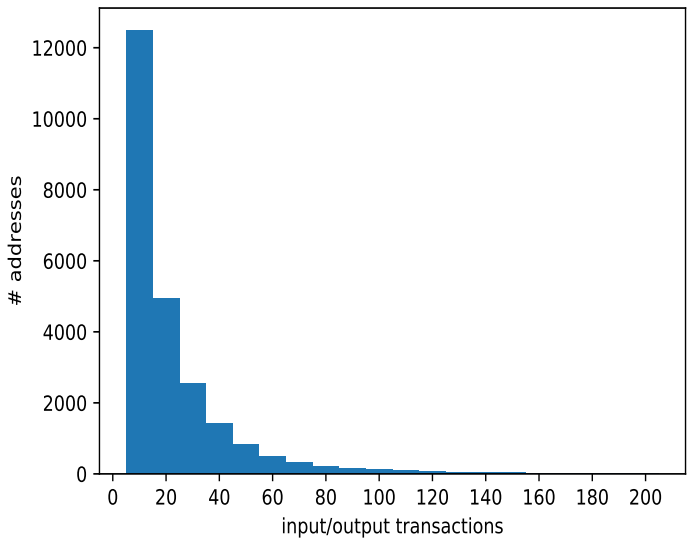


図 4.4: $D = 10$ のアドレスと取引回数 n

表 4.6: 取引回数と平均アドレス識別個数

識別手法	集合	取引回数 n										合計	平均
		10	20	30	40	50	60	70	80	90	100		
永田ら [5]	S	39.2	45.5	43.0	44.6	42.2	38.6	45.4	42.4	43.7	41.8	426	42.6
提案手法 1.	R	54.6	43.4	36.8	35.7	39.6	41.9	48.1	55.0	59.2	62.3	477	47.7
Meiklejohn ら [2]	I	49.8	48.6	46.7	48.7	49.7	54.4	52.5	60.9	64.1	69.1	545	54.5
提案手法 2.	O	60.4	45.3	37.1	39.5	43.9	49.7	54.7	66.8	71.2	78.6	547	54.7
	平均	51.0	45.7	40.9	42.1	43.9	46.1	50.2	56.3	59.5	62.9	-	-

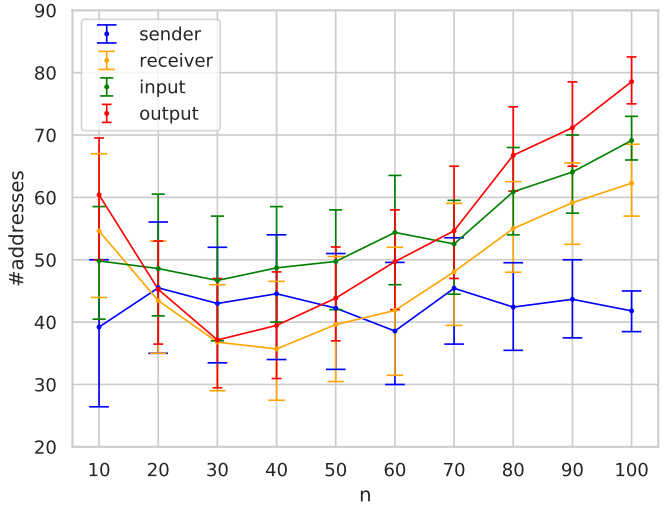


図 4.5: 取引回数 n についての平均アドレス識別個数

4.7 実験結果 1.

識別に成功した平均アドレス個数の結果を表 4.6 と図 4.5 に示す. 表 4.6 では 100 回の試行のうち, 識別に成功したアドレス数の平均値を示す. 図 4.5 では 4 つのアドレス集合の識別精度について,95% の信頼区間を表すエラーバーと共に示す.

4 つのアドレス集合のうち, 出力アドレス集合 O の 547 個が最も識別に成功したアドレス数が多い. この集合 O の Jaccard 係数の分布を図 4.6 の箱ひげ図に示す. $n = 10$ のとき,Jaccard 係数は 0.15 から 0.20 の値に分布しており, 分散が大きい. この分散は取引回数の増加に伴い, 小さくなっている.

$D = 10$ の長期間観測されたアドレスのうち, アドレス集合 R, I, O については取引回数 $n = 100$ の時に, 識別に成功したアドレス数が最も多い. 取引回数 $n = 100$ の Jaccard 係数の分布を図 4.7 に示す. ここで, 正しく識別したアドレスを True, 誤ったものを False で表す. 4 つの集合のうち識別に成功した sender_True,receiver_True,input_True,output_True の Jaccard 係数は 0.0 から 0.2 の範囲に分布しており, アドレス集合間に差異は見られない. $n = 100$ 以下の取引回数とアドレス識別に成功した数で大きな差異が見られなかった集合 S について, アドレス識別に失敗した sender_False の Jaccard 係数は 0.0 から 1.0 の範囲に分布しており, 分散が大きい. これは, 識別対象のアドレスと Jaccard 係数の値が同じとなる不正解のアドレスが多く存在していることを示している.

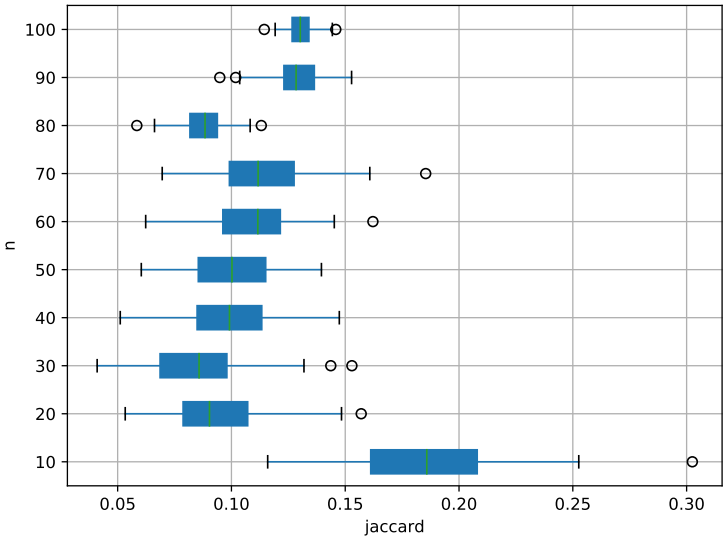


図 4.6: 集合 O における取引回数の Jaccard 係数の分布

表 4.7: 4 つの集合の平均アドレス識別数と標準偏差

識別手法	集合	合計	平均	標準偏差	変動係数
永田ら [5]	S	426	42.6	2.2	0.05
提案手法 1.	R	477	47.7	9.1	0.19
Meiklejohn ら [2]	I	545	54.5	7.2	0.13
提案手法 2.	O	547	54.7	13.4	0.25

アドレス集合 O は,$n = 30$ のときアドレス識別精度は 37.1 と最も低く,$n = 100$ のとき 78.6 まで増加している. 同様に, 集合 R, I の識別精度も $n = 30$ 前後で最小となり,$n = 100$ で最大となる. これは, 永田らの報告 [5] による取引回数とアドレスの識別率には相関がない結果と異なっている.

そこで, より精査して見るために, 表 4.7 に,4 つのアドレス集合の取引数と識別に成功したアドレス数の標準偏差を示す. 集合 S の標準偏差は 2.2 であり, 変動係数は 0.05 を示した. $n = 30$ 以上で識別精度が増加する集合 R の変動係数は 0.19, 集合 I は 0.13, 集合 O は 0.25 となり, 集合 S よりも大きな値を示した. よって, 集合 S は集合 R, I, O と比較して識別精度が取引回数に依存しないことがわかる.

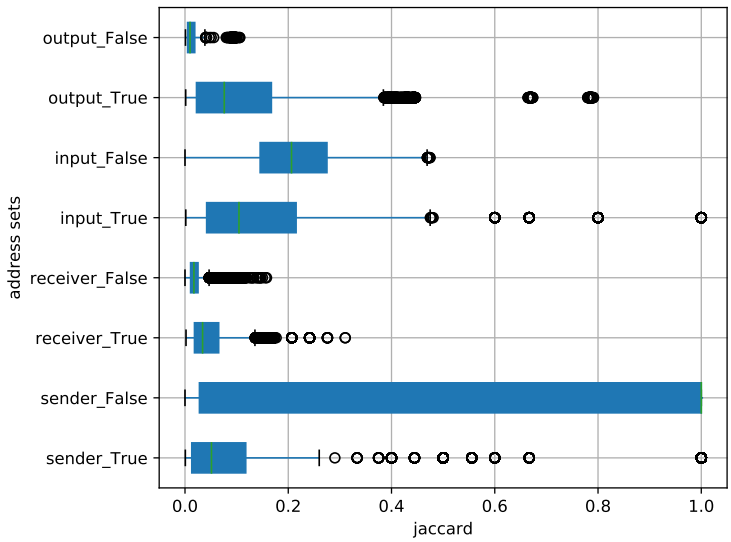


図 4.7: $n = 100$ における集合 S, R, I, O の Jaccard 係数の分布

表 4.8: 5 種類の利用目的の平均アドレス識別個数

識別手法	集合	利用目的					合計	平均
		BBS	ATM	Dark web	Exchange	Mining pool		
永田ら [5]	S	12.8	16.6	23.9	4.1	17.8	75	15.0
提案手法 1.	R	17.2	3.6	22.2	14.7	5.0	63	12.5
Meiklejohn ら [2]	I	17.6	16.5	22.3	12.6	15.0	84	16.8
提案手法 2.	O	19.5	4.3	20.9	22.6	5.0	72	14.5
	平均	16.8	10.2	22.3	13.5	10.7	-	-

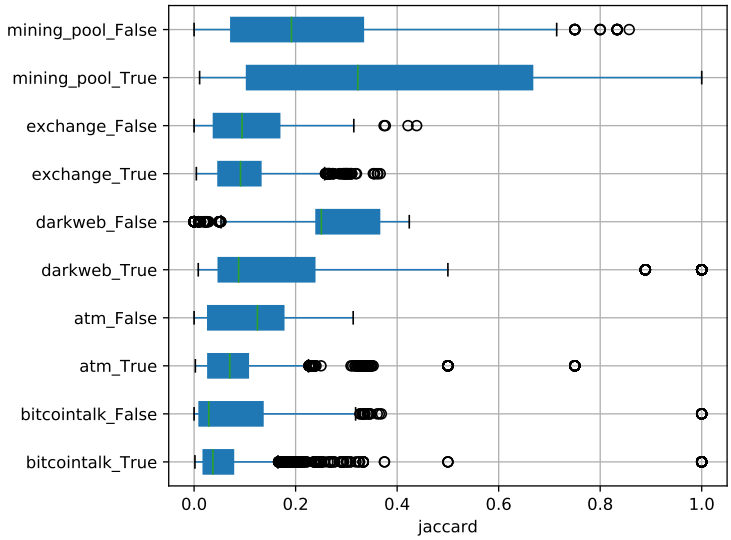


図 4.8: 集合 I における 5 種類の利用目的についての Jaccard 係数の分布

4.8 実験結果 2.

5 種類の利用目的に基づくアドレス識別結果を表 4.8 に示す．識別に成功したアドレス数が最も多い利用目的は Dark web であり,4 つのアドレス集合を用いた識別結果の平均個数は 22.3 個であった．また,4 つのアドレス集合のうち入力アドレス集合 I は最も識別精度が高い．

$D = 0.5$ の利用目的に考慮したアドレスのうち, 入力アドレス集合 I の 84 個が最も識別に成功したアドレス数が多い．この集合 I の Jaccard 係数の分布を図 4.8 に示す．ここで, 正しく識別したアドレスを True, 誤ったものを False で表す．5 種類の利用目的のうち, 最も識別に成功したアドレス数が多い dark web_True の Jaccard 係数は 0.0 から 0.2 の範囲に分布している．識別に失敗した dark web_False の Jaccard 係数は 0.2 から 0.4 の範囲に分布していることから, 識別に成功したアドレスと失敗したアドレスに交わりが少ないことを示している．

表 4.9: $D = 10$ の Bitcointalk アドレスに含まれる交換所アドレス数と取引回数

取引回数 n	Bitcointalk アドレス数				
	追加実験 1.		追加実験 2.		総アドレス数 (表 4.4)
	Bitcointalk のみ	サンプリング数	交換所アドレス	サンプリング数	
10	9,814	100	2,679	100	12,493
20	3,317	100	1,631	100	4,948
30	1,627	100	908	100	2,535
40	911	100	497	100	1,408
50	546	100	296	100	842
60	334	100	165	100	499
70	235	100	-	-	335
80	155	100	-	-	211
合計	16,939	800	6,176	600	-

表 4.10: 交換所アドレスを除いた取引回数と平均アドレス識別個数

集合	取引回数 n								合計	平均	標準偏差
	10	20	30	40	50	60	70	80			
S	43.2	53.5	53.4	54.0	50.4	46.8	53.6	47.4	402	50.3	3.8
R	56.9	47.1	43.2	43.4	46.3	47.8	53.6	61.4	400	50.0	6.2
I	51.2	53.4	52.5	52.6	51.2	56.3	55.4	63.7	436	54.5	3.9
O	62.3	51.5	48.3	50.9	55.4	61.2	64.2	74.9	469	58.6	8.2
平均	53.4	51.4	49.3	50.2	50.8	53.0	56.7	61.9	-	-	-

4.8.1 実験結果 3.

本実験で使用するアドレスと取引回数を表 4.9 に示す.

Bitcointalk のみで利用されていたアドレス数について, サンプリング数が 100 個を満たす十分なアドレス数は $n = 80$ までであった. そのため, 識別に使用可能なアドレス数が 100 個以上存在する $n = 80$ までのアドレスを識別対象としている.

また, 交換所アドレスの識別に使用可能なアドレス数について, サンプリング数が 100 個を満たす十分なアドレス数は $n = 60$ までであった. 識別に使用可能なアドレス数が 100 個以上存在する $n = 60$ までのアドレスを識別対象としている.

Bitcointalk アドレスのみを用いたアドレス識別の結果を表 4.10 と図 4.9 に示す. 表 4.10 の識別に成功したアドレス数の平均値は, 表 4.6 の,4 つのアドレス集合の平均と同じ, または, 増加している.

交換所アドレスを用いたアドレス識別の結果を表 4.11 と図 4.10 に示す. 最も識別精度が高い集合 R は, 平均 28.2 個のアドレスを識別している. $n = 10$ から $n = 60$ までの識別に成功したアドレス数は表 4.10 に示した Bitcointalk の識別結果と比較し, いずれも低い値となっている.

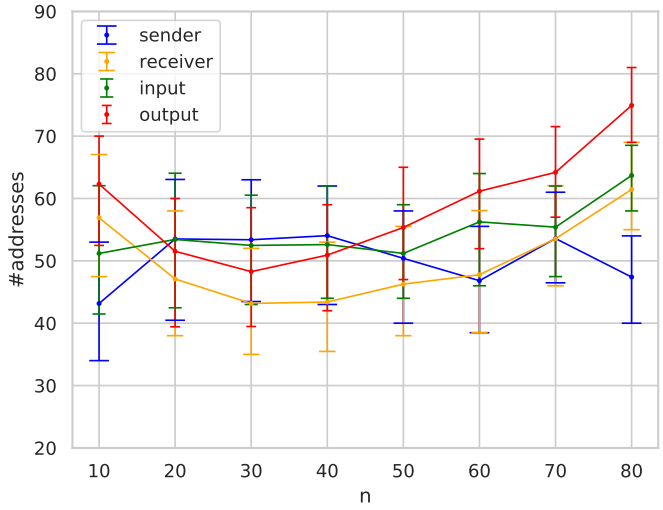


図 4.9: $D = 10$ の交換所アドレスを除いた Bitcointalk アドレスの識別結果

表 4.11: $D = 10$ の交換所アドレスの取引回数と平均アドレス識別個数

集合	取引回数 n						合計	平均	標準偏差
	10	20	30	40	50	60			
S	14.8	15.0	12.4	14.0	12.7	12.3	81	13.5	1.1
R	49.6	30.0	20.8	18.8	25.1	24.6	169	28.2	10.2
I	22.6	22.4	19.5	22.1	24.8	25.4	137	22.8	1.9
O	42.0	24.6	17.4	17.0	22.1	25.4	148	24.8	8.4
平均	32.2	23.0	17.5	18.0	21.1	21.9	-	-	-

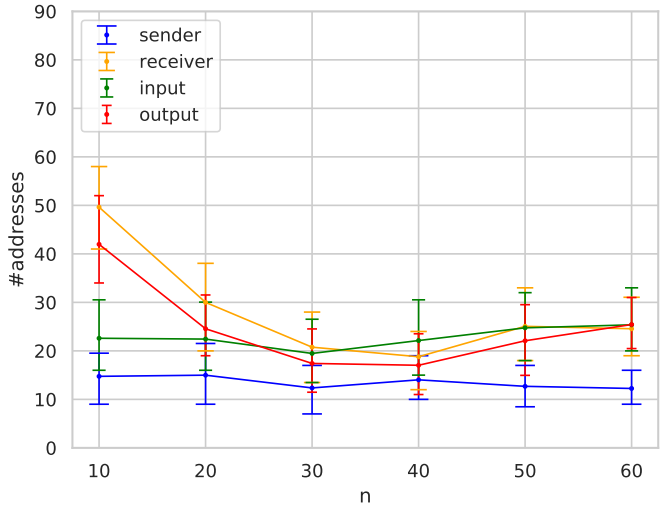


図 4.10: $D = 10$ の交換所アドレスかつ Bitcointalk アドレスの識別結果

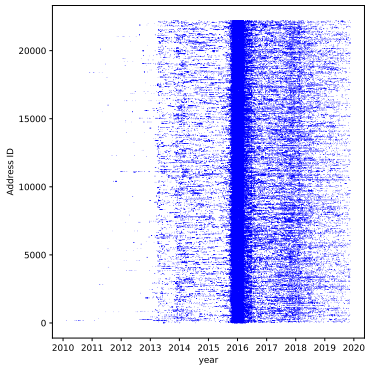


図 4.11: $D = 10$ 年の間におけるアドレス取引分布

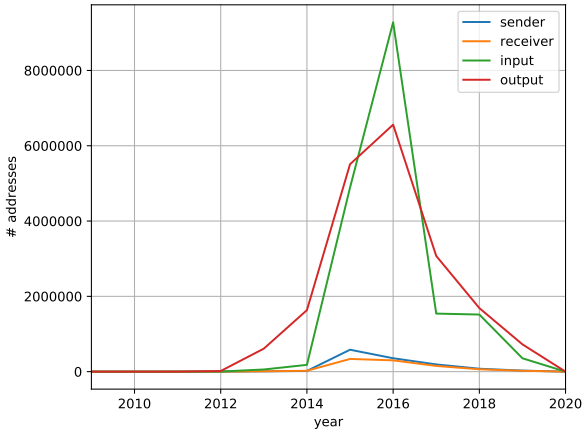


図 4.12: $D = 10$ 年の間アドレス集合の大きさの推移

4.9 評価と考察

4.9.1 $n = 30$ 付近での識別率の低下について

図 4.5 より,3 つのアドレス集合 R,I,O は $n = 30$ 付近において識別率のピークがあり,$n > 30$ では再び増加している. この非単調な識別率の原因として, 次が考えられる.

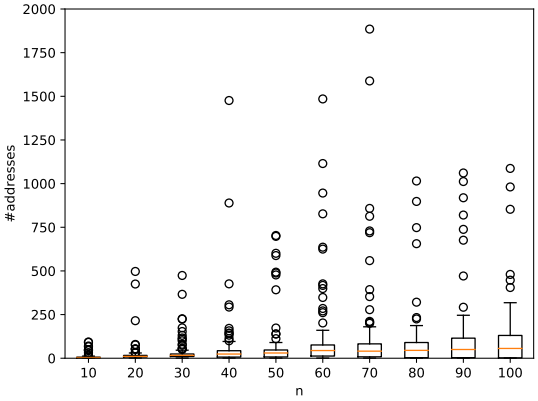
(1) $n < 30$ のアドレスに, アドレスが毎回更新される新しいウォレットで使われる割合が多いため.

(2) $n < 30$ のアドレスの $Input, Output$ に指定されたアドレス数 (学習量) が少ないため.

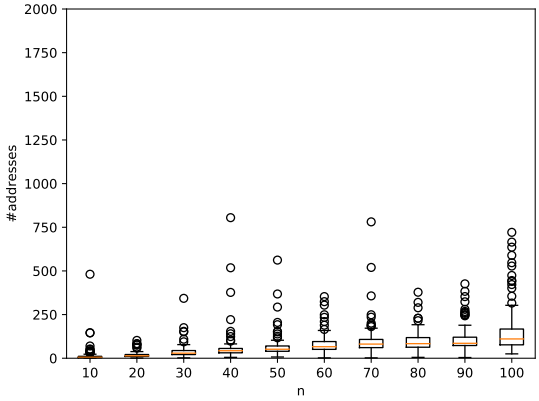
(3) $n < 30$ となるアドレスに, 特定の利用目的のものが偏っているため.

そこで,(1) を調査するために, 各取引数 n におけるアドレスの開始年度を調べた. 仮説が正しければ,$n < 30$ における年度に偏りが観えるはずである.

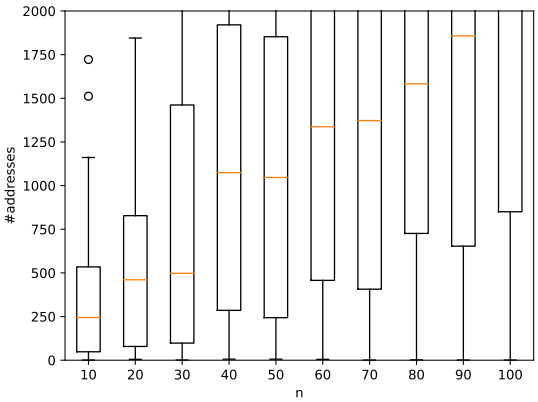
図 4.11 に識別対象のアドレスに関する取引日時の散布図を示す. 2016 年と 2018 年の前後で取引が頻繁に行われていたことがわかる. しかし, 取引の開始年度における著しい偏りはなく,(1) の仮説が原因とは考えにくい.



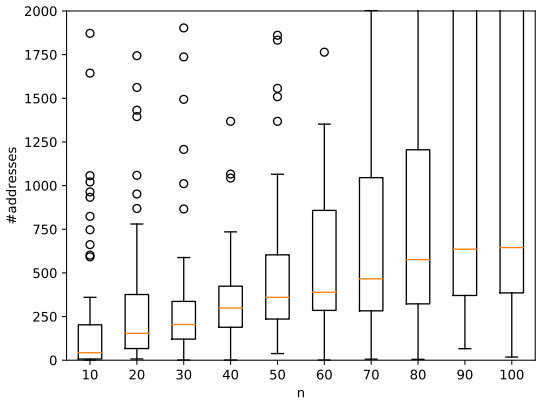
アドレス集合 S



アドレス集合 R



アドレス集合 I



アドレス集合 O

図 4.13: 4 つのアドレス集合における取引回数 n と学習アドレス数の分布

次に (2) を調査するため, 識別されるアドレスの集合の大きさと学習量を調査した. なぜならば, 識別されるアドレスの宛先, 送金元のアドレス数が多いほど過去に取引を行ったアドレス数 (学習量) が大きくなり, 取引回数が少ない場合でも十分他のアドレスと識別可能となる.

取引された期間と学習量の関係を分析するため,$D = 10$ の期間における,4 つのアドレス集合の大きさの推移を図 4.12 に示す. 集合の大きさは取引が行われた年で量子化している. 実験 1. の結果では, 集合 O が識別に成功したアドレス数が最も高い結果であったが,10 年間の推移では 2016 年の集合 I が最も大きい. また, 集合 S,R の大きさは同じ振る舞いとなっているが, 集合 R は取引回数 $n = 40$ より識別率が高くなっている.

アドレスの取引回数と学習量の変化の関係を分析するため,4 つの集合の取引回数 n と学習アドレス数の分布を図 4.13 に示す. 4 つの集合は取引回数が増加するごとに学習に利用したアドレス数は増加し, 分布は大きくなっていた.

アドレスの集合の大きさや学習量の大きさからは (2) の仮説が原因とは考えにくい.

次に (3) を調べるために, 利用目的の中の交換所のアドレスに注目する. 交換所のアドレスは取引パターンも固有で識別率も高いために, 全体の識別率を支配していると考えたためである. 4 つのアド

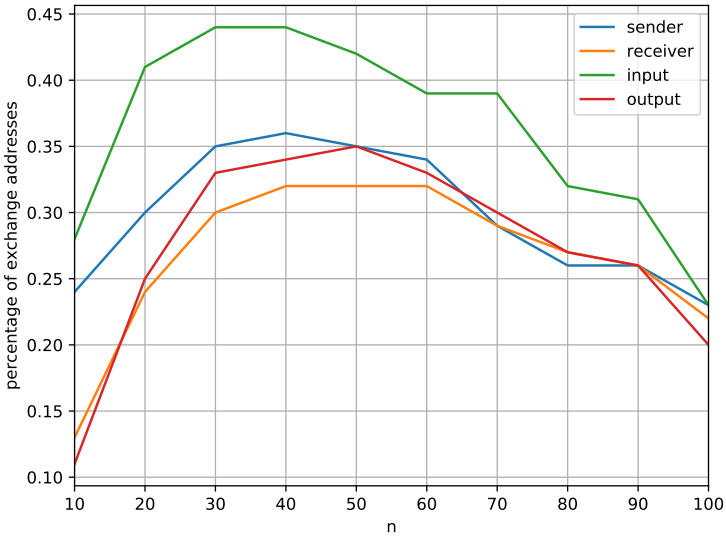


図 4.14: 識別成功アドレス中の交換所アドレスの割合

レス集合で識別に成功したアドレスのうち, 交換所のアドレスが含まれている割合を図 4.14 に示す. 4 つのアドレス集合では $n = 30$ から $n = 50$ 付近で交換所アドレスの割合が最も高い値を示している. 図 4.5 と図 4.14 を比較すると $n = 30$ 付近でピークとなること,$n = 10$ から $n = 100$ までの割合が類似していることから, 交換所のアドレスが識別率に影響を与えていると結論づける.

4.9.2 Dark web アドレスが最も識別率が高い原因について

表 4.8 の結果より, 最も識別率が高い利用目的は Dark web で利用されたアドレスであった. この要因の一つに Dark web の運用形態が原因であると考えられる. Dark web では違法商品の売買など法的に問題のあるサービスで利用されているので, 頻繁にサイトの公開と閉鎖が繰り返されており, 取引に利用された Bitcoin アドレスの寿命も短い. 従って, これが識別率に影響を与えていると考える.

4.9.3 提案手法の精度が高いことの仮説検定

提案方式 R, O のアドレス識別精度が従来手法 S, I より高いことを確かめるために平均値の t 検定 [24] を行う. 検定の結果を表 4.12 に示す. 検定には $n = 100$ における 4 つの方式の精度を比較している.

提案方式 2(出力アドレス O) は先行研究 S, I のいずれに対しても識別精度が高く, その差は十分に大きく,2 つのアドレス集合の識別率は一致しているという帰無仮説の p 値が 0.05 未満であり, 統計的に有意であることが示された.

表 4.12: 4 種類のアドレス集合と平均値の検定結果 ($n = 100$)

		平均値の差	統計量 t	p 値	
提案方式 1	R, S	20.5	57.0	2.2×10^{-16}	***
	R, I	−6.8	-	-	
	R, O	−16.3	-	-	
提案方式 2	O, S	36.8	135.1	2.2×10^{-16}	***
	O, I	9.5	34.9	2.2×10^{-16}	***
	O, R	16.3	45.5	2.2×10^{-16}	***

$***$：$P < 0.05$

4.9.4 アドレス識別の対策手法の検討

本稿で提案した集合 O を特徴量としたアドレス識別手法への対策は, 受け取り専用のアドレスを 1 つ定めて利用することである.

アドレス識別の問題は,“あるユーザが管理している複数のアドレスを与えて, そのユーザの管理する他の未知のアドレスを他人のアドレスから識別する問題” と定めている. 従って, 全ての取引で *Output* フィールドに自身のアドレスが 1 つのみ利用されている場合, 集合 O を用いた識別手法では識別することができない.

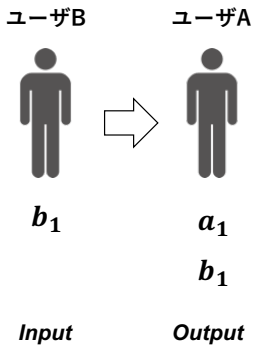
図 4.15 に示した例を元に, 自身のアドレス a_1 を受け取り専用アドレスと定めた場合を考える. 取引 1 は, ユーザ B のアドレス b_1 から自身のアドレス a_1 へ送金が行われる取引である. 取引の *Output* フィールドには,a_1 とユーザ B がお釣りを受け取るアドレス b_1 が指定される. このとき, ユーザ A のアドレスは *Output* フィールドに 1 つしか存在しないため, ユーザ A が管理する未知のアドレスが識別されることはない.

取引 2 は, ユーザ B がユーザ A に送金する際に, お釣りを受け取るアドレスに異なるアドレス b_2 を指定している. こちらも,*Output* フィールドにはユーザ A が管理している他のアドレスは指定されないため, 未知のアドレスは識別されない.

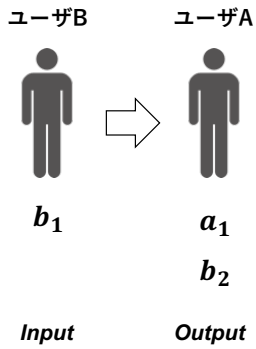
取引 3 は, ユーザ B が自分 (ユーザ A) とその他のユーザ C に対して,1 度に送金を行う取引である. 取引の *Output* フィールドには,a_1 とユーザ C のアドレス c_1, ユーザ B がお釣りを受け取るアドレス b_2 が指定される. 複数のユーザへ送金を行う取引においても, ユーザ A のアドレスは *Output* フィールドに 1 つだけ指定され, 未知のアドレスは識別されない.

また, ユーザ A の受け取り専用アドレス a_1 は, 別の取引時に送金を行うアドレスとして直接利用するのではなく, 自信が管理している他アドレス (a_2, a_3 など) に一度送金することで, アドレス識別は困難となる.

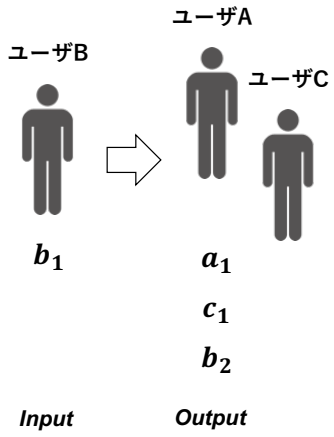
以上が, 本稿で提案した集合 O を特徴量としたアドレス識別手法の対策例となる.



取引 1:取引相手 (ユーザ B) が
お釣りアドレス b_1 を再利用



取引 2:取引相手 (ユーザ B) が
お釣りに別アドレス b_2 を利用



取引 3:取引相手 (ユーザ B) が
自分以外のユーザ C にも送金

図 4.15: 集合 O を利用したアドレス識別対策の取引例

第5章　アドレス利用目的推定実験

5.1　実験目的

Bitcoin アドレスの利用目的に注目し, あるアドレスの取引情報の特徴量から利用目的を推測可能であるか明らかにすることを目的とする. 利用目的は次の条件に大きく依存すると考えられる.

(1) Bitcoin を利用した商品やサービスの提供を行う業者のアドレス

(2) Bitcoin を利用した商品やサービスを享受するユーザのアドレス

本研究は,Bitcoin アドレスの管理者を業者とユーザのアドレスの 2 つに分ける. なぜならば,Bitcoin を使用したサービスを提供する事業者の取引と投機目的で Bitcoin を購入しているユーザの取引には取引方法に違いがあると考えたからである. (1) は 4.2 節で示した Bitcoin ATM,Dark web,Mining Pool のアドレスを対象とする. (2) は 4.2 節で示した Bitcointalk,Bitcoin ATM,Dark web,Exchange のアドレスを対象とする.

これらの条件を考慮して, アドレスの利用目的が推定可能か評価するために, 次の実験を行う.

実験 i アドレスの利用目的に関する取引情報の分析

実験 ii 決定木学習を用いた Bitcoin アドレスの利用目的の推定実験

5.2　データ収集

利用目的別に収集した Bitcoin アドレスと取引数を表 5.1 に示す. アドレスの利用目的は管理者を業者とユーザに区別した計 7 種類のアドレスを対象とする. アドレスは収集期間に 1 回以上取引を行った 4,049 個を収集した. 表 5.1 の取引数は利用目的別に集計した. 例えば,Bitcoin ATM の業者とユー

表 5.1: 利用目的別に収集した Bitcoin アドレス

利用目的	業者アドレス数	ユーザアドレス数	取引数	収集期間
Bitcointalk BBS	-	2,391	29,638	2019/4/1 — 9/30 (半年間)
Bitcoin ATM	3	452	26,849	
Dark web	26	67	35,076	
Exchange	-	1,012	33,351	
Mining Pool	98	-	24,876	
合計	127	3,922	149,790	-

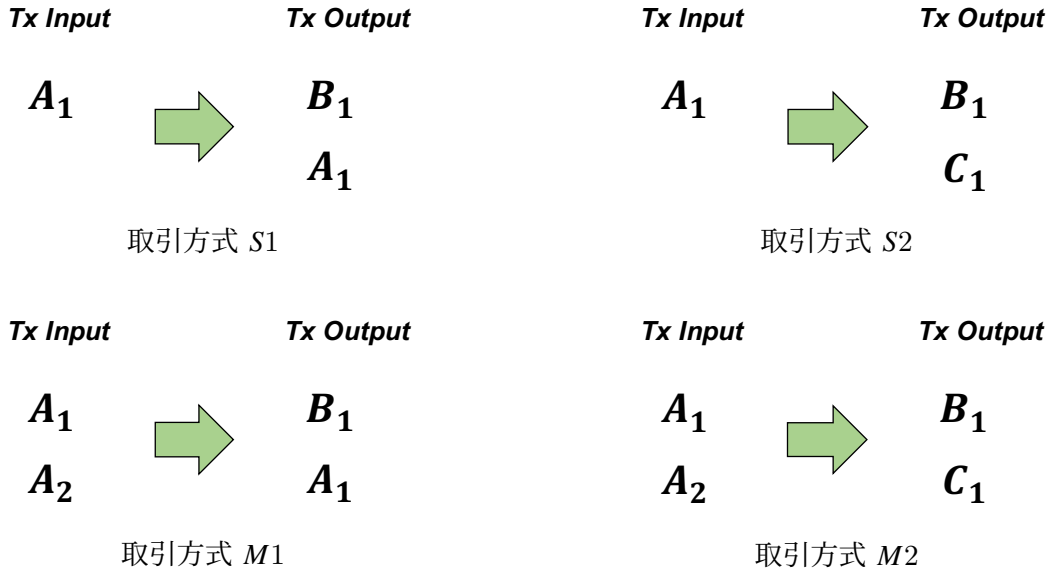


図 5.1: 送金アドレス数とおつりを受け取るアドレスに着目した取引構造の例

表 5.2: 4 つの取引方式の定義

取引方式	入力アドレス数	おつりアドレス	取引例
$S1$	1	入力アドレスを再利用	基本取引 ATM の「預金」取引
$S2$	1	なし (全額を送金)	一部ウォレットアプリによる取引
$M1$	2 以上	入力アドレスを再利用	交換所のユーザのアドレスを 交換所アドレスへまとめる取引
$M2$	2 以上	なし (全額を送金)	Mining Pool 業者からの マイナーへ報酬の支払い

ザとの間で行われた取引は 1 件の取引として集計している．また, 収集したアドレスからは複数の種類に重複しているものを取り除いている．例えば,Bitcointalk と Exchange(交換所) など 2 種類以上の利用目的で利用されたアドレスは含まれていない[1].

5.2.1 取引構造に基づく 4 つの取引方式の定義

7 種類のアドレスの利用目的を推定する特徴量として, 取引で利用される入力・出力アドレスの数とおつりを受け取るアドレスに着目した．なぜならば, 取引で利用されるアドレスは送金者が利用するウォレットアプリケーションに依存すると考えられるからである．例えば, モバイルウォレットの BitPay[2]では,Bitcoin の送金時におつりを受け取るアドレスを新たに自動で生成する．モバイルウォレットなど, 個人の端末で管理されているアドレスはユーザのアドレスである可能性が高いと考えた．これに対して, 業者によって管理されているアドレスは Bitcoin ATM など入金に対して自動で送金処

[1]業績 CSEC88,NBiS2020 の論文において “7 種のアドレスに重複はない” と報告した．
その後,Bitcointalk ユーザと Exchange ユーザに交わりがあるアドレスが 1 つ確認されたことをここに示す．
なお, 表 5.1 のアドレス数について, 実験結果に基づき訂正は行なっていない．

[2]BitPay - Buy Crypto (`https://apps.apple.com/jp/app/bitpay-buy-crypto/id1149581638`)

Listing 5.1: sourcecode

```
1    clf = tree.DecisionTreeClassifier(max_depth=5, min_weight_fraction_leaf=0.1)
```

図 5.2: 決定木学習のパラメータ

理を行うものや,Mining Pool など企業や組織が運営している複数の管理者が送金権限を持つアドレスである. システムで送金処理が管理されているアドレスは, 送金時のおつりを受け取るアドレスを使い回し, 長期間利用されるアドレスである可能性が高いと考えた.

そこで, 取引構造に基づく特徴量を図 5.1 と表 5.2 にて定義する. 任意の取引において, 入力アドレスの個数とおつりを受け取るアドレスを図 5.1 に示した 4 つの取引方式に分類する. 取引方式 $S1$ と $S2$ は, 入力アドレス数が単一である取引であり, 複数ある取引方式は $M1$ と $M2$ とする. 入力アドレスと出力アドレスに同じアドレスが指定されている取引方式を $S1$ と $M1$ とし, それ以外, すなわち入力アドレスと出力アドレスに重複はないものを $S2$ と $M2$ とする. 取引方式 $S1$,$M1$ のように, 入力アドレスに利用されたアドレスが, 再び出力アドレスに指定される場合, 送金者が取引で生じた “おつり” を受け取っていることを意味する. 例えば, 取引方式 $S1$,$M1$ では, アドレス A_1 がおつりを受け取るアドレスとして再利用されている. 取引方式 $S2$,$M2$ では, 入力アドレスに使用されたアドレスは利用されていない. 送金者が入力アドレスとは別のアドレスを利用しておつりを受け取っている場合においても, おつりを受け取っていない (全金額を送金している) とみなす.

5.3 実験 i. アドレスの取引情報の分析

7 種類のアドレス利用目的の推定に使用する特徴量を, 収集したアドレスの取引情報より検討する.

取引数が多いアドレスほど学習に使用する特徴量が多く, 利用目的の推定精度が高くなると考えた. そこで,7 種類の利用目的のアドレスに対して取引数の統計量に基づく分析を実施する. 取引数の統計量は取引回数の平均値, 最小値, 中央値, 最大値, 標準偏差である.

また, 業者とユーザによって管理されているアドレスは, 使用されているウォレットに依存するなど, 取引構造に異なる振る舞いが見られると考えた. そこで, 表 5.2 で定義された 4 つの取引方式について,7 種類のアドレスが含まれる取引数と割合を求める.

5.4 実験 ii. アドレスの利用目的の推定

決定木学習を用いた 7 種類のアドレス利用目的の推定手法を以下に定める.

(i) 7 種類の利用目的のアドレスを対象とする.

(ii) アドレスの取引記録から特徴量を求める.

表 5.3: 取引の特徴量一覧

特徴量	統計量	説明
取引件数	5	取引を行った総回数
送金回数	5	Bitcoin の送金取引を行った総回数
受け取り回数	5	Bitcoin の受け取り取引を行った総回数
取引の入力アドレス数	5	取引時に入力アドレスに使用されたアドレス数
取引の出力アドレス数	5	取引時に出力アドレスに使用されたアドレス数
取引で利用されたアドレス数	1	取引の入力, 出力アドレスに使用されたアドレス数
再利用入力アドレス数	1	異なる取引に繰り返し使用された入力アドレス数
再利用出力アドレス数	1	異なる取引に繰り返し使用された出力アドレス数

表 5.4: 利用目的の推定に使用するアドレス数

	利用目的	学習アドレス数	評価アドレス数
業者	Bitcoin ATM	2	1
	Dark web	18	8
	Mining Pool	69	29
ユーザ	Bitcointalk BBS	1,674	717
	Bitcoin ATM	316	136
	Dark web	47	20
	Exchange	708	304
	合計	2,834	1,215

(iii) 推定に使用するアドレスを利用目的ごとに学習アドレス数と評価アドレス数と同数となるように 100 回ランダムサンプリング[3]する.

(iv) 決定木学習を用いてアドレスの利用目的を推定する.

決定木学習には Python の scikit-learn ライブラリ[4]より CART アルゴリズムを使用する. 図 5.2 に決定木学習のパラメータを示す. パラメータは作成する決定木の深さの最大値を 5 に設定し,1 つの葉に属する必要があるサンプルの割合の最小値が 10%以上なるように設定した [25]. 7 種類のアドレス推定精度は正解率, 適合率, 再現率で評価する.

5.5 実験結果 i.

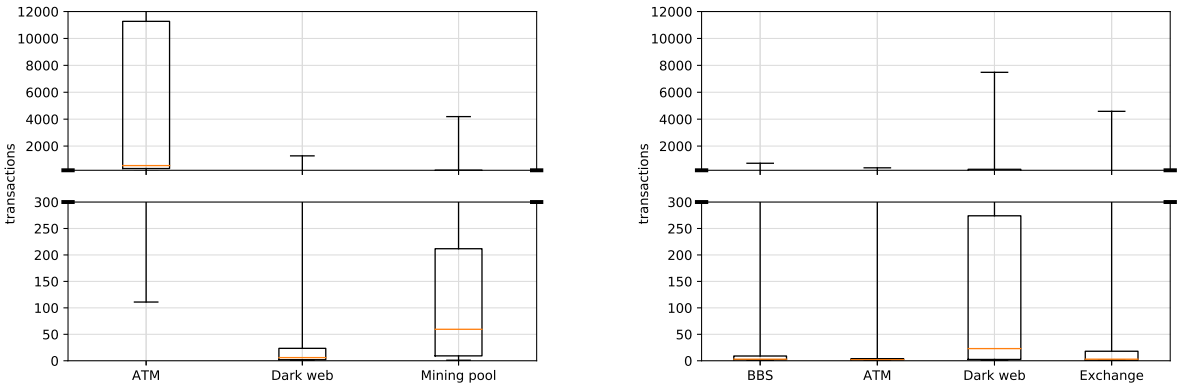
表 5.2, 表 5.3 に 7 種類の利用目的の推定に使用した特徴量を示す. 表 5.3 の統計量は平均値, 最小値, 中央値, 最大値, 標準偏差の 5 種類とする.

[3]業績 CSEC88,NBiS2020 の論文において “3-クロスバリデーションを実施” と報告したが, 正しくは “ランダムサンプリングを実施” に訂正する.

[4]scikit-learn Machine Learning in Python (https://scikit-learn.org/stable/)

表 5.5: アドレスの取引数の統計量

	利用目的	平均	最小	中央値	最大	標準偏差
業者	Bitcoin ATM	7,551	111	549	21,993	12,509
	Dark web	74	1	6	1,272	250
	Mining Pool	271	1	60	4,190	668
ユーザ	Bitcointalk BBS	13	1	3	722	42
	Bitcoin ATM	12	1	2	383	34
	Dark web	503	1	23	7,482	1,228
	Exchange	45	1	3	4,582	239



業者アドレスの取引分布　　ユーザアドレスの取引分布

図 5.3: 7 種類の利用目的の取引分布

7 種類のアドレスに関する取引回数の統計量を表 5.5 に示す．図 5.3 のユーザアドレスの取引回数分布について,BBS,ATM,Exchange アドレスは取引回数が 1 以上 25 未満で分布していることがわかる．ユーザの取引に利用されているアドレスは繰り返し使用されることが少ない (繰り返しアドレスを利用しない) ことを示している．ユーザのアドレスに対して, 図 5.3 の業者アドレスの取引回数分布は ATM と Mining Pool のアドレスには 100 回以上の取引を行っているアドレスの分布が見られる．ユーザのアドレスと比較し, 業者のアドレスは取引回数が多く, 繰り返し利用されていることを示している．

また, 図 5.2 で定義した 4 つの取引方式について,7 種類のアドレスにおける取引数と割合を表 5.6 に示す．3 つの業者のアドレスについて,3 つそれぞれの取引方式 $M1$ の割合は取引全体の 1%未満であり, 取引方式 $MS2$ の割合は取引全体の 7%未満である．業者のアドレスが行う取引は半数以上が取引方式 $S1$ を使用し,ATM は 98.5%,Dark web は 64.4%,Mining は 78.7%の割合を示している．業者のアドレスが使用するウォレットは取引の *Input* に単一のアドレスを使用する取引方式 $S1$ が使用されていた．

表 5.6: 4 つの取引方式の取引数と割合

	利用目的	S1		S2		M1		M2	
		取引数	%	取引数	%	取引数	%	取引数	%
業者	Bitcoin ATM	22,319	98.5	135	0.6	174	0.8	25	0.1
	Dark web	1,24	64.4	557	28.9	3	0.2	127	6.6
	Mining	19,569	78.7	2,845	11.4	410	0.2	2,052	6.6
ユーザ	Bitcointalk BBS	6,978	23.5	10,704	36.1	1,478	5.0	10,478	35.4
	Bitcoin ATM	1,700	33.3	2,033	39.9	44	0.9	1,323	25.9
	Dark web	7,627	23.0	12,546	37.8	1,264	3.8	11,711	35.3
	Exchange	8,730	26.2	11,269	33.8	2,908	8.7	10,444	31.3

表 5.7: 7 種類の利用目的の推定結果

利用目的	正解率 [%]		適合率 [%]		再現率 [%]	
	業者	ユーザ	業者	ユーザ	業者	ユーザ
Bitcointalk BBS	-	77	-	65	-	63
Bitcoin ATM	99	91	16	45	22	40
Dark web	98	93	6	49	9	36
Exchange	-	85	-	80	-	79
Mining Pool	92	-	70	-	65	-
全体		81		49		39

5.6 実験結果 ii.

7 種類の利用目的の推定結果を表 5.7 に示す. 3 つの評価指標のうち, 業者アドレスの正解率は全て 90%以上であり, 全体の正解率 81%よりも高い値を示している. しかし,Bitcoin ATM 業者の適合率は 16%, 再現率は 22%,Dark web 業者の適合率は 6%, 再現率は 9%であり, 全体の適合率 49%と再現率 39%よりも低い値を示している. 7 種類の利用目的において,3 つの評価指標のうち Exchange アドレスは正解率, 適合率, 再現率がいずれも全体の評価指標を上回っている.

推定精度を確認するため,7 種類の利用目的の推定アドレス数を表 5.8 に示す. 表 5.8 は 100 回実施した推定アドレス数の一例である. 表 5.8 の推定結果と 7 種の利用目的の推定アドレス数の割合を表 5.9 に示す. ユーザアドレスの Bitcointalk と Bitcoin ATM は利用目的を正しく推定できたアドレスの割合が最も高く,88%を示している. 3 つの業者アドレスは正しく推定できたアドレスの割合が低く,Mining 業者は 7%,Bitcoin ATM と Dark web は正しく推定できたアドレスはない.

推定に利用された特徴量を分析するため, 決定木学習によって作成されたモデルを図 5.4 に示す. 図 5.4 は 100 回実施したモデルの一例である.

表 5.8: 7 種類の利用目的の推定アドレス数の例

	利用目的	予測							合計
		業者			ユーザ				
		ATM	Dark web	Mining	BBS	ATM	Dark web	Exchange	
業者	ATM	0	0	0	1	0	0	0	1
	Dark web	0	0	0	8	0	0	0	8
	Mining	0	0	2	19	8	0	0	29
ユーザ	BBS	0	0	0	633	31	0	53	717
	ATM	0	0	0	16	119	0	1	136
	Dark web	0	0	0	12	3	2	3	20
	Exchange	0	0	0	56	9	0	239	304

表 5.9: 表 5.8 の推定アドレス数の割合

	利用目的	真陽性 (TP)		偽陽性 (FP)		偽陰性 (FN)	
		アドレス数	%	アドレス数	%	アドレス数	%
業者	Bitcoin ATM	0	0	0	0	1	100
	Dark web	0	0	0	0	8	100
	Mining	2	7	0	0	27	93
ユーザ	Bitcointalk BBS	633	88	112	22	84	12
	Bitcoin ATM	119	88	51	5	17	13
	Dark web	2	10	0	0	18	90
	Exchange	239	79	57	6	65	21

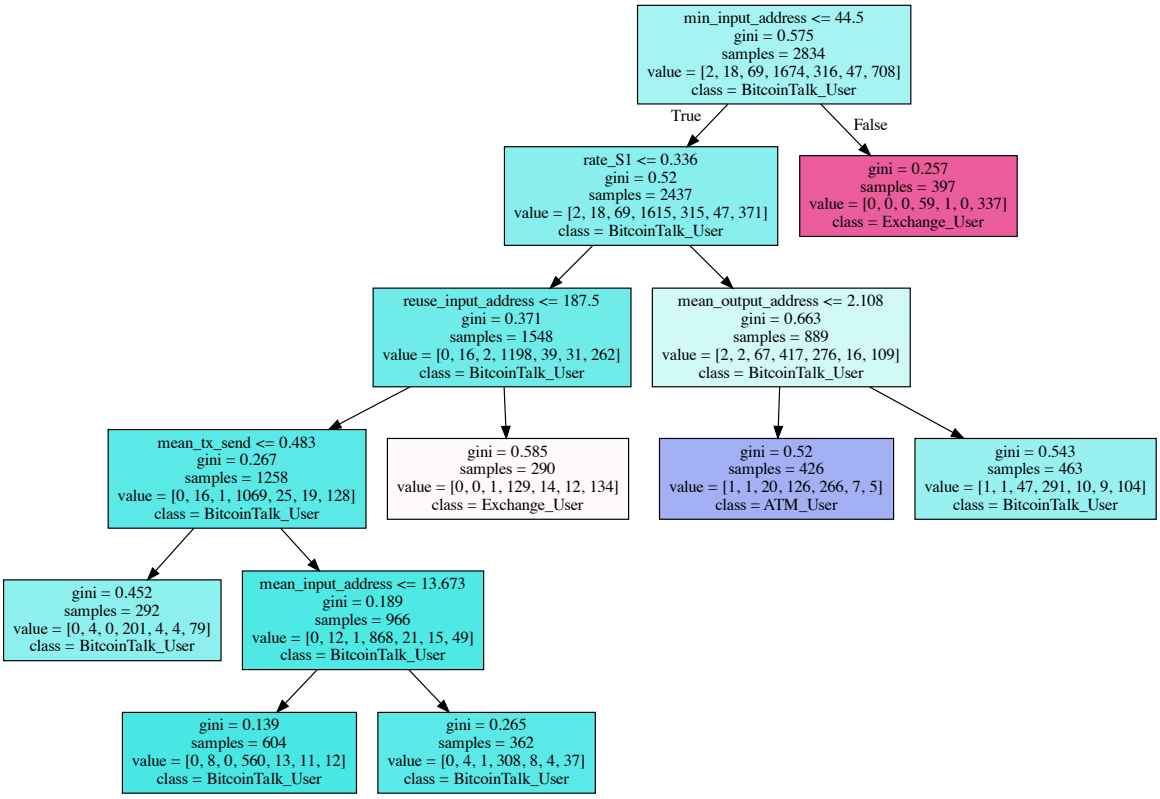


図 5.4: scikit-learn による決定木モデル例

表 5.10: 入力アドレスの最小個数の統計量

	利用目的	平均	最小	中央値	最大	標準偏差
業者	Bitcoin ATM	1	1	1	1	0
	Dark web	1.9	1	1	17	3.2
	Mining Pool	1	1	1	1	0
ユーザ	Bitcointalk BBS	7	1	1	676	40.1
	Bitcoin ATM	1.3	1	1	112	5.2
	Dark web	1.7	1	1	12	2.3
	Exchange	137.9	1	10.5	662	190

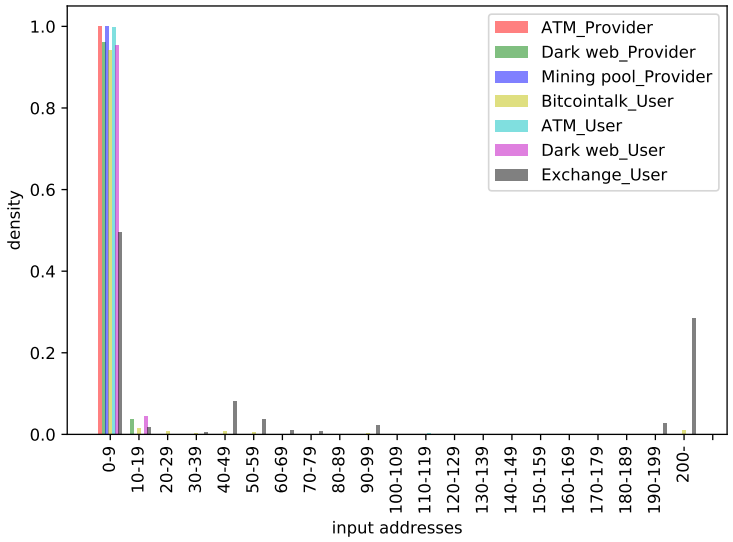


図 5.5: 入力アドレスの最小個数と 7 種類の利用目的のヒストグラム

5.7 考察

5.7.1 利用目的の推定に有効な特徴量の分析

図 5.4 に示したモデルについて, 決定木のルートに使用されている特徴量に注目した. なぜならば, ルートに利用される特徴量は全体の分類を行う際に最も効果的な特徴量が選択されるからである. ルートに使用された特徴量の統計量を表 5.10 に示す. 選択された特徴量は入力アドレスの最小個数であった. この特徴量は, 推定されるアドレスの全て取引において, 入力アドレス (*Input* フィールド) に指定されたアドレス数が最も少ない値を示している. 7 種類の利用目的のアドレスについて, 入力アドレスの最小個数のヒストグラムを図 5.5 に示す. 7 種類の利用目的のうち, 特徴量の値が 40 以上を取る利用目的は Exchange ユーザのアドレスであることが示されている.

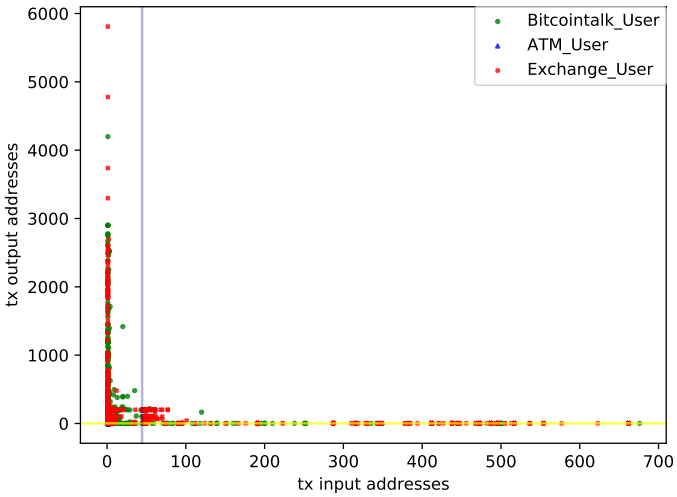


図 5.6: 入力アドレスの最小個数と出力アドレスの平均個数の散布図

5.7.2 推定可能な利用目的の分析

図 5.4 に示したモデルでは,7 種類の利用目的のうち Exchange ユーザアドレス (赤色),Bitcoin ATM ユーザアドレス (青色),Bitcointalk ユーザアドレス (水色) の 3 種類のみ推定できることを示している. 3 種類のアドレスを分類するため, 特徴量として使用された入力アドレスの最小個数と出力アドレスの平均個数に注目した. 出力アドレスの平均個数とは, 推定されるアドレスの全て取引において, 出力アドレス (*Output* フィールド) に指定されたアドレス数の平均値である. 2 つの特徴量の散布図を図 5.6 に示す. $x = 44.5$ の値をとる青色線,$y = 2.1$ の値をとる黄色線は, 図 5.4 に示したモデルで設定された閾値である.

5.7.3 業者アドレスが正しく推定できなかった原因

7 つの利用目的を定義するにあたり, 本実験ではアドレスの管理者が業者とユーザで異なる仮説を立てた. しかし, 表 5.6 で示された結果のように, ユーザが使用しているアドレスは $S1$,$S2$,$M2$ の取引頻度が高く, 業者が使用しているアドレスは $S1$,$S2$ が高い. $M2$ の取引を行っているかでユーザと業者のアドレスを推定できると予測した. しかし, 表 5.7 や表 5.8 の推定精度では, 業者のアドレスの推定精度が低く Bitcoin ATM や Dark web 業者のアドレスは正しく推定したアドレス数が 0 個となることもあった. これは, 収集したアドレス数の偏りが原因であると考える. 実際に表 5.1 に示した利用目的別のアドレス数では, 業者のアドレス数が 127 個に対してユーザのアドレス数が 3,922 個と 10 倍以上の差があることを示していた. 特に, 推定精度が低い Bitcoin ATM 業者のアドレスは 3 個,Dark web 業者のアドレスは 26 個と, 他の利用目的と比較してアドレス数が極端に少ない. 加えて, 表 5.9 で正しく推定に成功したアドレスの割合が 75%以上の Bitcointalk ユーザ,Bitcoin ATM ユーザ,Exchange ユーザのアドレスは, 収集したアドレス数が上位 3 つの利用目的であった. 従って, 分類に使用したアドレス数の差が推定精度に大きく影響を与えていると考える.

この対策として, 利用目的の間でアドレス数の差が小さくなるようにアドレスの収集やオーバーサンプリング手法を活用したデータの合成手法などを検討する.

5.7.4 入力アドレスの最小個数が有効な特徴量となった要因

図 5.4 で示した学習モデルではルート特徴量として入力アドレスの最小個数が採用されていた. 表 5.10 に示した特徴量の統計量においても Exchange ユーザは中央値が 10.5, 標準偏差が 190 と値の分散が大きい. この特徴量を使用することで,Exchange ユーザとその他の利用目的を分類している.

入力アドレスの最小個数が特徴量として利用された要因に, 交換所が行う複数のアドレスを用いた送金取引を以下に述べる.

(1) 交換所に登録しているユーザのアドレスから交換所アドレスに送金を行う取引

(2) 交換所に登録しているユーザが外部のアドレスへ送金する取引

(1) は交換所に登録しているユーザのアドレスが所有している Bitcoin を交換所で運営, 管理されているアドレスへ送金する取引である. これは, 複数のユーザが所有する Bitcoin を交換所で管理しているアドレスに送金することで,Bitcoin の管理を容易にするためであると考えられる.

(2) は交換所に登録しているユーザが所有している Bitcoin を外部のアドレス (同一の交換所で管理されてないアドレス) に送金する取引である. ユーザは送金時, マイナーへ取引の承認を依頼するため, 取引のデータサイズに応じて手数料を支払う必要がある. 交換所は複数のユーザからの送金依頼を一定期間集めることで,1 ユーザあたりの取引に必要とされていた手数料を抑えることができる. また, ユーザが支払った手数料の一部は交換所の利益となるため, 全ての交換所アドレスで同様の取引が行われていると考えられる.

(1),(2) の送金取引は交換所アドレスが頻繁に行う特有の振る舞いである. また, 特定の交換所による特異な取引ではないことからも,Exchange アドレスの推定精度が高い要因であると考える.

第6章　議論

6.1　制限と今後の研究課題

これまでに,Bitcoin アドレスの新たな識別手法の提案とアドレスの利用目的に着目した推定実験を実施した. 本章では,2 つの実験結果に対して, 本実験における制限事項とその改善方法を検討する.

6.1.1　利用したデータセットの制限

Bitcointalk のアドレスはユーザが自身のプロフィールページにアドレスを記載するため, 自身が交換所で作成したアドレスを記載していることが多くあった. また, 取り除いた交換所のアドレスは WalletExploer に掲載されているアドレスであり, 収集の対象外となる交換所 (coincheck のアドレス等) は含まれている可能性がある. そのため,7 種類の利用目的の推定で使用した表 5.1 の 4,049 個のアドレスについても, 同一の利用目的に属するアドレスが含まれている可能性が残っている. 同様に,Bitcoin ATM を利用したユーザや Dark web 上で取引が行われたアドレスは複数の利用目的で再利用されるため, 正しく評価できないことが考えられる. この問題に対して, 交換所アドレスや Bitcoin ATM の機器に設定されたアドレスなど, 複数の利用目的で使用される可能性が少ないアドレスの利用を検討する.

6.1.2　アドレス識別手法の制限

新たに提案した 2 つのアドレス集合 R,O を用いた識別手法について, 本実験では Jaccard 係数を利用した識別精度の評価を実施した. Jaccard 係数を用いた評価手法は永田ら [5] による宛先アドレス集合を用いた識別にならった手法であり, 機械学習などを用いた識別手法との精度の評価ができていない. また, アドレス識別成功の定義として,Jaccard 係数が最も高い値を持つアドレスが複数存在する場合に, 確率的な識別精度の評価を実施していない. 例えば, 本手法を用いて識別するアドレス a_1 について,Jaccard 係数が 1.0 となる予測アドレスが 2 つ存在する場合は,a_1 は $\frac{1}{2}$ の確率で識別できる. 今後は, 本実験で得られた結果を元に機械学習などを用いた識別手法や確率的な識別精度の評価手法を検討する.

6.2　倫理面に関する言及

本稿で提案した 2 つのアドレス識別手法を用いることでアドレスの所有者や個人の特定を行うことはできない. 識別に利用するアドレスについては, 取引記録をブロックチェーン上より収集し,$Input$,$Output$ に指定されたアドレスを使用して識別を実施している. 取引記録には $Input$,$Output$ フィールドに指

定されているアドレスや取引時刻, 送金額などが含まれており, いずれもブロックチェーン上に公開されている情報である.

また, 収集した情報はコンピュータセキュリティシンポジウム 2020 のサイバーセキュリティ研究における倫理的配慮のためのチェックリスト [26] を用いて倫理的配慮が必要なデータであるか確認を行った. 以下にチェックリストの内容を示す.

(1) 基本的確認

(1-1) 情報処理学会の倫理綱領を確認した.

(2) 実験のために収集した機微な情報に関して

(2-1) 個人を特定可能な情報 (PII, Personally Identifiable Information) を含む機微な情報の取り扱いに配慮したこと, およびその配慮をどのように実施したかについて, 文中に明記した.

(3) 実験の実施や論文の公開による “ネガティブな影響” について

(3-1) 事前に (製品名・サービス名や, 攻撃対象・攻撃手法などの公開に伴う)“ネガティブな影響” の検討を行った.

(3-2) 検討結果を踏まえて, 関係者への通知 (直接通知 or 届出制度を利用) を事前に行った.

(3-3) 文中に製品・サービスの具体名を表記している, もしくは, 容易に推測できる記述がある場合, そのように記述することの妥当性を検討した.

(3-4) 上述の “ネガティブな影響” を最小化するための対策について, また論文で取り上げた対象以外に他の製品・サービス等への影響についても検討した.

(3-5) (3-1)?(3-4) の検討内容に関して, 必要の程度で文中に明記した.

チェックリストへの対応について, 項目 (1-1) は情報処理学会の web ページより, 倫理綱領 [27] を確認した. 項目 (2-1) は本稿で使用した Bitcoin アドレスデータと取引データは該当しない. 項目 (3-1),(3-4) は本稿の実験結果は,Bitcoin やその他の暗号資産, ブロックチェーンシステムの脆弱性ではないと判断した. 項目 (3-2),(3-3) は本稿の報告によって特定の製品やサービスに与える影響はないと判断した. 最後に, 項目 (3-5) は本節にて詳細を明記した.

第7章　結論

本稿では Bitcoin アドレスを識別する新たに 2 つのアドレス集合 R,O を特徴量とする方式を提案した. 10 年間のアドレス用いたアドレス識別実験により, アドレス集合 O を用いた提案手法の識別結果が最も精度が高いことを示した. これは, 既存手法のアドレス集合 S, I に対して, 統計的に有意な水準で高い. また, アドレスが再利用されることで, 集合 O では最も識別精度が低い取引回数 $n = 30$ の 37.1 に対して $n = 100$ は 78.6 まで識別率が増加することを示した.

取引回数が増加するごとに識別率が高くなる, という実験結果の予測に対して, 取引回数 $n = 30$ の前後で識別率が低下する実験結果が得られた. 識別率を低下させる要因として, 交換所のアドレスが原因である結果を示した.

5 種類の利用目的を用いたアドレス識別実験では Dark web で利用されているアドレスの識別率が最も高いことを示した. また,4 つのアドレス集合のうち, 最も高い識別精度は,Meiklejohn ら [2] の集合 I が平均で 16.8 を示した. しかし, 利用目的別に識別精度の評価を行うと, 提案手法 O は Bitcointalk の 19.5 と Exchange の 22.6 で最も高い値となった. 従って, アドレスの識別を行う際には, アドレスの利用目的を考慮し,4 つの集合より適切に精度が評価できる手法を選択する必要がある.

本稿で提案した集合 O を特徴量としたアドレス識別手法への対策は, 取引の受け取り専用のアドレスを 1 つ作成し, 再利用することで識別が困難であることを示した.

利用目的別に観測される取引の特徴量では,Exchange アドレスの入力アドレスの最小個数が他の利用方法と比較してアドレス数が多いことを示した.

アドレス識別実験の結果より, 匿名性が高いとされていた Bitcoin アドレスが識別される可能性を示した. また, アドレスの利用目的を推定実験結果より, 交換所で利用されているアドレスの宛先アドレス数に特徴的な振る舞いがあることが明らかになった. 本稿で報告した内容は,Bitcoin やブロックチェーンシステムの脆弱性に該当するものではない. しかし, 公開されている取引記録やアドレス情報を用いることで, アドレス識別のリスクなどが存在していることを報告した.

“アドレスの管理者が特定できないことでユーザが匿名性が高い” とされている暗号資産について, 非公開ではあるが, 交換所などアドレスと個人情報が結びついた情報に留意する必要がある. また, 交換所へのクラッキング被害により, アドレスと個人情報が結びついたデータが流出する可能性もある. ユーザが正しい情報を学び, 安心して Bitcoin やブロックチェーンシステムを利用できる環境が望まれる.

謝辞

本稿は筆者が明治大学先端数理科学研究科先端メディアサイエンス専攻博士前期課程に在学した 2 年間の研究成果をまとめたものである．本稿が完成するまでに多くの方々のご指導, ご協力を賜りました.

指導教員である菊池浩明教授よりは, 研究に関して右も左も分からない筆者に対し, 継続的かつ熱心なご指導をいただきました．本稿を一つの研究成果としてここにまとめることができたことを深く感謝申し上げます．また, 合同ゼミやワークショップにおいて的確かつ有益なご助言をいただいた静岡大学の西垣正勝教授や大木哲史先生, 国立中山大学の Chun-I Fan 教授に心から感謝致します.

2 年間の研究活動に際して有益な意見を与えてくれた明治大学菊池研究室院生の皆様, 様々な研究会を通してご意見をいただいた方々に感謝致します．特に, 筆者が博士前期課程 1 年時に研究内容への活発な議論をいただいた, 菊池研究室の山崎孝順氏, 草野蘭之介氏, 井垣秀星氏, 永田倖大氏に心から感謝いたします.

最後に, 博士前期課程への進学にあたり, 新たな研究環境で挑戦する機会を与えてくださった両親に心から感謝致します.

参考文献

[1] Satoshi Nakamoto, “Bitcoin: A Peer-to-Peer Electronic Cash System”, (2008).

[2] Sarah Meiklejohn, Marjori Pomarole, Grant Jordan, Kirill Levchenko, Damon McCoy, Geoffrey M. Voelker, Stefan Savage, “A fistful of bitcoins: characterizing payments among men with no names”, In Proceedings of the 2013 conference on Internet measurement conference(IMC ’13), ACM, pp. 127–140, (2013).

[3] Ron D., Shamir A., “Quantitative Analysis of the Full Bitcoin Transaction Graph”, Financial Cryptography and Data Security(FC 2013). Lecture Notes in Computer Science, vol 7859. Springer, Berlin, Heidelberg. pp. 6-24, (2013).

[4] George Kappos, Haaroon Yousaf, Mary Maller, Sarah Meiklejohn, “An Empirical Analysis of Anonymity in Zcash”, In the Proceedings of the 27th USENIX Security Symposium(USENIX Security ’18), pp. 463–477, (2018).

[5] 永田 倖大, 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 第 80 回コンピュータセキュリティ研究発表会 (CSEC-80), pp. 1-6, (2018).

[6] Jules Dupont, Anna C Squicciarini, “Toward De-Anonymizing Bitcoin by Mapping Users Location”, In Proceedings of the 5th ACM Conference on Data and Application Security and Privacy(CODASPY ’15), pp. 139–141, (2015).

[7] 井垣 秀星, 永田 倖大, 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, pp. 481-482, (2019).

[8] 山崎 孝順, 草野 蘭之介, 松本 寛輝, 井垣 秀星, 菊池 浩明, “取引件数の時間分布の相関を用いた Bitcoin 取引所のユーザの属性推定”, 情報処理学会 第 82 回全国大会, pp. 407-408, (2020).

[9] Mikkel Alexander Harlev, Haohua Sun Yin, Klaus Christian Langenheldt, Raghava Rao Mukkamala, Ravi Vatrapu, “Breaking Bad: De-Anonymising Entity Types on the Bitcoin Blockchain Using Supervised Machine Learning”, Proceedings of the 51st Hawaii International Conference on System Sciences(HICSS 2018), pp. 3497–3506, (2018).

[10] 廣澤 龍典, 上原 哲太郎, “ビットコインのミキシングにおける資金移動の分析”, 情報処理学会 第 81 回コンピュータセキュリティ・第 41 回インターネットと運用技術合同研究発表会, pp. 1-8, (2018).

[11] D. Y. Huang et al., “Tracking Ransomware End-to-end”, 2018 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, 2018, pp. 618-631, doi: 10.1109/SP.2018.00047, (2018).

[12] 坂間 潤一郎, 金岡 晃, “ビットコインにおけるデジタル署名の乱数分析”, 情報処理学会 第 87 回コンピュータセキュリティ研究発表会, pp. 1-5, 2019.

[13] 井垣 秀星, 松本 寛輝, 菊池 浩明, “カナダにおける Bitcoin ATM の利用者調査”, 情報処理学会第 82 回全国大会, pp. 411-412, (2020).

[14] Alex Biryukov, Dmitry Khovratovich, Ivan Pustogarov, “Deanonymisation of Clients in Bitcoin P2P Network”, Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS ’14), Pages 15–29, 2014

[15] Garba, A., Guan, Z., Li, A. and Chen, Z., “Analysis of Man-In-The-Middle of Attack on Bitcoin Address”, In Proceedings of the 15th International Joint Conference on e-Business and Telecommunications (ICETE 2018), pp. 388-395, (2018).

[16] Shaileshh Bojja Venkatakrishnan, Giulia Fanti, Pramod Viswanath, “Dandelion: Redesigning the Bitcoin Network for Anonymity”, In Proceedings of the ACM on Measurement and Analysis of Computing SystemsJune 2017, 34 pages, (2017).

[17] Bitcoin.org プライバシーの保護 (https://bitcoin.org/ja/protect-your-privacy) (参照 2020-6-26)

[18] Coincheck 仮想通貨のウォレットとは?特徴や種類は? (https://coincheck.com/ja/article/143) (参照 2021-1-9)

[19] ITmedia NEWS ビットコインの Mt.Gox がアクセス不能——大手 6 社が共同声明 (https://www.itmedia.co.jp/news/articles/1402/25/news139.html) (参照 2021-1-9)

[20] Coincheck 仮想通貨 NEM の不正送金に関するご報告と対応について (https://corporate.coincheck.com/2018/03/08/46.html) (参照 2021-1-9)

[21] PRTIMES テックビューロ株式会社 仮想通貨の入出金停止に関するご報告、及び弊社対応について (https://prtimes.jp/main/html/rd/p/000000093.000012906.html) (参照 2021-1-9)

[22] ビットポイント 仮想通貨流出に関する 現状報告および今後の対応方針 (https://www.bitpoint.co.jp/news/info/info-2019071602/) (参照 2021-1-9)

[23] Jaccard 係数 (https://mathwords.net/jaccardkeisu) (参照 2021-1-25)

[24] 平均値の検定 (http://www.aoni.waseda.jp/abek/document/t-test.html) (参照 2020-11-28)

[25] scikit-learn で決定木分析 (CART 法) (https://pythondatascience.plavox.info/scikit-learn/scikit-learn で決定木分析) (参照 2020-2-18)

[26] コンピュータセキュリティシンポジウム 2020 サイバーセキュリティ研究における倫理的配慮のためのチェックリスト (https://www.iwsec.org/css/2020/ethics_list.html) (参照 2021-1-27)

[27] 情報処理学会倫理綱領 (https://www.ipsj.or.jp/ipsjcode.html) (参照 2021-1-27)

業績

国際会議論文 (査読あり)

1. Hiroki Matsumoto, Shusei Igaki, Hiroaki Kikuchi, Address Usage Estimation Based on Bitcoin Traffic Behavior, The 23rd International Conference on Network-Based Information Systems(NBiS-2020), pp. 188-199, 2020.

国内研究会

1. 松本 寛輝, 菊池 浩明, Bitcoin 取引履歴の特徴量に基づくアドレス識別リスクの評価, コンピュータセキュリティシンポジウム 2020(CSS2020), pp. 512-518, 2020.
2. 松本 寛輝, 井垣 秀星, 菊池 浩明, “Bitcoin サービス業者と利用者アドレスの種類の推定と評価”, 情報処理学会 第 182 回マルチメディア通信と分散処理・第 88 回コンピュータセキュリティ合同研究発表会 (CSEC-88), pp. 1-7, (2020).
3. 井垣 秀星, 松本 寛輝, 菊池 浩明, “カナダにおける Bitcoin ATM の利用者調査”, 情報処理学会 第 82 回全国大会, pp. 411-412, (2020).
4. 山崎 孝順, 草野 蘭之介, 松本 寛輝, 井垣 秀星, 菊池 浩明, “取引件数の時間分布の相関を用いた Bitcoin 取引所のユーザの属性推定”, 情報処理学会 第 82 回全国大会, pp. 407-408, (2020).