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

井垣秀星 †　　永田倖大 ‡　　菊池浩明 †

明治大学総合数理学部 †

1　はじめに

近年, 暗号通貨の利用者が増加している. その中でも 2009 年 から運用が開始された Bitcoin[1] のウォレットを保持しているユーザ数は，日本の金融庁に暗号通貨取引所への登録が開始された 2017 年 9 月 29 日の時点では 17, 222, 399 だったが, 2018 年は 28, 869, 795 と一年間で 1,000 万以上増加している. その理由として 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, 資産価値が高いという特徴がある.

しかし, そのプライバシー保護は十分ではなく, 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]. しかしながら全てのユーザが所属するタイムゾーンの昼の時間帯に活動するわけではなく, 平均推定精度は 72% にとどまっていた. さらに, 取引が 0 の時間帯を特徴量としていたので, 1 回でも取引をすると誤りを引き起こしていた.

そこで, 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.

2　実験

2.1　概要

本実験では, アドレスデータセットと取引データセットを使用し, 提案手法である平均取引時間分布を Bitcoin ユーザの 25% から作成, 残りの Bitcoin ユーザの取引時間分布との相関係数を用い Bitcoin ユーザのタイムゾーン推定の精度を明らかにする. 本手順のシステム構成図を図 1 に示す.

2.2　方法

2.2.1　アドレスデータセットの取得

Bitcoin のオンラインフォーラムである Bitcointalk*のプロフィールページにて公開されている address と Location データをスクレイピングにて取得し, address テーブルに表 1 に格納する. トランザクション内の同一 Input に格納されているアドレスは同一ユーザのものであると

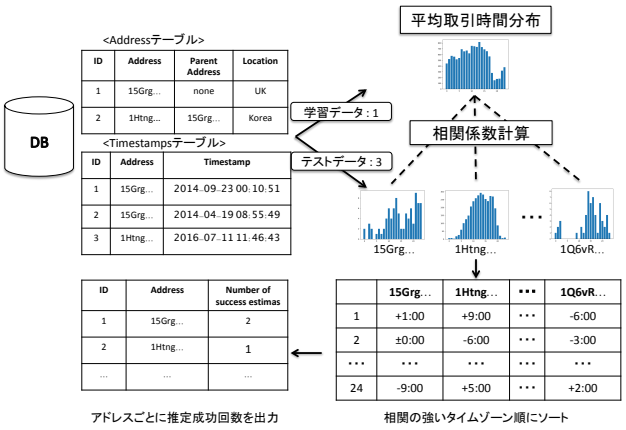

図 1　システム構成図

仮定し, 同一ユーザのアドレスを補完する. Bitcoin の取引所 BLOCKCHAIN†の取引データから, 同一 Input の収集元アドレスを Parent address と呼び, address テーブルに追加する. これらをアドレスデータセットとする.

2.2.2　取引データセットの取得

BLOCKCHAIN 取引データから取得する. アドレスデータセットの address が Input に入っている transaction 時刻の値を取得し, 表 2 の Timestamps テーブルに格納する.

2.2.3　タイムゾーンデータセットの取得

全タイムゾーンのデータを timeanddate‡から取得し, time zone テーブルに格納する.

2.2.4　平均取引時間分布データの作成

平均取引時間分布データは, アドレスデータセットの 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.

2.2.5　各アドレスの取引時間分布と相関の出力

2.2.4 節にて作成した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

Time zone estimation of Bitcoin user based on correlation of distributions of transaction time

†Shusei Igaki Department of Frontier Media Science, School of Interdisciplinary Mathematical Science, Meiji University.

‡Graduate School of Advanced Mathematical Sciences, Meiji University

*https://bitcointalk.org/

†https://www.blockchain.com/explorer

‡https://www.timeanddate.com/

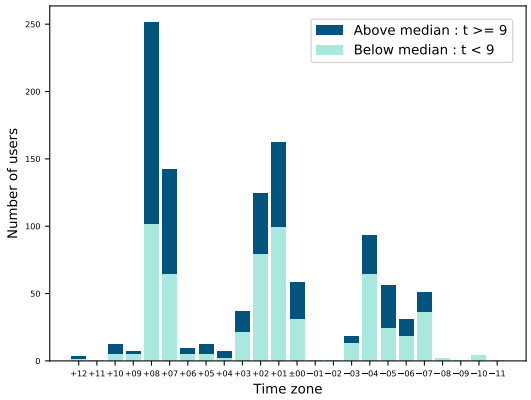

図 2　タイムゾーンごとのユーザ数

表 1　Address テーブルの例

Address	Parent address	Location
1CasperDEhyGD81WNPo9qkaFnWxUSWmrqk	none	Hong Kong
1FdxQxtzkRRcCApy7AFGroUgjesyLKRENK	1CasperDEhyGD81WNPo9qkaFnWxUSWmrqk	Hong Kong

表 2　Timestamps テーブルの例

Address	Timestamp
1CasperDEhyGD81WNPo9qkaFnWxUSWmrqk	2012-02-18 00:57:33
1CasperDEhyGD81WNPo9qkaFnWxUSWmrqk	2012-02-18 04:24:46
1FdxQxtzkRRcCApy7AFGroUgjesyLKRENK	2012-06-23 23:45:23

表 3　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

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 $c(fi, f_*), c(fi, f_* + 1), \ldots, c(fi, f_* + 24)$ を求め, 最大化するタイムゾー i_* をユーザ i のタイムゾーンと推定する. すなわち,

$$j_* = \underset{j\in\{0,\ldots,24\}}{\arg\max}\, c(f_i, f_* + j) \tag{1}$$

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 θ[時] 以内を推定成功とする. これをユーザごとに求める.

2.2.6　推定成功率の出力

本実験では, 2.2.4, 2.2.5 節の手順を 1000 回実施し, 推定成功回数の平均を求める.

2.3　結果と推定成功率の評価

取得したデータセットの概要を表 3, タイムゾーンごとのユーザ数の分布を図 2 に示す. ここで, 全ユーザの取引回数の中央値である 9 回を基準にし, 取引回数に応じてユーザを 2 色に分けている.

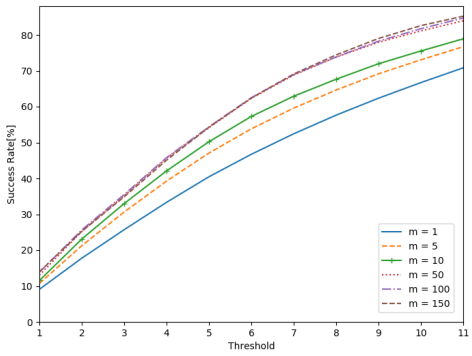

図 3　推定成功率

推定成功回数閾値 θ 回以上の推定成功回数のユーザ数の割合を推定成功率

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

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. 結果を図 3 に示す.

2.4　考察

本実験では, 推定成功率は 77% で Dupont らの手法よりも推定成功率が高い. このことから, 多くの取引を行う場合は複数のアドレスを並行して使用し, またその場合は同一ユーザのものとされないために同一 Input に入るような使用はしないようにする必要があると考える.

3　まとめ

本実験では Bitcoin アドレスの平均取引分布に基づくタイムゾーンの推定を行った. その結果, 相関を用いた推定成功率は Dupont らの手法を上回る結果であることを確認した.

今後の課題として, データをより多く取得することをあげる.

参考文献

[1] S. Nakamoto, ”Bitcoin: A Peer-to-Peer Elecctronic Cash System”, 2008.

[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), pp.139-141, ACM, 2015.