履歴データに対する匿名化モデル k-concealment の改良手法の提案

伊藤 聡志[1,a)]　菊池 浩明[1,b)]

概要：匿名加工は，個人が識別されることを防ぐために個人情報を加工する技術である．企業などの組織が収集したビッグデータ (顧客情報や位置情報など) を分析・利活用することにより，我々は様々な恩恵を得ることができるが，そのためにはデータの匿名加工とリスク評価が不可欠である．個人データを匿名化する手法として，2012 年に Tamir らによって k-concealment が提案されている．この手法は代表的な k-anonymity と同程度の安全性で有用性が高い特長がある．Tamir らは，顧客数がレコード数に等しいデータを仮定しており，購買履歴や位置履歴などの様に，顧客当たりのレコード数が可変なデータには適用できない制約があった．そこで，本研究では，動的データ（顧客数 < レコード数）にも適用可能な k-concealment 手法を提案する．一般化やレコード削除を用いることによって，動的データに対しても効果的に k-concealment を満たす手法を提案する．

キーワード：匿名加工，k-anonymity，k-concealment，履歴データ

Proposal on Modified k-concealment Model for Anonymized Histories of Records

Satoshi Ito[1,a)]　Hiroaki Kikuchi[1,b)]

Abstract: De-identification is a process to prevent individuals from being identified from original transaction data by processing personal identification information. Companies are required to assess the re-identi cation risks when employing big data extensively in their businesses. In 2012, Tamir proposed k-concealment method that is extension of k-anonymity. This method ensures same security and better utility for de-identified data. However, the history data like purchase history can not be de-identified by this method because this method supposed for only static data (the numbers of user and record are same). In this paper, we propose a modified k-concealment method for dynamic data (the number of record is greater than that of user).

Keywords: De-identification, k-anonymity, k-concealment, History data

1. はじめに

企業などの組織が収集したビッグデータ (顧客情報や位置情報など) を分析・利活用することにより，我々は様々な恩恵を得ることができるが，そのためにはデータの匿名加工とリスク評価が不可欠である．匿名加工は個人が識別されることを防ぐために個人情報を加工する技術であり，匿名加工されたデータから個人を識別しようとする攻撃を再識別という．匿名加工されたデータは有用性・安全性の 2 点で評価されることが多く，評価指標が盛んに研究されている．

代表的なデータの評価指標に，Sweeny によって提案された k-anonymity[1] がある．この指標はある公開データ中の少なくとも k レコード（人）の区別がつかないことを

[1] 明治大学大学院先端数理科学研究科
Graduate School of Advanced Mathematical Sciences, Meiji University
[a)] mmhm@meiji.ac.jp
[b)] kikn@meiji.ac.jp

保証するものである．k-anonymity はそのわかりやすさからデータの評価に広く用いられており，日本では k-匿名性と呼ばれ，データが k-匿名性を満たすように加工することを k-匿名化と呼んでいる．しかし，多くの研究者がこの k-anonymity の弱点を指摘しており，代替案や改善案を提案している [2][3]．

k-anonymity では「最低でも k レコード（人）の区別がつかない」という状態を満たすために大きさ k のグループを作り，それに属するレコードを加工して 出来るだけ等しくしていたが，この手法では大きさが $k+1$ 以上のグループができてしまうことがある．Tamir はこの k-anonymity の無駄を指摘し，それを改善・拡張した指標である k-concealment[4] を提案した．k-concealment は元データと加工データの関係を 2 部グラフに置き換えて定めた指標であり，k-anonymity と同等の安全性を保ちつつ，k-anonymity より有用性を高めることができることが特徴である．しかし，k-concealment は静的データ（レコード数 = 顧客数）を評価・加工する場合のみを想定されたものであり，動的データ（レコード数 > 顧客数）の場合は想定されていない．

また，動的データを k-匿名化する際にはさらに多くの無駄が生じる．前述した大きさが $k+1$ **以上のグループができてしまう** 問題に加え，動的データでは顧客ごとにレコード数が異なるので，k 人の区別がつかなくするためには **レコードを削除したり，ダミーレコードを足して** k 人のレコード数をそろえる必要がある．例えば 2018 年度に開催された匿名加工・再識別コンテスト PWSCUP では動的データ（購買履歴データ）が取り扱われたが，レコード数が最も多い顧客と 2 番目に多い顧客の区別がつかないようにするためには，5,890 レコード中 2,688 レコードを削除する必要があった [5]．

問題点をまとめると以下の 3 点である．

(1) k-concealment では動的データは想定されていない
(2) 動的データ内の k 人の区別がつかなくするためにはレコード削除，またはダミーレコード追加を行う必要がある
(3) k-匿名化では $k+1$ 以上の冗長なグループが生じることもある

本稿ではこれらの問題の解決を研究目的とし，レコードや顧客の追加や削除をせずに，k-concealment のアイデアに基づいて動的データ内の k 人の区別がつかなくする手法を提案する．また，動的データを k-concealment 化する手法によって人流データを加工する一つのアルゴリズムを実装し，その効果を評価する実験を行う．これらの提案手法・実験結果を本稿の貢献とする．

本稿では，2 章で既存研究の説明をし，3 章で動的データに対して k-concealment 化を行う手法を提案し，4 章で人流データを用いた評価実験を行う．

表 1 Original Data T_1

name	age	zipcode
Alice	30	10055
Bob	21	10055
Carol	21	10023
David	55	10165
Eve	47	10224

表 2 Anonymized Data T_2

id	age	zipcode
1	21-55	10***
2	21-55	10***
3	21-55	10***
4	21-55	10***
5	21-55	10***

表 3 Anonymized Data T_3

id	age	zipcode
1	21-30	100**
2	21-30	100**
3	21-30	100**
4	47-55	10***
5	47-55	10***

2. 既存研究

2.1 *k*-anonymity

k-anonymity[1] は Sweeny によって提案された匿名性指標であり，ある公開データのすべてのレコードが少なくとも他の $k-1$ レコードと区別できないことを保証するものである．説明のためのデータ T_1, T_2, T_3 を表 1,2,3 に示す．T_1 は 5 人分の年齢と郵便番号のデータであり，T_2, T_3 は T_1 を加工したデータである．T_1 はすべてのレコードが異なっているので 1-anonymity しか満たしていないが，T_2, T_3 はそれぞれ 5-anonymity, 2-anonymity を満たしている．

k-anonymity は非常にシンプルでわかりやすい指標であり，広く用いられている．しかしながら k-anonymity を満たすようにデータを加工する際に無駄が生じることを弱点として指摘されている [4]．例えば 2-anonymity を満たすように T_1 を加工する場合，T_1 は 5 レコードのデータであるため，どうしても T_3 のように 3 レコードの区別がつかないグループ（レコード 1,2,3）を作らなくてはならない．「少なくとも 2 レコードの区別がつかない」という状態さえ満たせば 2-anonymity を満たしているといえるため，3 レコードの区別がつかないグループは無駄であるといえる．

2.2 *k*-concealment

k-concealment[4] は Tamir によって提案された匿名性指標であり，k-anonymity を改善・拡張したものである．k-concealment は元データと加工データの関係が 2 部グラ

フ [6] [*1]に置き換えられることに注目した指標であり，元データの全レコードが少なくとも k 種類の完全マッチングの辺（match）を持つことを保証するものである．本稿では k-concealment を満たすようにデータを加工することを k-concealment 化と呼ぶ．

定義 2.1 (k-concealment) 表 D とその一般化を D'，E を D と D' の間の辺とし，(D, D', E) を 2 部グラフとする．全ての $d \in D$ と $d' \in D'$ について，$(d, d') \in E$ であり，(d, d') を含むある完全マッチングが存在するとき，d は match を持つという．全ての $d \in D$ について，少なくとも k 個の異なる match を持つとき，D' を D の k-concealment と呼ぶ．

例 2.1 T_1 と T_3 の間の完全マッチングの一例を図 1 に示す．これらの辺はすべて完全マッチングの一部であるため，match である．また，T_1 と T_3 の間のレコード関係を 2 部グラフ (T_1, T_3, E) で表したものを図 2 に示す．元データの各レコードと加工後候補として当てはまるレコードの間に辺が張られている．例えば，元データ T_1 の Alice のレコードの加工後候補として，加工後データ T_3 では id=1,2,3 のレコードが当てはまるため，Alice のレコードからは 3 本の辺が張られている．図 2 のすべての辺も完全マッチングの一部になりうるため match である．よって，元データ T_1 の各レコードは加工後データ T_3 との間に少なくとも 2 本の match を持つため，T_3 は 2-concealment を満たしているといえる．（元データと加工後データ間の完全マッチングを攻撃者の再識別パターンと考えるとよい）

新たな加工後データとして表 4 の T_4 を考える．T_4 は T_3 の一部（1,3 行目）を加工前に戻したデータであり，k-anonymity の観点でこのデータを評価すると，1-anonymity しか満たしていないが，k-concealment の観点で評価をするとこのデータは 2-concealment を満たしている．T_1 と T_4 の関係を示す 2 部グラフを図 3 に示す．これらの辺はすべて match であり，データの一部を元に戻したことによって辺の総数は減っているが，各レコードが少なくとも 2 本の match を持っている．そのため，T_4 も「最低でも 2 人の区別がつかない」という状態を満たしている．

2-anonymity と 2-concealment を満たす T_3 と，1-anonymity と 2-concealment を満たす T_4 を比較してみよう．どちらも「最低でも 2 人の区別がつかない」という状態を満たす一方，T_4 は T_3 の一部を加工前に戻したデータであるため，T_3 よりも高い有用性を持っている．例えば加工されたセル数でデータの有用性を評価すると，T_3 の有用性は 10，T_4 は 8 であり，T_4 の方が元データに近い（値が小さいほど有用性が高い）．このように，k-concealment の観点でデータを評価・加工することにより，k-anonymity の場合と同等の安全性（最低でも k 人の区別がつかない）で，より高い有用性を持つデータを作ることができる．

[*1] グラフ (V, E) において，V の分割 V_1, V_2 に対して全ての辺 $e \in E$ が V_1 と V_2 に属すとき，その時に限り 2 部グラフ (bipartite graph) という．

name	age	zipcode
Alice	30	10055
Bob	21	10055
Carol	21	10023
David	55	10165
Eve	47	10224

id	age	zipcode
1	21-30	100**
2	21-30	100**
3	21-30	100**
4	47-55	10***
5	47-55	10***

図 1 T_1 と T_3 の間の完全マッチングの一例

name	age	zipcode
Alice	30	10055
Bob	21	10055
Carol	21	10023
David	55	10165
Eve	47	10224

id	age	zipcode
1	21-30	100**
2	21-30	100**
3	21-30	100**
4	47-55	10***
5	47-55	10***

図 2 T_1 と T_3 の関係を示す 2 部グラフ (T_1, T_3, E)

表 4 Anonymized Data T_4（1-anonymity, 2-concealment）

id	age	zipcode
1	21-30	10055
2	21-30	100**
3	21	100**
4	47-55	10***
5	47-55	10***

name	age	zipcode
Alice	30	10055
Bob	21	10055
Carol	21	10023
David	55	10165
Eve	47	10224

id	age	zipcode
1	21-30	10055
2	21-30	100**
3	21	100**
4	47-55	10***
5	47-55	10***

図 3 T_1 と T_4 の関係を示す 2 部グラフ

2.3 動的データの k-匿名化

定義 2.2 顧客数とレコード数が等しいデータを静的データ，レコード数が顧客数より多い（1 顧客が複数のレコードを持つ）データを動的データとよぶ．

動的データを k-匿名化する場合，k レコードではなく k 人の顧客の区別がつかなくなるようにする必要がある．また，動的データでは 1 顧客ごとのレコード数が異なる場合が多いため，k-匿名化する場合にはレコードを削除または追加し，レコード数をそろえる必要がある．

動的データの例として，購買履歴データ T_5 を表 5 に示す．T_5 は顧客 3 人の 4 日分の購買履歴であり，このデータを 3-匿名化する場合を考える．Alice は 3 レコード，Bob と Carol はそれぞれ 2 レコードを持っているため，この 3 人の区別がつかなくするためにはダミーレコードを追加

表 5　購買履歴データ T_5

name	date	goods
Alice	12/1	a
Alice	12/2	b
Alice	12/3	c
Bob	12/2	d
Bob	12/3	e
Carol	12/3	f
Carol	12/4	g

表 6　3-匿名化された購買履歴データ T_6

id	date	goods
1	12/1	a
1	12/2-12/3	b,d,f
1	12/3-12/4	c,e,g
*2	12/1	a
2	12/2-12/3	b,d,f
2	12/3-12/4	c,e,g
*3	12/1	a
3	12/2-12/3	b,d,f
3	12/3-12/4	c,e,g

し，すべての顧客の持つレコード数を 3 にそろえてやる必要がある．T_5 にダミーレコードを追加して 3-匿名化したデータ T_6 を表 6 に示す．id に*印がついているレコードがダミーレコードであり，この場合 2 レコードを追加してデータを一般化することによって 3 顧客の区別がつかなくなっている．また前述したように，このデータを完全 2-匿名化することはできない．(1 人余ってしまうため)

3. 動的データの k-concealment 化

3.1 アイデア

2 章で紹介した k-concealment は静的データを評価・加工する場合を想定したものであり，動的データを評価・加工することは考えられていなかった．

前述したように，**動的データを k-匿名化するためにはレコードを削除・追加して，顧客ごとのレコード数をそろえる必要がある**と考えられており，これについて 2018 年に開催された匿名加工・再識別コンテスト PWSCUP-2018[5] にて議論されている．この大会では顧客 400 人の 1 年分の購買履歴データ（81,776 レコード）が加工対象として用いられており，上位チームの加工データには k-匿名化されたものが多かった．しかしながら，データが大規模であるほど顧客ごとのレコード数の違いも大きく，このデータ内で最もレコード数が多い顧客（4,289 レコード）と 2 番目に多い顧客（1,601 レコード）の区別をつかなくするには，2,688 レコードを削除する必要があった．この大会では，レコードを削除するとデータの有用性が大きく下がるようにルールが定められていたため，レコード数が多く k-匿名化にコストのかかる顧客を見捨てる（加工をしない）チームもあった．

表 7　仮名が一般化された購買履歴データ T_7

id	date	goods
1,2,3	**12/1**	**a**
1	12/2-12/3	b,d,f
1	12/3-12/4	c,e,g
2	12/2-12/3	b,d,f
2	12/3-12/4	c,e,g
3	12/2-12/3	b,d,f
3	12/3-12/4	c,e,g

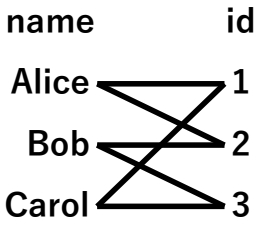

図 4　T_5 を 2-concealment した場合の 2 部グラフ

しかしながら，**動的データを「最低でも k 人の区別がつかない」状態にするためにはレコードの追加・削除や顧客の削除が必須なのだろうか？**本研究では，ここに k-concealment のアイデアを導入することにより，この問に否定的に回答する．キーとなる手法は，レコード数の異なる複数の顧客に対する「仮名の一般化」と「レコードの k-concealment 化」である．次節ではその手法・アルゴリズムを提案する．

3.2 仮名の一般化，レコードの k-concealment 化

加工データの仮名を一般化することにより，追加・削除するレコードの数を減らすことができる．例えば T_6 は 3-匿名化された T_5 であるが，これの仮名を表 7 の T_7 の 1 レコード目のように一般化することによって，ダミーレコードの数を 0 にすることができる．仮名 1,2,3 は実際には別のランダムな ID，例えば 4 とみなせばよい．仮名 4(1,2,3) は，Alice,Bob,Carol のどの顧客にも当てはまるようにすることで，加工レコードと真のレコードの関係を識別不能にする．

T_5 を 2-匿名化するためには顧客の削除が不可欠であるが，仮名の一般化だけではこの問題は解決できないため，さらにレコードの k-concealment 化を行う．表 8 の T_8 のように加工することにより，顧客と仮名が図 4 の関係を満たすように T_5 を 2-concealment 化することができる．T_8 は顧客 2 人の区別がつかないのに加え，2 レコードの区別がつかない状態にもなっており，図 4 の通り Alice と Carol が id1 に，Alice と Bob が id2 に，そして Bob と Carol が id3 に一般化されている．これは T_7 と比較すると加工の幅が小さくなっていることがわかる．また，T_5 と T_8 の間のレコードの対応を表す 2 部グラフを図 5 に示す．

表 8　2-concealment 化された購買履歴データ T_8

id	date	goods
1,2	12/1-12/3	a,f
1	12/2-12/3	b,f
1	12/3-12/4	c,g
2	12/1-12/2	a,b,d
2	12/3	c,e
3	12/2-12/3	d,f
3	12/3-12/4	e,g

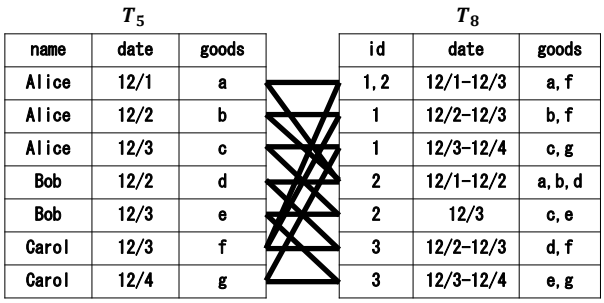

図 5　T_5 と T_8 間のレコードの対応を示す 2 部グラフ

3.3　基礎定義

動的データに対する k-concealment 手法を提案するために，以下の定義を行う．

定義 3.1　動的データを D，D のレコード数を m，顧客数を n，属性数を ρ とする．顧客の集合を $U = \{u_1, u_2, \ldots, u_n\}$ とし，D のうち顧客 u_i が持つ r_i 個のレコードの集合を $D_i = \{d_1^{(i)}, d_2^{(i)}, \ldots d_{r_i}^{(i)}\}$ とする．ここで，j 番目の要素は $d_j^{(i)} = (u_i, d_{j,2}^{(i)}, \ldots, d_{j,\rho}^{(i)})$ である．
$D = D_1 \cup D_2 \cup \ldots \cup D_n$ であり，$r_1 + \cdots + r_n = m$ である．また，D の 1 列目の属性は顧客の識別子であることを仮定している．

定義 3.2　仮名化された D を D' とし，U を仮名化した仮名の集合を $V = \{v_1, v_2, \ldots, v_{n'}\}$ とする．ここで，仮名 v_i が持つレコードの集合は D'_i とする．ここで，その j 番目の要素は $d'^{(i)}_j = (v_i, d_{j,2}^{(i)}, \ldots, d_{j,\rho}^{(i)})$ である．

1 人の顧客に複数の仮名を振ることを許す．すなわち，$n \leq n'$ となることがあることに注意せよ．

例 3.1　$D = T_5$ の場合, $m = 7, n = 3$ であり, U =\{Alice, Bob, Carol\} である．u_1 =Alice とすると $r_1 = 3$ であり，D_1 は T_5 の 1,2,3 レコード $\{d_1^{(1)}$=(Alice,12/1,A), $d_2^{(1)}, d_3^{(1)}\}$ を要素とする集合である．この場合，U は $V = \{1, 2, 3\}$ に仮名化される．

定義 3.3　D_i と D_j について定められる顧客 u_i と u_j 間の距離を $m_{i,j}$ で表し，U の顧客間の距離行列を M とする．M は n 次正方行列である．レコード数の異なる顧客間の距離は，D_i と D_j の全レコードを一般化した場合を考えて求め，余ったレコードは削除したときのコストで定める．

例 3.2　例えば，T_5 についての距離行列 M は表 9 のようになる．PWSCUP-2018 で濱田らは一般化されたデータと元データの距離を，一般化した集合の大きさに応じて，真の値となる期待値で定義している [5]．（ただし，表 9 はこの定義で求めたものではない）

定義 3.4 (顧客間対称 k-concealment)　顧客集合 U とその仮名集合 V となる k-concealment な 2 部グラフ $G = (U, V, E)$ において，$u_i \in U$ と $v_j \in V$ について $(u_i, v_j) \in E$ ならば $(u_j, v_i) \in E$ であるとき，G は対称であるという．また，G は M をもとにして距離を最小化されたものである．

例 3.3　図 4 と図 6 はどちらも 2-concealment を満たす 2 部グラフであるが，図 4 は対称であり，図 6 は対称ではない．

定義 3.5 (レコード間 k-concealment)　$G = (U, V, E)$ を k-concealment を満たす対称な 2 部グラフ, D を U が有するレコード集合, D' をその仮名化とする．全ての $(u_i, v_j) \in E$ について，$\{d_1^{(i)}, \ldots, d_{r_i}^{(i)}\} \subset D$ と $\{d'^{(j)}_1, \ldots, d'^{(j)}_{r_j}\} \subset D'$ が 2 部グラフとなる辺集合 $E_{i,j} \subset E'$ が存在し，$(d_x^{(i)}, d'^{(i)}_y)$ $(x \neq y) \in E_{i,j}$ は存在しないとき，$H = (D, D', E')$ をレコード k-concealment という．

例 3.4　図 6 は定義 3.5 の条件を満たしているため，レコード 2-concealment である．

定義 3.6　レコード $(d_{x,1}^{(i)}, \ldots, d_{x,\rho}^{(i)})$ と $(d_{y,1}^{(j)}, \ldots, d_{y,\rho}^{(j)})$ の一般化とは，レコード $(v, g_2, \ldots, g_\rho)$ とする．ここで，$\ell = 2, \ldots, \rho$ について $g_\ell = g_{x,\ell}^{(i)} \cup g_{y,\ell}^{(j)}$ である．また，仮名 v は

$$v = \begin{cases} v^{(i)} & (r_i \geq r_j) \\ \{v^{(i)}, v^{(j)}\} & (otherwise) \end{cases}$$

とする．ここで，$v^{(i)}$ は v_i に割り当てた一意でランダムな仮名である．

3.4　提案アルゴリズム

3.1,3.2 節にて提案した，動的データに対する k-concealment 化手法をアルゴリズム 1 に示す．入力する D は元データ，k は顧客に対する安全性のパラメータであり，出力される D'' は k-concealment を満たす加工された動的データである．

例 3.5　例として，アルゴリズム 1 を用いて，T_5 を 2-concealment 化する場合を考える．入力するものは $D = T_5, k = 2$ とする．

Step 1.
D の顧客をレコード順にソートする．この場合 $U = \{u_1, u_2, u_3\}$ =\{Alice,Bob,Carol\} であり，$r_1 = 3, r_2 = 2, r_3 = 2$ である．また，U は $V = \{v_1, v_2, v_3\} = \{1, 2, 3\}$ に仮名化するとし，仮名化された D を D' とする．

Step 2.
D_1, D_2, D_3 の距離を測り，距離行列 M を作る．この場合，M は表 9 のようになるとする．

Algorithm 1 提案手法

Input: 動的データ D, integer k，顧客集合 U

Output: 加工データ D''

Step 1.

D を仮名化して D' とする．D' の仮名集合を V とする．

Step 2.

$D_1, \ldots, D_n$ の距離を $n \times n$ の距離行列 M で表す．

Step 3.

M をもとにマッチングコストを最小化した，顧客間の対称かつ k-concealment な 2 部グラフ $G = (U, V, E)$（定義 3.4）を作る．

Step 4.

G をもとに，レコード間 k-concealment な D と D' の 2 部グラフ $H = (D, D', E')$（定義 3.5）を作る．

Step 5.

$d'^{(j)}_{\ell} \in D'$ について（$\ell = 1, \ldots, m$），$(d, d'^{(j)}_{\ell}) \in E'$ となる全ての $d \in D$ の一般化を g_ℓ とする（定義 3.6）．

Step 6.

$g_1, \ldots, g_m$ を結合し，D'' を作る．

表 9　T_5 についての距離行列 M

name \id	1	2	3
Alice	0	1	5
Bob	1	0	4
Carol	5	4	0

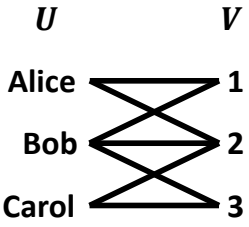

図 6　M をもとに T_5 の顧客と仮名の対応を示した 2 部グラフ G

Step 3.

M をもとに対称かつ k-concealment な 2 部グラフ $G = (U, V, E)$ を作る．この場合，G は図 6 のようになるとする．

Step 4.

G をもとにレコードの 2 部グラフ $H = (D, D', E')$ を作る．この場合 H は図 7 のようになる．実線はマッチングの際に D' から張られた辺であり，点線は実線を反転することによって張られた辺である．赤実線はマッチングで余ってしまうレコード $(d'^{(1)}_3)$ から張られている辺であり，このようなレコードの仮名が一般化される．

Step 5,6.

D' の各レコードを，H で辺がつながっているレコードと一般化し，D'' を作る．この場合，D'' は表 10 のようになり，このデータは 2-concealment を満たしている．

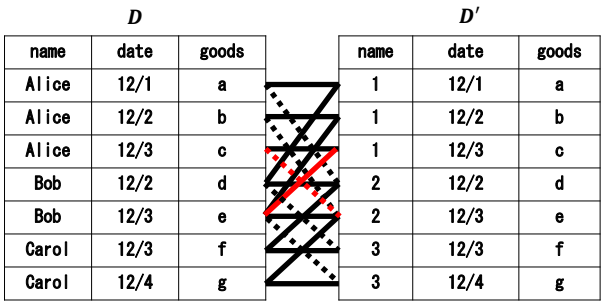

図 7　T_5 におけるレコード間 2 部グラフ H

表 10　H をもとに 2-concealment 化されたデータ D''

id	date	goods
1	12/1-12/2	a,d
1	12/2-12/3	b,e
1,2	12/3	c,e
2	12/1-12/3	a,d,f
2	12/2-12/4	b,c,e,g
3	12/2-12/3	d,f
3	12/3-12/4	e,g

4. 実験

本章では動的データを k-concealment 化することによって，データの特性がどう変化するかを実験によって評価する．しかし，この章で用いる k-concealment 化手法は 3 章で提案したものと厳密に同じではなく，レコードを補間することによる簡易的な手法であることに注意せよ．

4.1 レコード補間 k-concealment 化手法

動的データの特徴は顧客ごとにレコード数が異なることであるが，レコードを補間してレコード数を揃えることによって静的データのように扱うことができ，容易に k-concealment 化をすることができる．T_5 を 2-concealment 化する場合を考える．このデータは顧客 3 人の 12/1〜12/4 の購買履歴データであるが，4 日分すべてのデータを持っている顧客はいない．そこで，各顧客の欠けているレコードを補間して，すべての顧客が 4 日分のデータ（4 レコード）を持つようにデータを変形する．T_5 のレコードを補間したデータ T_9 を表 11 に示す．例として，図 4 のような関係になる加工データを作る場合を考えると，Alice と Carol を一般化した id 1，Alice と Bob を一般化した id 2，Bob と Carol を一般化した id 3 を作成すれば 2-concealment を満たすことができる．T_9 を図 4 を満たすように一般化したデータ T_{10} を表 12 に示す．この手法は，顧客間の距離を測ることとレコード間の対応を決めることが容易である．

4.2 疑似人流データ

本章では，ナイトレイ社から公開されている疑似人

表 11　レコードを補間した購買履歴データ T_9

name	date	goods
Alice	12/1	a
Alice	12/2	b
Alice	12/3	c
Alice	12/4	*
Bob	12/1	*
Bob	12/2	d
Bob	12/3	e
Bob	12/4	*
Carol	12/1	*
Carol	12/2	*
Carol	12/3	f
Carol	12/4	g

表 12　2-concealment を満たす購買履歴データ T_{10}

id	date	goods
1	12/1	a,*
1	12/2	b,*
1	12/3	c,f
1	12/4	g,*
2	12/1	a,*
2	12/2	b,d
2	12/3	c,e
2	12/4	*
3	12/1	*
3	12/2	d,*
3	12/3	e,f
3	12/4	g,*

表 13　疑似人流データと実験データの統計量

データ	m	n
疑似人流データ	6,432	901,465
D_{10}	10	1,402
D_{50}	50	6,608
D_{100}	100	13,503
D_{500}	500	68,699
D_{1000}	1,000	141,511

流データ [7] を用いる．このデータから 10 人,50 人,100 人,500 人,1000 人の顧客をランダムに抽出したものを順に $D_{10}, D_{50}, D_{100}, D_{500}, D_{1000}$ とし，これらを実験に用いる．疑似人流データと 5 つの実験用データの統計量を表 13 に示す．

疑似人流データは 9 属性（顧客 ID，性別，日時，緯度，経度，位置カテゴリ 1，位置カテゴリ 2，状態，カテゴリ ID）のデータであるが，本稿ではそのうち 4 属性（顧客 ID，日時，緯度，経度）だけを用いている．D_{100} をレコード補間したのち，各顧客間の緯度・経度のユークリッド距離の分布を図 8 に示す．

4.3　評価実験

k-anonymity と k-concealment を比較するために，いくつかの実験を行った．k-anonymity 化されたデータと k-concealment 化されたデータの有用性と n の関係を図 9 に示す．ここで，有用性は元データと加工データの距離のユーザ平均値で定める誤差で評価する．青線が 2-anonymity 化されたデータの，赤線が 2-concealment 化されたデータの各々の有用性を示している．この場合 n の値にかかわらず，2-anonymity 化されたデータの方が有用性が高い（= 誤差が小さい）ことがわかる．図 10 に $n = 100$ のときの誤差の分布を示す．図 9 と同じく，青線が 2-anonymity 化されたデータ，赤線が 2-concealment 化されたデータの分布を示している．2-concealment 化されたデータは誤差の分布が横に広がっており，誤差の平均値が大きい．

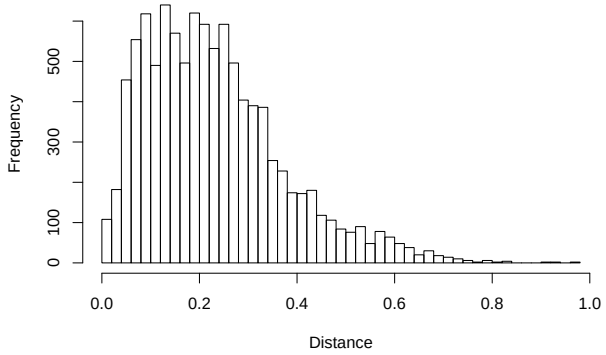

図 8　D_{100} の 2 顧客間の距離の分布

D_{100} を 2-anonymity 化したデータと 2-concealment 化されたデータを，実際の位置情報に投影して可視化した結果の一部をそれぞれ図 11,12 に示す．図中の黒点は元の座標情報を示しており，赤四角は一般化された範囲を示している．2-anonymity の方は独立した 2 点が一般化されているが，2-concealment の方は 2 つの四角（一般化された範囲）が 1 つの点を共有している場合がある．しかし，一般化後にできる四角形が大きいほど有用性は低くなり，図 12 の方が明らかに四角形の面積は大きい．

この実験では k-concealment の有用性は k-anonymity よりも低いという結果が出たが，マッチングの方法や 2 部グラフの作り方に改良の余地が残っている．

5.　おわりに

本稿では**動的データを「最低でも k 人の区別がつかない」状態にするためにはレコードの追加・削除や顧客の削除が必須なのだろうか？**という問題を解決するために，「仮名の一般化」と「レコード間の k-concealment」を用いた動的データの k-concealment 化手法を提案した．我々の提案手法を用いることにより，顧客やレコードの削除・追加をすることなく，動的データを「最低でも k 人の区別がつかない」状態にすることが可能である．

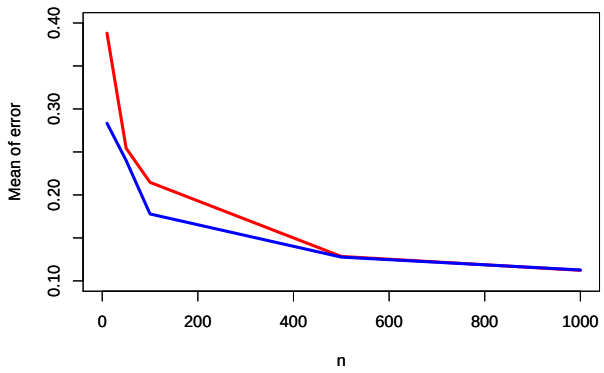

図 9　2-anonymity 化されたデータと 2-concealment 化されたデータの顧客数 n と有用性の関係

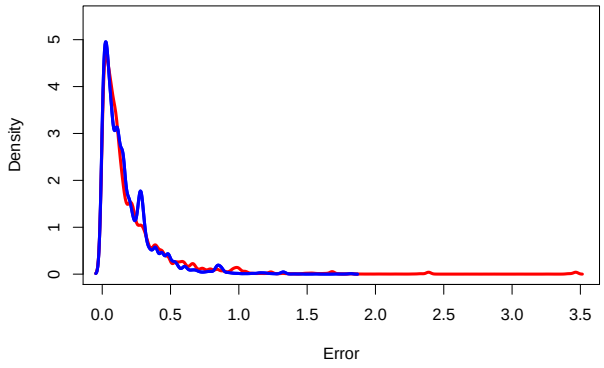

図 10　2-anonymity 化されたデータと 2-concealment 化されたデータの誤差の分布

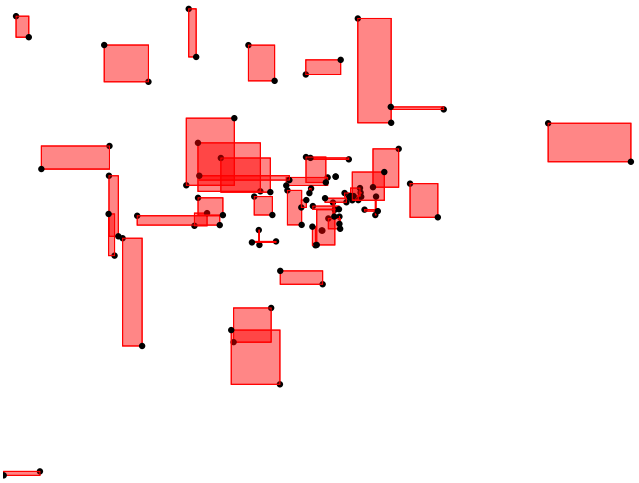

図 11　2-anonymity 化された D_{100} の可視化

また，ナイトレイ社から公開されている疑似人流データを用いて，提案する手法の簡易版の実装をし，その手法の評価をした．k-anonymity と比較すると k-concealment 化されたデータの有用性は低い．しかしながら，マッチングや 2 部グラフ作成のアルゴリズムの改善を今後の課題とする．

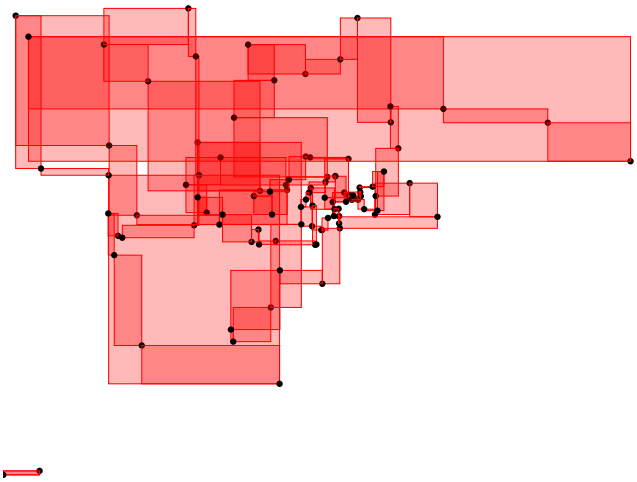

図 12　2-concealment 化された D_{100} の可視化

参考文献

[1] L. Sweeny, “k-anonymity: a model for protecting privacy”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), pp.557–570. (2006)

[2] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “l-Diversity: Privacy beyond k-anonymity”, In International Conference on Data Engineering (ICDE), page 24. (2006)

[3] T. Truta, A. Campan, and P. Meyer, “Generating microdata with p-sensitive k-anonymity property”, In Secure Data Management (SDM), pages 124–141. (2007)

[4] Tamir Tassa, Arnon Mazza, Aristides Gionis, “k-Concealment: An Alternative Model of k-Type Anonymity”, TRANSACTIONS ON DATA PRIVACY 5, pp. 189–222. (2012)

[5] 濱田浩気，荒井ひろみ，小栗秀暢，菊池浩明，黒政敦史，中川裕志，西山賢志郎，波多野卓磨，村上隆夫，山岡裕司，山田明，渡辺知恵美，「PWS Cup 2018: 匿名加工再識別コンテストの設計 ～履歴データの一般化・再識別～」，コンピューターセキュリティシンポジウム（CSS 2018），pp.935–940．(2018)

[6] 高橋磐郎，藤重悟，「離散数学」，岩波講座情報科学 17，pp.147–162．(1981)

[7] 疑似人流データ，ナイトレイ社，`https://nightley.jp/archives/1954/`，2019 年 8 月 22 日参照．