**CSS-2019**

# 履歴データに対する匿名化手法$k$-concealmentの改良手法の提案

**伊藤聡志(明治大学大学院)**

**菊池浩明(明治大学)**

# 研究概要

- **データにはレコード数と顧客数が等しい静的データと，レコード数が顧客数より多い動的データがある**

- **これまで，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，ダミーレコードの追加やレコード削除が必要だと考えられていた**

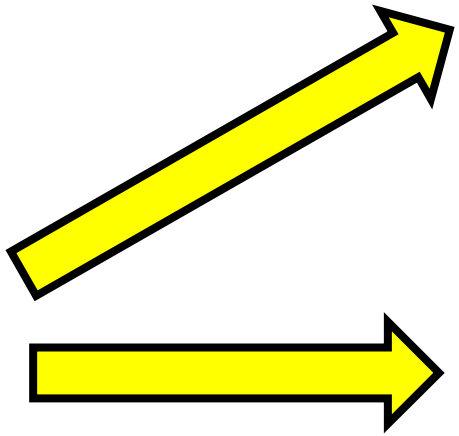
- **本研究では，レコードの追加や削除をせずに，動的データを「最低でも$k$人の区別がつかない」状態にする手法を提案する**

# 既存研究：$k$-anonymity

Sweeneyによって提案された匿名性指標
ある公開データにおいて，最低でも$k$人の
区別がつかない(≒データが等しい)とき，
そのデータは$k$-anonymityを持つという．

元データ

| 名前 | 年齢 | 郵便番号 |
|---|---|---|
| Alice | 30 | 10055 |
| Bob | 21 | 10055 |
| Carol | 21 | 10023 |
| David | 55 | 10165 |
| Eve | 47 | 10224 |



加工データ

| 仮名 | 年齢 | 郵便番号 |
|---|---|---|
| 1 | 21-55 | 10*** |
| 2 | 21-55 | 10*** |
| 3 | 21-55 | 10*** |
| 4 | 21-55 | 10*** |
| 5 | 21-55 | 10*** |

5-anonymityを持つ

| 仮名 | 年齢 | 郵便番号 |
|---|---|---|
| 1 | 21-30 | 100** |
| 2 | 21-30 | 100** |
| 3 | 21-30 | 100** |
| 4 | 47-55 | 10*** |
| 5 | 47-55 | 10*** |

2-anonymityを持つ

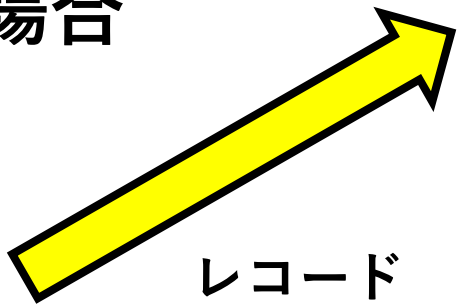
# $k$-anonymityの問題点

**$k$の値によってはレコードの追加/削除が必要になる**

**例：顧客5人のデータを**
**3-匿名化したい場合**

元データ

| 名前 | 年齢 | 郵便番号 |
|---|---|---|
| Alice | 30 | 10055 |
| Bob | 21 | 10055 |
| Carol | 21 | 10023 |
| David | 55 | 10165 |
| Eve | 47 | 10224 |



レコード
削除

ダミーレコード
追加

| 仮名 | 年齢 | 郵便番号 |
|---|---|---|
| 1 | 21-30 | 100** |
| 2 | 21-30 | 100** |
| 3 | 21-30 | 100** |

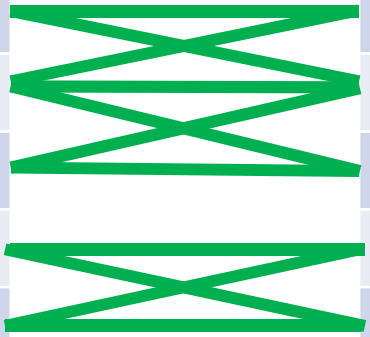
| 仮名 | 年齢 | 郵便番号 |
|---|---|---|
| 1 | 21-30 | 100** |
| 2 | 21-30 | 100** |
| 3 | 21-30 | 100** |
| 4 | 47-55 | 10*** |
| 5 | 47-55 | 10*** |
| 6 | 47-55 | 10*** |

# 既存研究：$k$-concealment

**2012年にTamirによって提案された匿名性指標**
**元データと加工データのレコード関係を2部グラフに表し，元データの各レコードが，加工データとの間に少なくとも$k$種類の完全マッチングの辺を持つとき，加工データは$k$-concealmentを持つ．　※完全マッチング＝攻撃者の回答**

元データ

| 名前 | 年齢 | 郵便番号 |
|---|---|---|
| Alice | 30 | 10055 |
| Bob | 21 | 10055 |
| Carol | 21 | 10023 |
| David | 55 | 10165 |
| Eve | 47 | 10224 |



加工データ

| 仮名 | 年齢 | 郵便番号 |
|---|---|---|
| 1 | 21-30 | 10055 |
| 2 | 21-30 | 100** |
| 3 | 21 | 100** |
| 4 | 47-55 | 10*** |
| 5 | 47-55 | 10*** |

2-concealmentを満たす

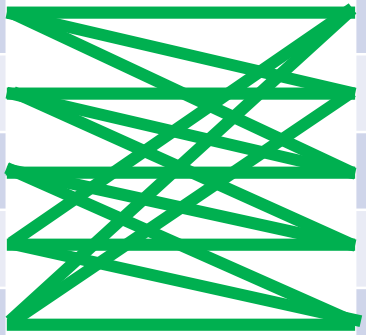
# $k$-concealmentの特徴 1

**任意の$k$を選べるようになる**

**例：顧客5人のデータの3-concealment化**

元データ

| 名前 | 年齢 | 郵便番号 |
|---|---|---|
| Alice | 30 | 10055 |
| Bob | 21 | 10055 |
| Carol | 21 | 10023 |
| David | 55 | 10165 |
| Eve | 47 | 10224 |



加工データ

| 仮名 | 年齢 | 郵便番号 |
|---|---|---|
| 1 | 30-55 | 10*** |
| 2 | 21-47 | 10*** |
| 3 | 21-30 | 100** |
| 4 | 21-55 | 10*** |
| 5 | 21-55 | 10*** |

**データ中の$k$人の区別をつかなくするために，**
**レコードの削除/追加を行う必要がなくなる！**

# $k$-concealmentの特徴 2

**$k$人の区別がつかない状態を保ちつつ，$k$-匿名化されたデータより有用性が高いデータを作れる**

元データ

| 名前 | 年齢 |
|---|---|
| Alice | 30 |
| Bob | 21 |
| Carol | 21 |
| David | 55 |
| Eve | 47 |

2-anonymityを満たしている加工データ

| 仮名 | 年齢 |
|---|---|
| 1 | 21-30 |
| 2 | 21-30 |
| 3 | 21-30 |
| 4 | 47-55 |
| 5 | 47-55 |

2-concealmentを満たしている加工データ

| 仮名 | 年齢 |
|---|---|
| 1 | 21-30 |
| 2 | 21-30 |
| 3 | 21 |
| 4 | 47-55 |
| 5 | 47-55 |

どちらも「少なくとも2人の区別がつかないデータ」だが，
右のほうが無加工の1セル分有用性が高い

# 動的データに対する$k$-concealment 1

**$k$-concealmentは静的データ(レコード数$m$＝顧客数$n$)にしか対応しておらず，動的データ(レコード数$m$>顧客数$n$)には未対応である**

静的データ

| 名前 | 年齢 | 郵便番号 |
|---|---|---|
| Alice | 30 | 10055 |
| Bob | 21 | 10055 |
| Carol | 21 | 10023 |

$n = 3, m = 3$

動的データ

| 名前 | 日付 | 商品 |
|---|---|---|
| Alice | 12/1 | 2 |
| Bob | 12/2 | 3 |
| Alice | 12/2 | 1 |
| Carol | 12/2 | 2 |
| Bob | 12/3 | 1 |
| Bob | 12/4 | 4 |

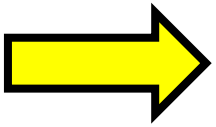
$n = 3, m = 6$

# 動的データに対する$k$-concealment 2

**動的データの複数の顧客の区別がつかなくするためには，**
**大量のレコードを追加/削除する必要がある！**
**例：PWSCUP-2018ではレコード数の多い上位2顧客の区別を**
**つかなくするために，2,688レコードの追加/削除が必要**

| 名前 | 日付 | 商品 |
|---|---|---|
| Bob | 12/2 | 3 |
| Bob | 12/3 | 1 |
| Bob | 12/4 | 4 |

| 名前 | 日付 | 商品 |
|---|---|---|
| Carol | 12/2 | 2 |



| 仮名 | 日付 | 商品 |
|---|---|---|
| 2 | 12/2 | {2, 3} |
| 2 | 12/3 | 1 |
| 2 | 12/4 | 4 |

| 仮名 | 日付 | 商品 |
|---|---|---|
| 3 | 12/2 | {2, 3} |
| 3 | 12/3 | 1 |
| 3 | 12/4 | 4 |

ダミー
レコード

# ここまでのまとめと研究目的

**問題点**

- **$k$-anonymityの$k$の値の制約等を改善した $k$-concealmentは，動的データに対応していない**
- **動的データを$k$人の区別がつかない状態に加工しようとすると，大量のレコード追加/削除が必要**

**研究目的**

- **レコードの追加や削除なしに，動的データを $k$人の区別がつかない状態に加工する手法の提案**

**解決手法**

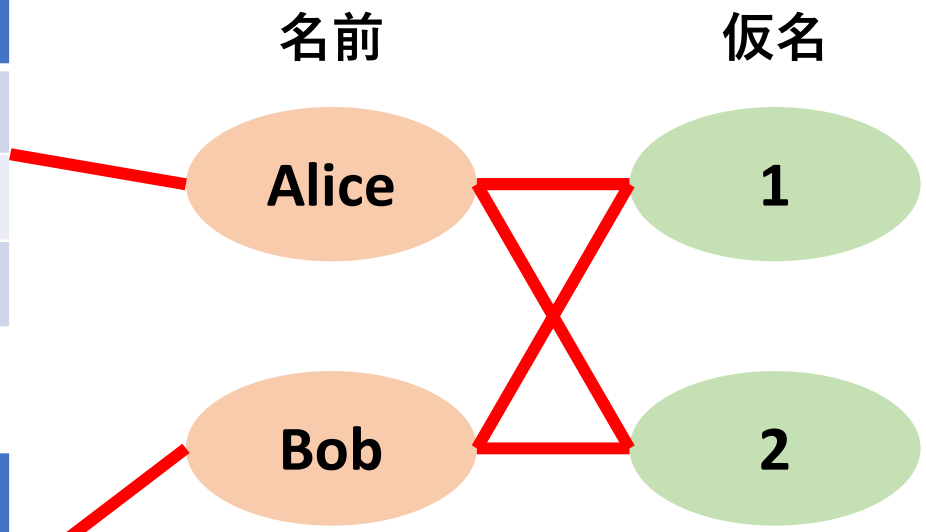
- **2重$k$-concealment化 + 仮名の一般化**

# 解決手段：2重$k$-concealment 1

**動的データは静的データと違い，顧客ごとにレコード数が異なる場合があるので，顧客の$k$-concealment化だけでは不十分**

| 名前 | 日付 | 商品 |
|---|---|---|
| Alice | 12/2 | 3 |
| Alice | 12/3 | 1 |
| Alice | 12/4 | 4 |

| 名前 | 日付 | 商品 |
|---|---|---|
| Bob | 12/1 | 2 |
| Bob | 12/2 | 5 |



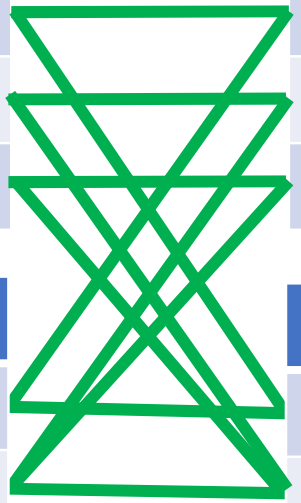

**2-concealment**
**(少なくとも顧客2人の区別がつかない)**

# 解決手段：2重$k$-concealment 2

**もしレコードについても$k$-concealment化することができれば**
**動的データを$k$-concealment化することができる**

| 名前 | 日付 | 商品 |
|---|---|---|
| Alice | 12/2 | 3 |
| Alice | 12/3 | 1 |
| Alice | 12/4 | 4 |

| 名前 | 日付 | 商品 |
|---|---|---|
| Bob | 12/1 | 2 |
| Bob | 12/2 | 5 |



| 仮名 | 日付 | 商品 |
|---|---|---|
| 1 | 12/1-12/2 | {2,3} |
| 1 | 12/2-12/3 | {1,2} |
| 1 | 12/2-12/4 | {4,5} |

| 仮名 | 日付 | 商品 |
|---|---|---|
| 2 | 12/1-12/2 | {2,3} |
| 2 | 12/2-12/4 | {1,4,5} |

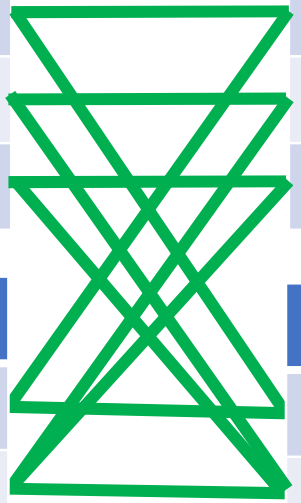
**2-concealment**
**(少なくともレコード2つの区別がつかない)**

# 解決手段：仮名の一般化

**仮名を一般化することにより，レコードの削除や追加をせずに動的データの複数の顧客の区別がつかないようにできる**

| 名前 | 日付 | 商品 |
|---|---|---|
| Alice | 12/2 | 3 |
| Alice | 12/3 | 1 |
| Alice | 12/4 | 4 |

| 名前 | 日付 | 商品 |
|---|---|---|
| Bob | 12/1 | 2 |
| Bob | 12/2 | 5 |



| 仮名 | 日付 | 商品 |
|---|---|---|
| 1 | 12/1-12/2 | {2,3} |
| 1 | 12/2-12/3 | {1,2} |
| 1,2 | 12/2-12/4 | {4,5} |

| 仮名 | 日付 | 商品 |
|---|---|---|
| 2 | 12/1-12/2 | {2,3} |
| 2 | 12/2-12/4 | {1,4,5} |

**2-concealment**
**(少なくともレコード2つの区別がつかない)**

# 0. 準備

**用意するもの**

- **パラメータ$k(\leq n)$**
- **元データ$T$とそれを仮名化したデータ$T'$**

**元データ$T$**

| 名前 | 日付 | 商品 |
|---|---|---|
| Alice | 12/2 | B |
| Alice | 12/3 | E |
| Alice | 12/4 | F |
| Bob | 12/1 | A |
| Bob | 12/2 | C |
| Carol | 12/2 | D |

**加工データ$T'$**

| 名前 | 日付 | 商品 |
|---|---|---|
| 1 | 12/2 | B |
| 1 | 12/3 | E |
| 1 | 12/4 | F |
| 2 | 12/1 | A |
| 2 | 12/2 | C |
| 3 | 12/2 | D |

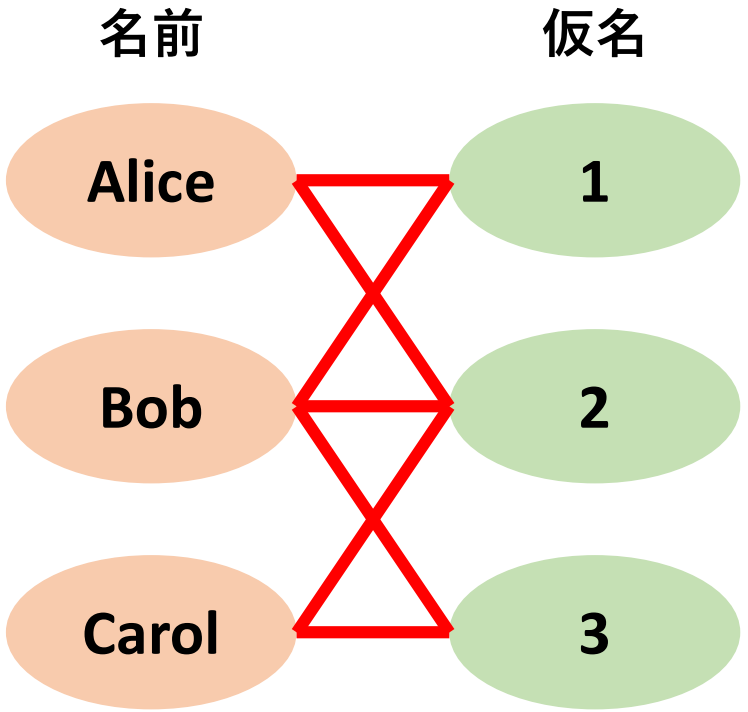
**パラメータ**
$k = 2$

# 1. 顧客の$k$-concealment化

**$T$の顧客と$T'$の仮名の間に辺を張り，$k$-concealmentを満たすような対称な2部グラフ$G$を作成する**

元データ$T$

| 名前 | 日付 | 商品 |
|---|---|---|
| Alice | 12/2 | B |
| Alice | 12/3 | E |
| Alice | 12/4 | F |
| Bob | 12/1 | A |
| Bob | 12/2 | C |
| Carol | 12/2 | D |



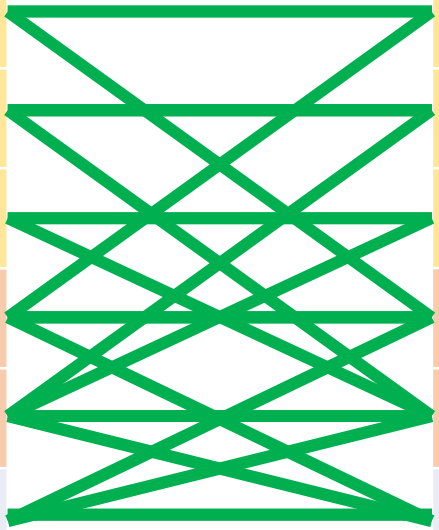

顧客についての2部グラフ$G$
(2-concealmentを満たす)

# 2. レコードの$k$-concealment化

**$T$のレコードと$T'$のレコードの間に辺を張り，$G$と矛盾せず$k$-concealmentを満たすような2部グラフ$H$を作る**

元データ$T$

| 名前 | 日付 | 商品 |
|---|---|---|
| Alice | 12/2 | B |
| Alice | 12/3 | E |
| Alice | 12/4 | F |
| Bob | 12/1 | A |
| Bob | 12/2 | C |
| Carol | 12/2 | D |



加工データ$T'$

| 名前 | 日付 | 商品 |
|---|---|---|
| 1 | 12/2 | B |
| 1 | 12/3 | E |
| 1 | 12/4 | F |
| 2 | 12/1 | A |
| 2 | 12/2 | C |
| 3 | 12/2 | D |

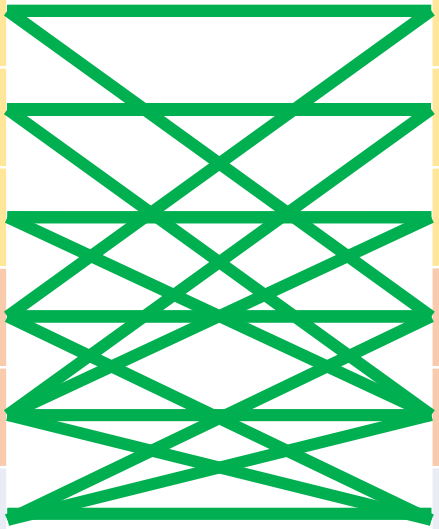
**レコードについての2部グラフ$H$**
**(2-concealmentを満たす)**

# 3. データの一般化

**$H$に矛盾しないようにデータを一般化する**

**元データ$T$**

| 名前 | 日付 | 商品 |
|---|---|---|
| Alice | 12/2 | B |
| Alice | 12/3 | E |
| Alice | 12/4 | F |
| Bob | 12/1 | A |
| Bob | 12/2 | C |
| Carol | 12/2 | D |



**加工データ$T'$**

| 名前 | 日付 | 商品 |
|---|---|---|
| 1 | [12/1-12/2] | {A, B} |
| 1 | [12/2-12/3] | {C, E} |
| 1 | [12/2-12/4] | {C, F} |
| 2 | [12/1-12/2] | {A, B, D} |
| 2 | [12/2-12/4] | {C, D, E, F} |
| 3 | [12/1-12/2] | {A, C, D} |

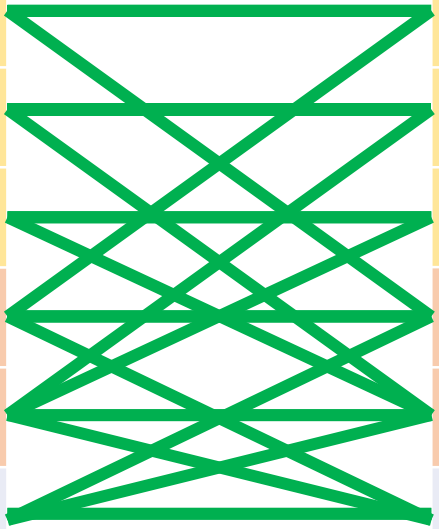
**レコードについての2部グラフ$H$**
**(2-concealmentを満たす)**

# 4. 仮名の一般化

## $T'$の一部のレコードの仮名を一般化する

元データ$T$

| 名前 | 日付 | 商品 |
|---|---|---|
| Alice | 12/2 | B |
| Alice | 12/3 | E |
| Alice | 12/4 | F |
| Bob | 12/1 | A |
| Bob | 12/2 | C |
| Carol | 12/2 | D |



加工データ$T'$

| 名前 | 日付 | 商品 |
|---|---|---|
| 1 | [12/1-12/2] | {A, B} |
| 1 | [12/2-12/3] | {C, E} |
| 1,2 | [12/2-12/4] | {C, F} |
| 2,3 | [12/1-12/2] | {A, B, D} |
| 2 | [12/2-12/4] | {C, D, E, F} |
| 3 | [12/1-12/2] | {A, C, D} |

レコードについての2部グラフ$H$
(2-concealmentを満たす)

# $k$-concealmentの安全性

元データ

| 名前 | 年齢 |
|---|---|
| Alice | 30 |
| Bob | 21 |
| Carol | 21 |
| David | 55 |
| Eve | 47 |

2-anonymityを満たしている加工データ

| 仮名 | 年齢 |
|---|---|
| 1 | 21-30 |
| 2 | 21-30 |
| 3 | 21-30 |
| 4 | 47-55 |
| 5 | 47-55 |

2-concealmentを満たしている加工データ

| 仮名 | 年齢 |
|---|---|
| 1 | 21-30 |
| 2 | 21-30 |
| 3 | 21 |
| 4 | 47-55 |
| 5 | 47-55 |

識別される人数の期待値

$$\frac{1}{3} * 3 + \frac{1}{2} * 2 = 2$$

$$\frac{1}{3} * 2 + \frac{1}{2} * 1 + \frac{1}{2} * 2 = \frac{13}{6}$$

# 提案手法の安全性

**レコードの持ち主が確定してしまう場合がある**

**加工データ$T'$**

| 名前 | 日付 | 商品 |
|---|---|---|
| 1 | [12/1-12/2] | {A, B} |
| 1 | [12/2-12/3] | {C, E} |
| 1,2 | [12/2-12/4] | {C, F} |
| 2,3 | [12/1-12/2] | {A, B, D} |
| 2 | [12/2-12/4] | {C, D, E, F} |
| 3 | [12/1-12/2] | {A, C, D} |

**回答パターン**

| 1 | 2 | 3 |
|---|---|---|
| A | B | A |
| A | B | A |
| A | A | A |
| B | A | B |
| B | A | C |
| C | C | B |

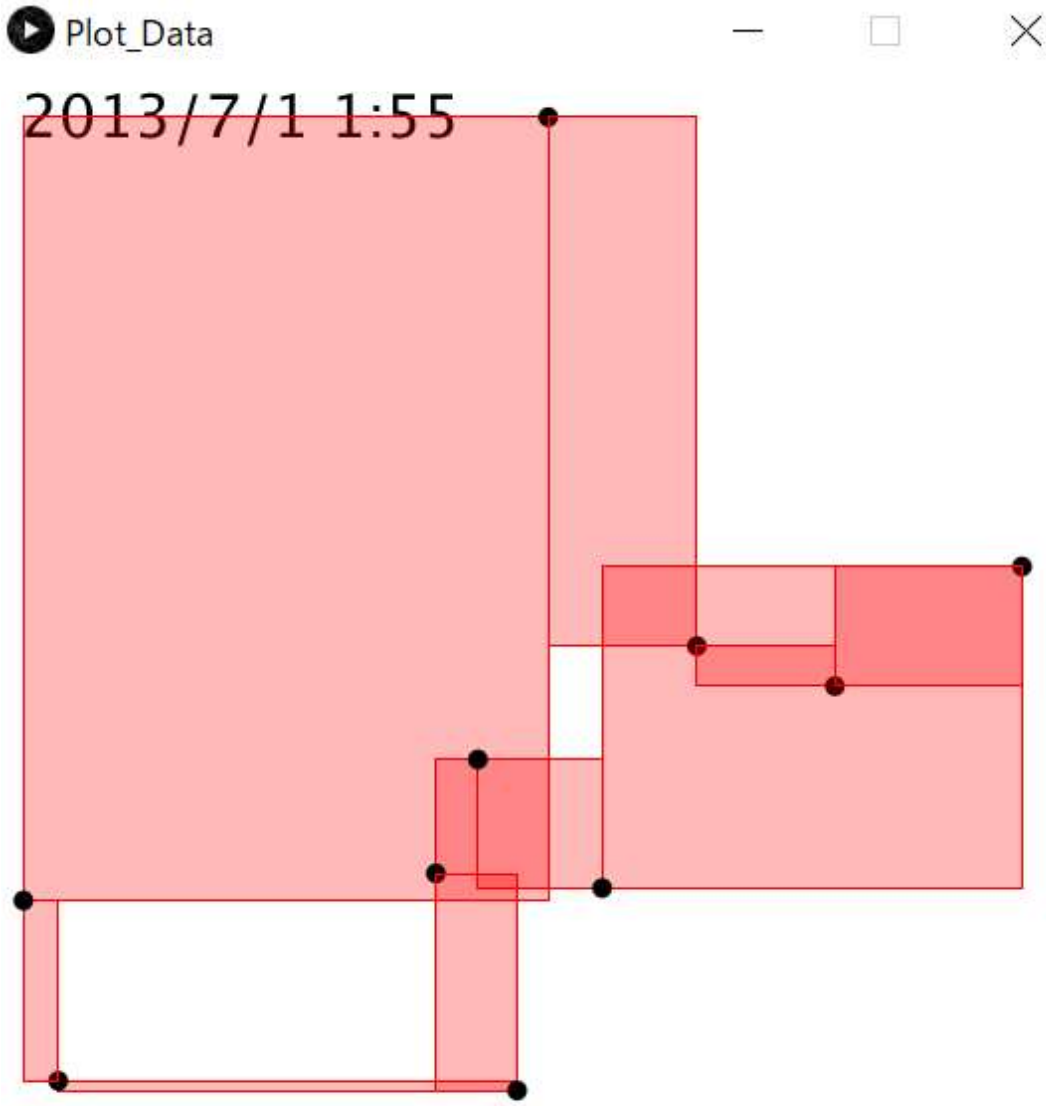
# おまけ

# 一般化された位置情報の可視化

**一般化された位置情報を可視化した**

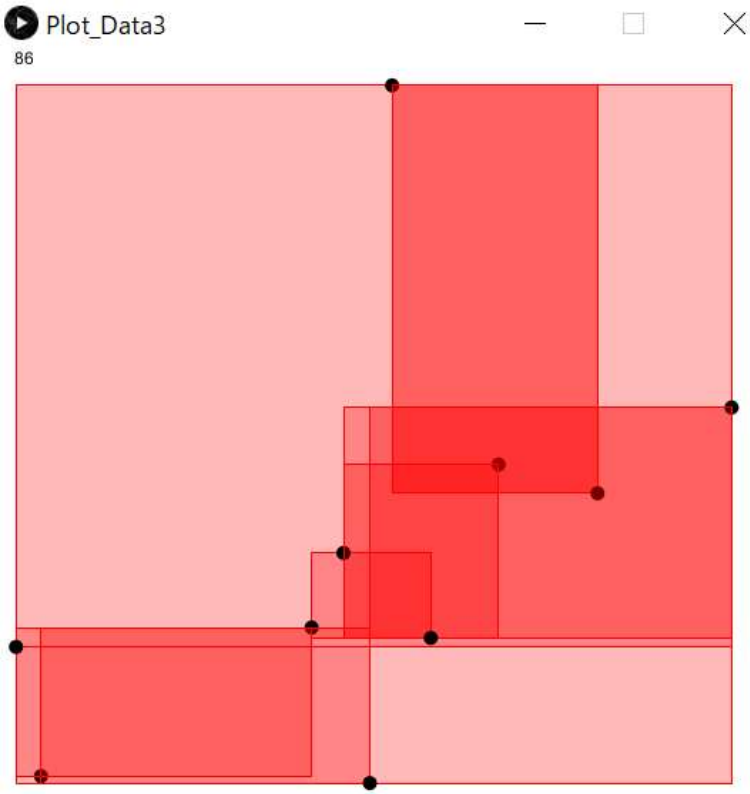
**例：右図**
**顧客数10の人流データを2-concealment化した場合**

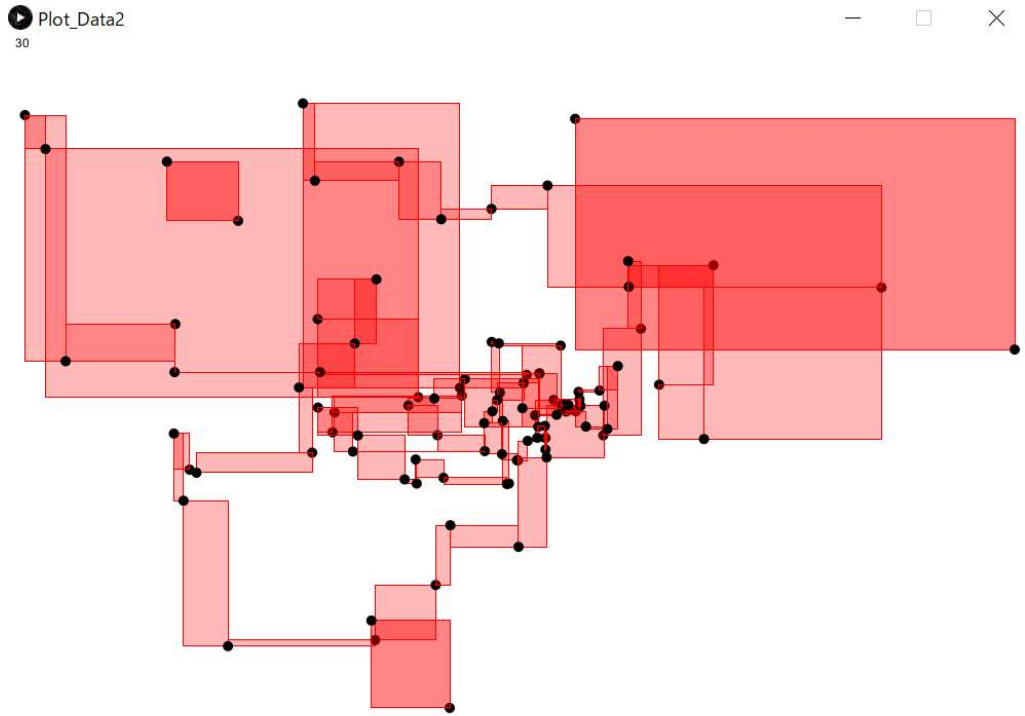
**黒点　：元情報**
**赤四角：加工情報**



# 一般化された軌跡情報の可視化 1

**$k$-concealmentされたデータの可視化**

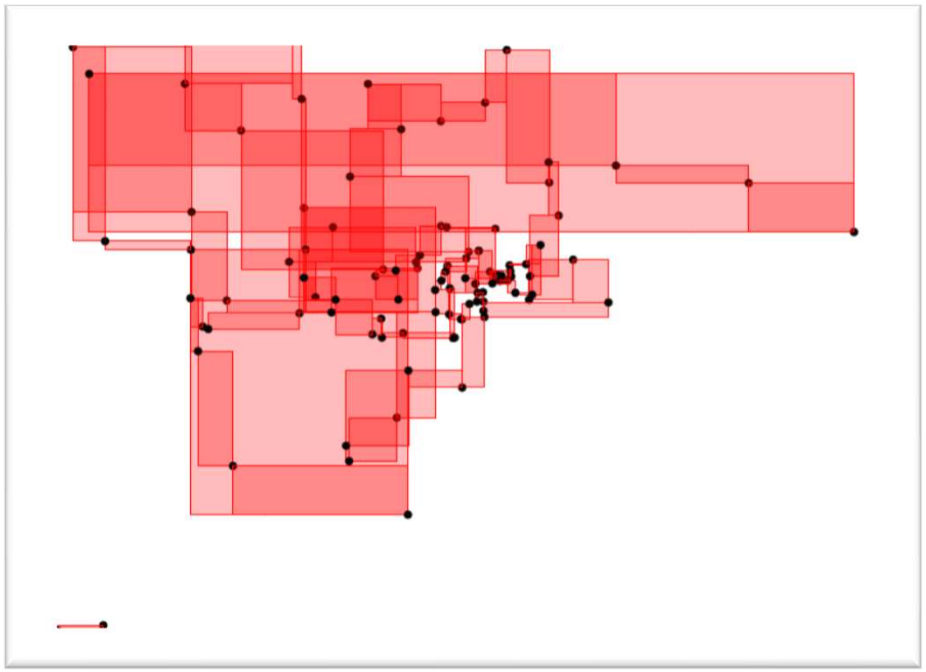


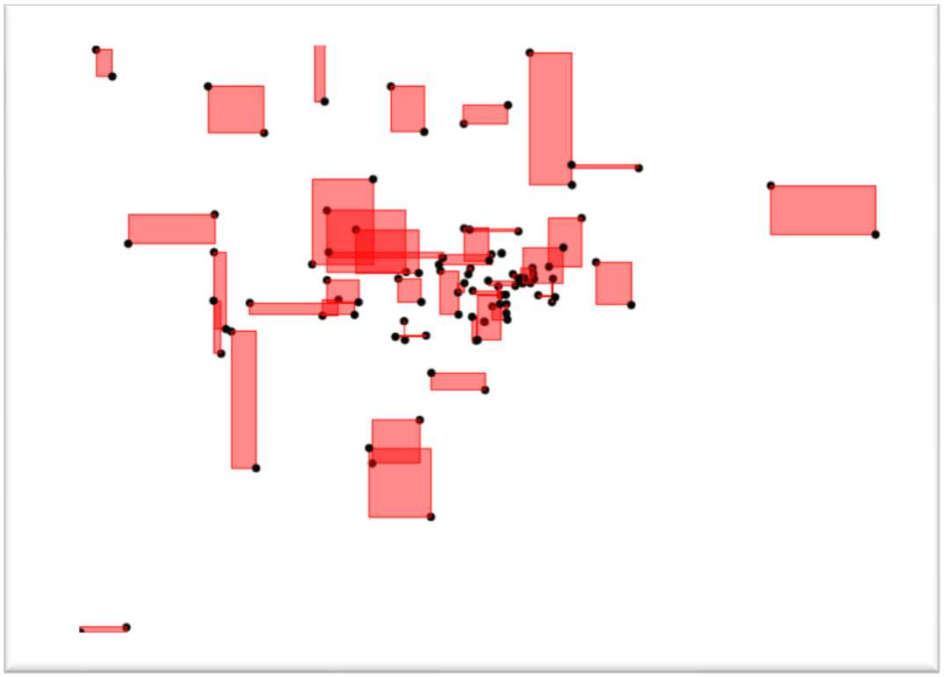

$n = 10, k = 3$

$n = 100, k = 2$

# 一般化された軌跡情報の可視化 2



$D_{100}$，2-concealment



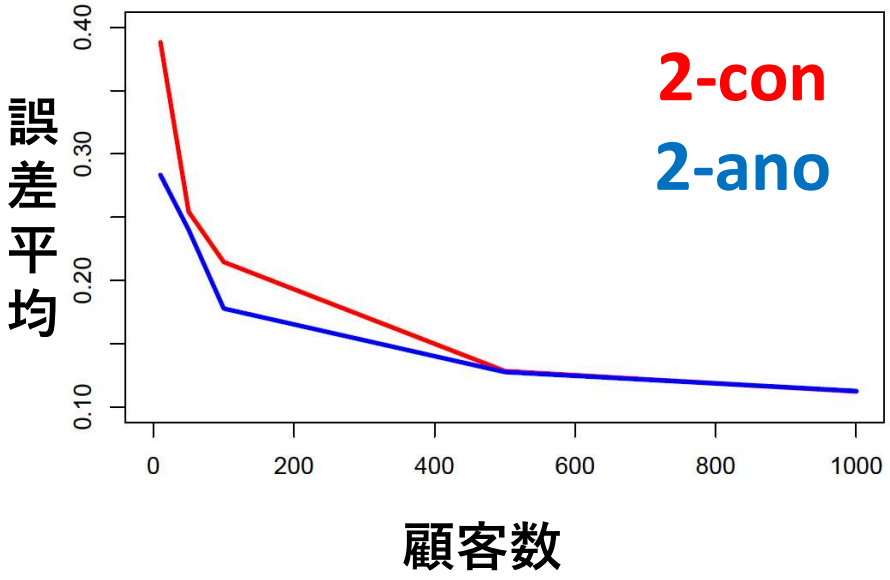
$D_{100}$，2-anonymity

**※どちらも「顧客2人の区別がつかない」データなので，**
**四角の面積(有用性)が小さいほうがよい**

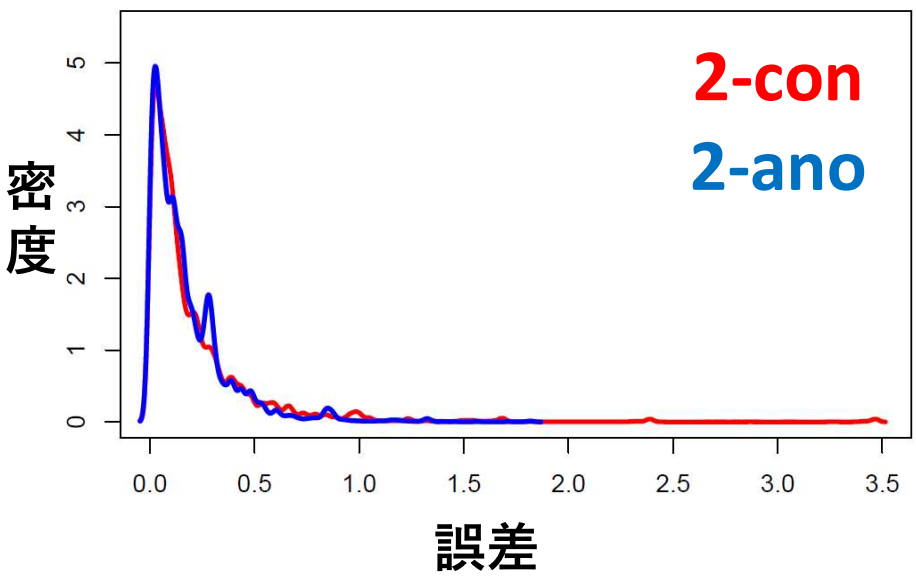
# $k$-anonymity VS $k$-concealment

2-anonymity化したデータと2-concealment化したデータの有用性(ユークリッド距離による誤差)を比較してみた

顧客数と有用性の関係




$D_{100}$の誤差の分布



# まとめ・今後の課題

**まとめ**

- ***k*-anonymityの「*k*の値によってはレコード削除/追加が発生してしまう問題」は*k*-concealmentによって解決できるが，これは動的データには対応していない**
- **2重*k*-concealment化と仮名の一般化を行うことにより，レコードの追加や削除をすることなく動的データを*k*-concealment化する手法を提案した**
- **簡易的な*k*-concealmentを実装し，それによって加工したデータを評価する実験を行った**

**今後の課題**

- **提案手法とアルゴリズムの改善**
- **提案手法の実装と評価実験**