CSS 2021

完全k-concealment匿名化を求める精度の高いアルゴリズムの評価

伊藤聡志，菊池浩明
明治大学

研究背景（匿名化とk-anonymity）

匿名化：データから個人が識別されることを防ぐために情報を加工する技術.

k-anonymity：Sweeneyによって提案された匿名性指標. データ中の最低でもk人の区別がつかないとき，そのデータはk-anonymityを満たす.

k-anonymityを満たすようにデータを加工することはk-匿名化と呼ばれている.

元データ

名前	年齢	郵便番号
Alice	30	10055
Bob	25	10055
Carol	21	10023
David	55	10165
Eve	47	10224

2-匿名化された加工データ

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***

5-匿名化された加工データ

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***

「最低でも2人の区別がつかない」状態(2-匿名化)のためには
3人のグループは加工の無駄ではないか？

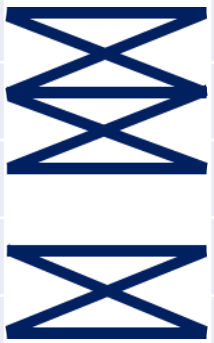
研究背景（k-concealment）

k-concealment：

2012年にTamirら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を満たす．（2部グラフ＝加工の設計図，完全マッチング＝攻撃者の回答）
データをk-concealmentを満たすように加工することをk-concealment化とする．

元データ

名前	年齢	郵便番号
Alice	30	10055
Bob	25	10055
Carol	21	10023
David	55	10165
Eve	47	10224



2-concealment化された
加工データ

仮名	年齢	郵便番号
1	25-30	10055
2	21-30	100**
3	21-25	100**
4	47-55	10***
5	47-55	10***

攻撃者の回答パターン

Ans1	Ans2	Ans3	Ans4	Ans5	Ans6
Alice	Bob	Alice	Alice	Bob	Alice
Bob	Alice	Carol	Bob	Alice	Carol
Carol	Carol	Bob	Carol	Carol	Bob
David	David	David	Eve	Eve	Eve
Eve	Eve	Eve	David	David	David

研究背景（k-anonymity VS k-concealment）

- k-匿名化されたデータとk-concealment化されたデータは，どちらも「最低でもk人の区別がつかない」という条件を満たしている．
- kが等しい場合，k-concealment化はk-匿名化よりも有用性が高いデータを作成することができる．
- k-concealment化では，レコード数以下の任意のkを選べる．

元データ

名前	年齢	郵便番号
Alice	30	10055
Bob	25	10055
Carol	21	10023
David	55	10165
Eve	47	10224

2-concealment化された加工データ

仮名	年齢	郵便番号
1	25-30	10055
2	21-30	100**
3	21-25	100**
4	47-55	10***
5	47-55	10***

2-匿名化された加工データ

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***

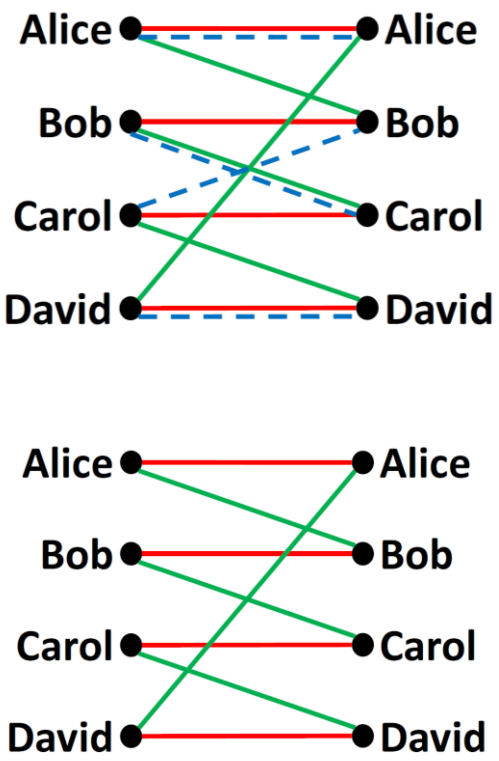
完全k-concealment化

k-匿名化の無駄を完全に解消するためには，k-concealment化を使って「全員等しくk人の区別がつかない」状態のデータを作成したい．
この加工を完全k-concealment化とする．

元データ

名前	年齢	性別
Alice	10	女
Bob	20	男
Carol	40	男
David	50	女

2部グラフ作成
(設計図作成)



2-concealment化

データ作成

完全
2-concealment化

仮名	年齢	性別
1	10-50	女
2	10-40	男,女
3	20-40	男
4	40-50	男,女

仮名	年齢	性別
1	10-50	女
2	10-20	男,女
3	20-40	男
4	40-50	男,女

本研究では，2部グラフの辺の距離の総和を加工コストと定義する

研究概要

リサーチクエスチョン

- **k-匿名化の無駄を完全に解消するためには,**
  どのように完全k-concealment化をしたらよいのか？

研究目的

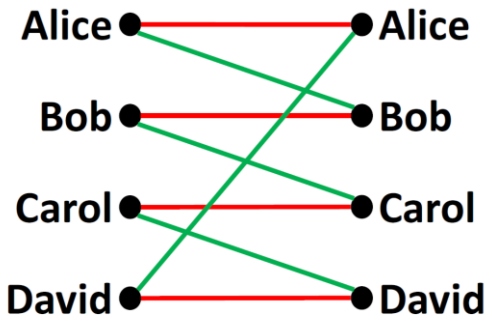
- **データを低コストで完全k-concealment化する**

解決手法

- **TSP解法やクラスタリングを用いたアルゴリズムを提案する**

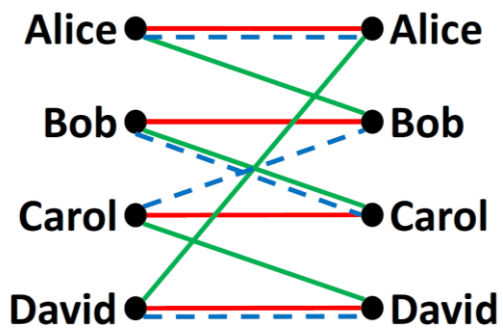
完全k-concealment化の方法

データを完全k-concealment化するためには,
2部グラフ内にk種類の「辺の重複しない完全マッチング」を作成すればよい.



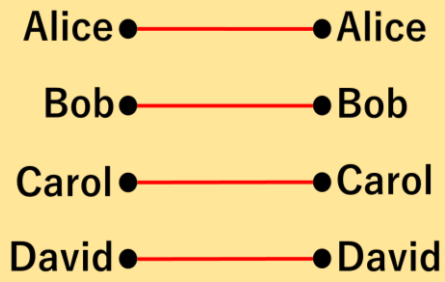
辺の重複しない2種類の完全マッチング
(赤マッチングと緑マッチング)
が2部グラフ内にあるため,
完全2-concealment化ができる

低コストの完全k-concealment化
||
辺が重複しないように
コストが小さい完全マッチングを
k種類作成する



3種類の完全マッチング
(赤,青,緑マッチング)
が2部グラフ内にあるが,
辺が重複しているので
完全3-concealment化にはならず
2-concealment化になる

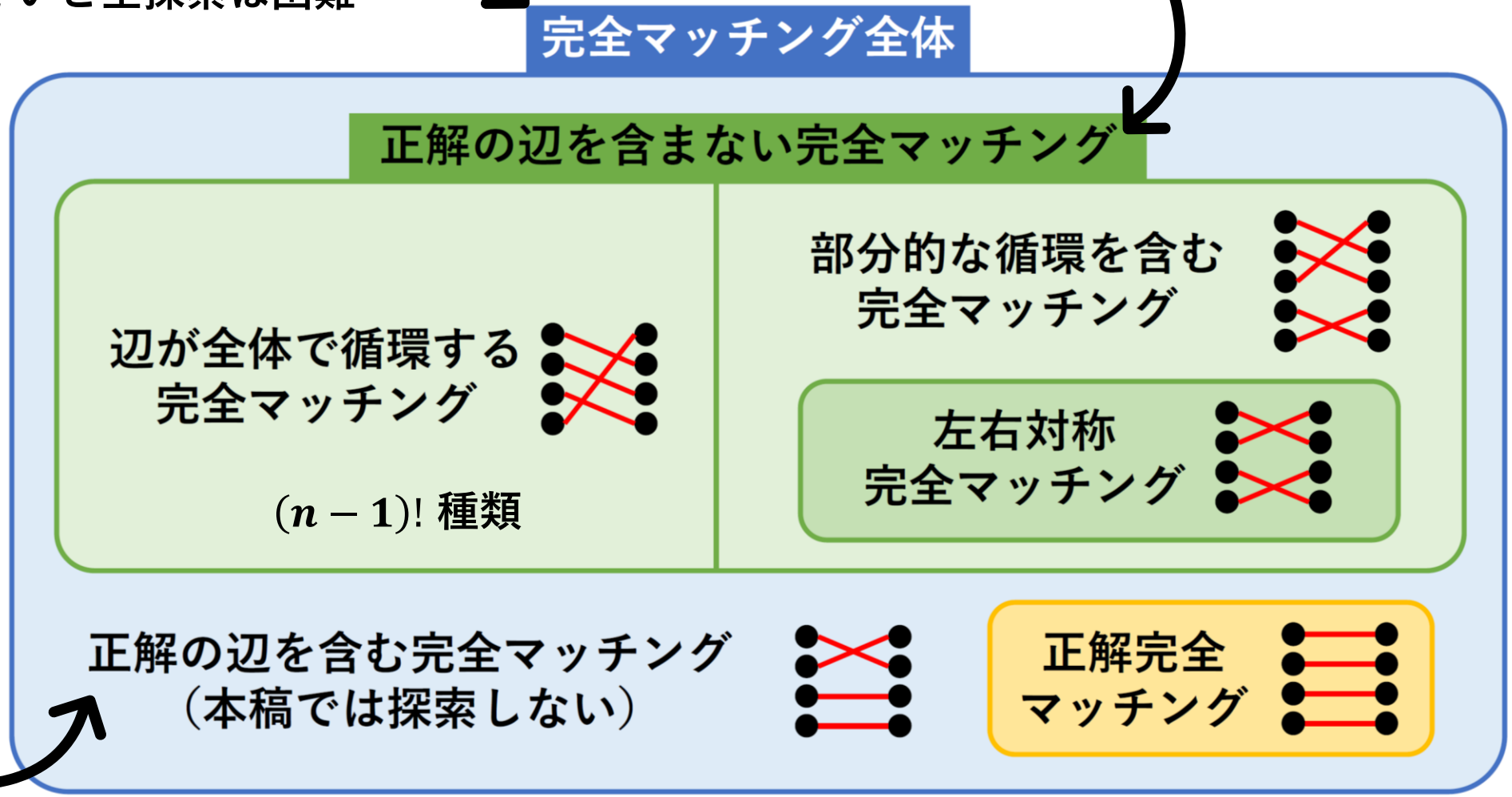
正解完全マッチング



完全マッチング探索の困難さ

完全マッチングは全部で $n!$ 種類
あるため，nが大きいと全探索は困難

本研究の提案手法で
探索するのはこの部分のみ

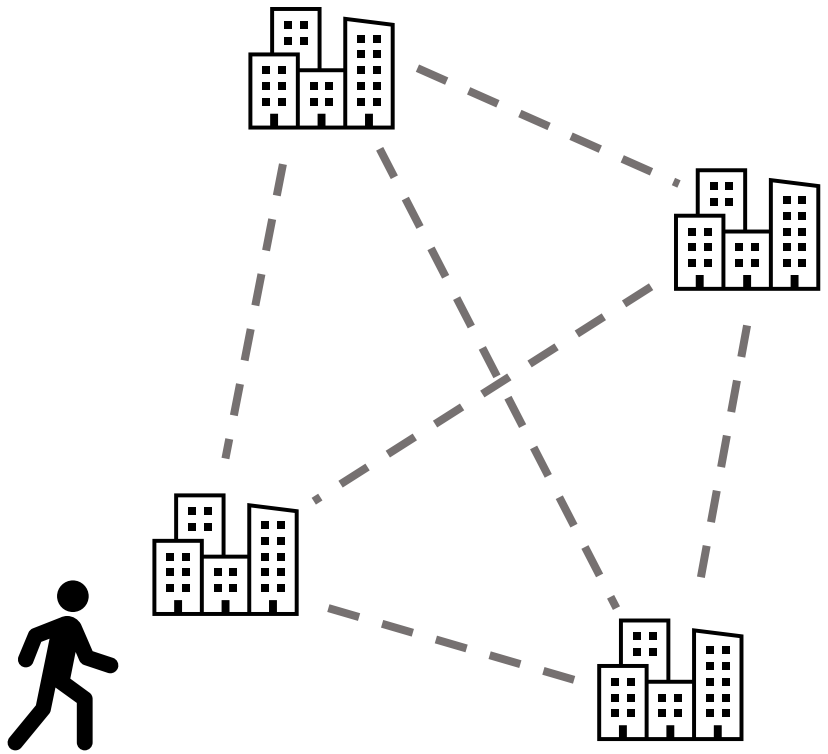


正解完全マッチングと
辺が重複するので
この部分は探索しない

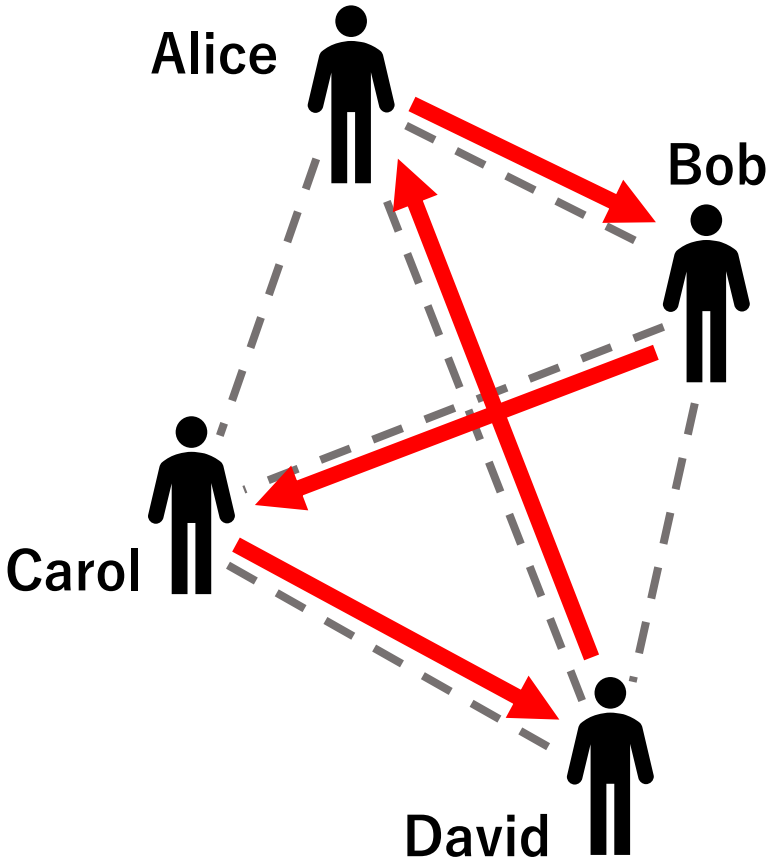
アイデア：TSP解法と完全マッチング探索

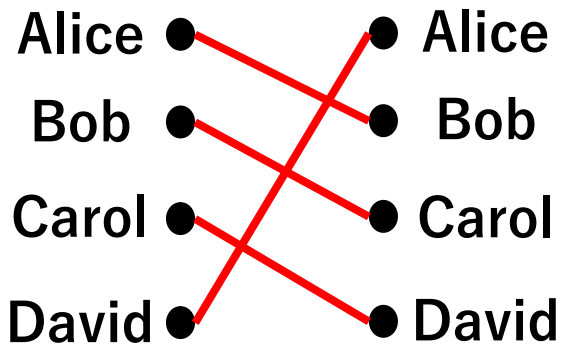
巡回セールスマン問題 (TSP)

セールスマンが全ての都市を1回ずつ巡回する場合の最短経路を求める問題



低コストの循環する完全マッチングなら，
TSPの解法で(高速に)求められるのではないか？





R言語のTSPパッケージを使って完全マッチングをk種類作成した

提案手法

辺が重複しないように辺の総和が小さい完全マッチングを
k種類作成する手法を，本研究では3つ比較する．

手法１：貪欲法
　　個人から最も近い個人へ辺を張って完全マッチングをk種類作る．
手法２：くじ引き法
　　ランダムに複数回完全マッチングを作成し，最もコストが低いものを
　　(辺が重複しないように) k種類採用する．
提案手法：TSP解法手法
　　巡回セールスマン問題の解法を使って完全マッチングをk種類作る．

また，手法２と提案手法にクラスタリングを組み合わせた結果も評価する．

データと評価実験の概要

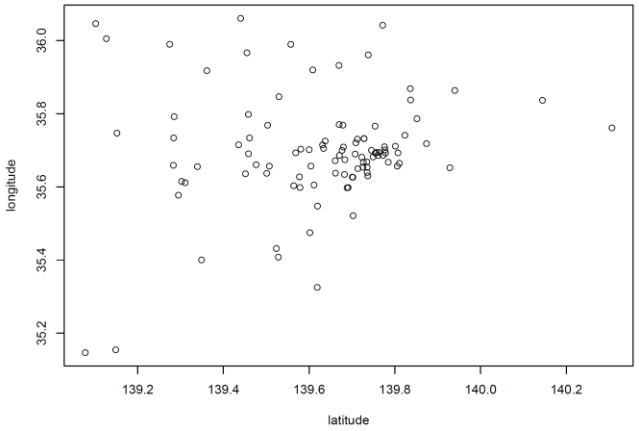
実験に用いるデータ

1. **疑似人流データ**
   → ナイトレイ社より公開されている位置情報データ
   ランダムに抽出した100人分の緯度経度のみ(0時0分)を用いる
2. **世帯収入データ**
   → UCIより公開されている32,561世帯の収入データ
   連続値6属性(年齢など)と離散値9属性(職種など)の計15属性

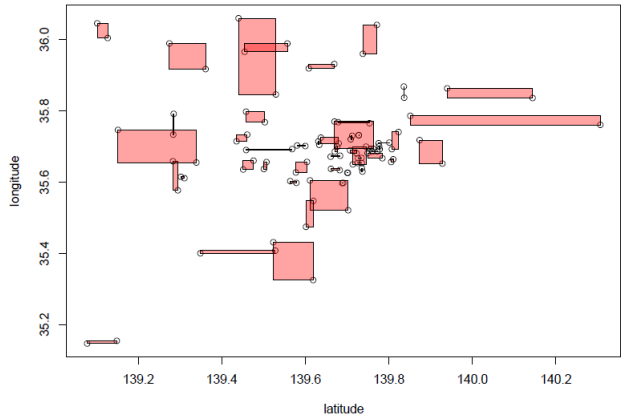
評価実験

1. **手法1,2,提案手法+クラスタリングで完全2-concealment化を行い,**
   貪欲法を用いた完全2-匿名化手法も加えた計6手法の比較を行う.
2. **手法2,提案手法+クラスタリングの計4手法で完全k-concealment化を**
   行い，kを変えた時の加工コストの変化を調べる.

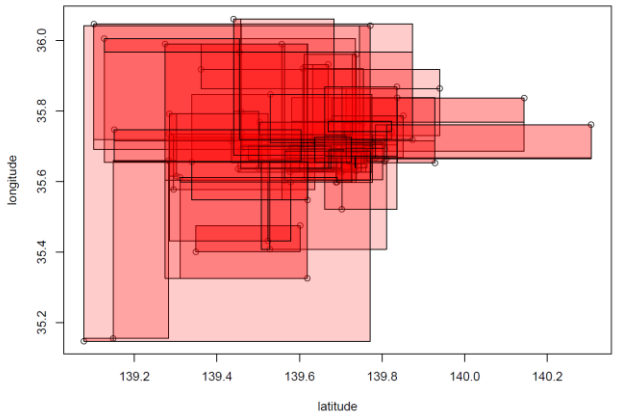
実験１結果（疑似人流データ，$k = 2$）



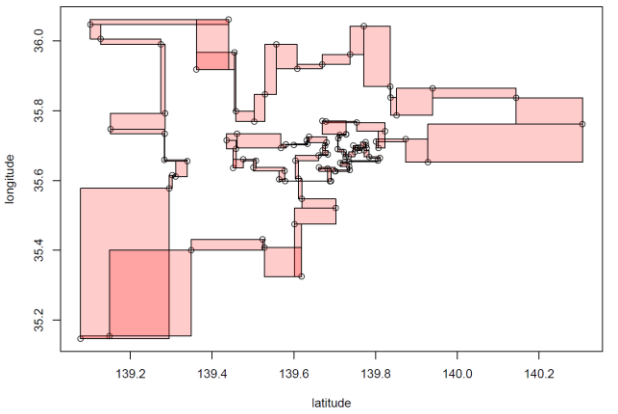
元データ



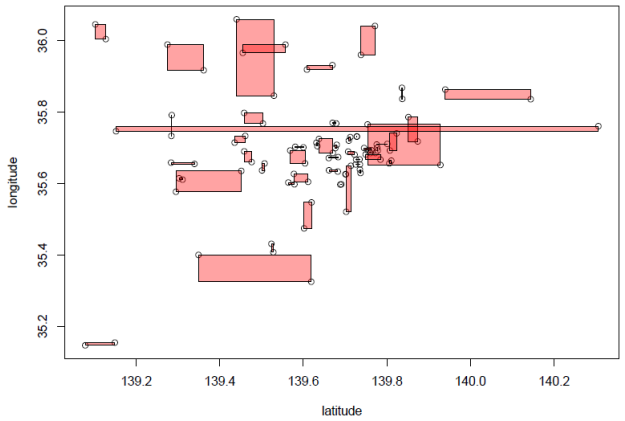
完全2-匿名化(貪欲)
コスト：0.2145



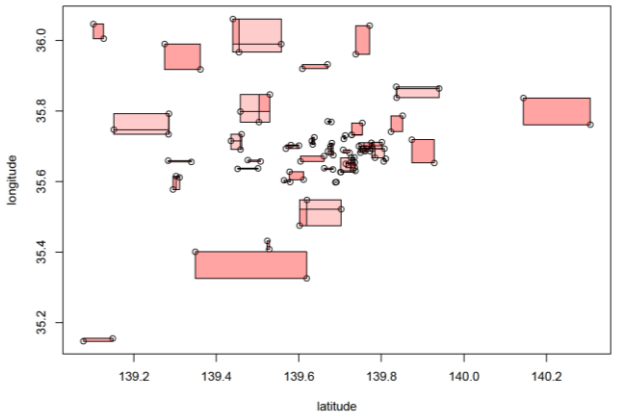
手法2(くじ引き法)
コスト：3.1490



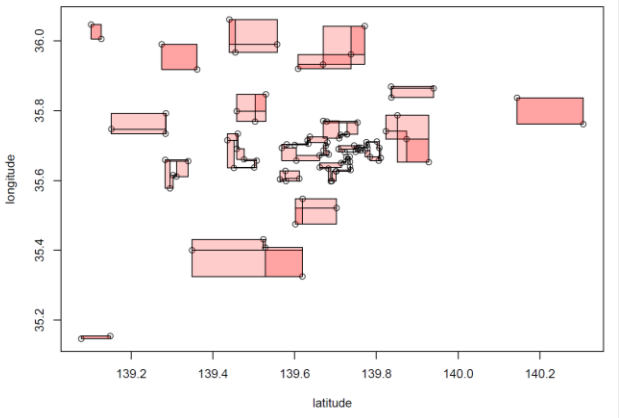
提案手法(TSP解法)
コスト：0.3186



手法1(貪欲法)
コスト：0.2328



手法2+クラスタリング
コスト：0.1493

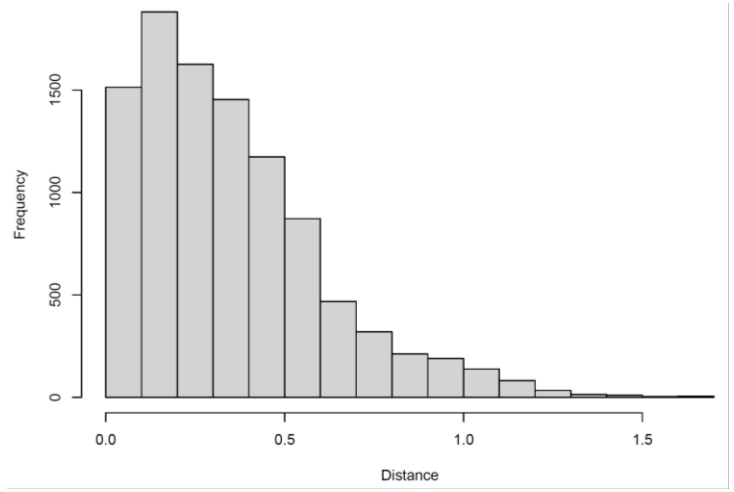


提案+クラスタリング
コスト：0.1640

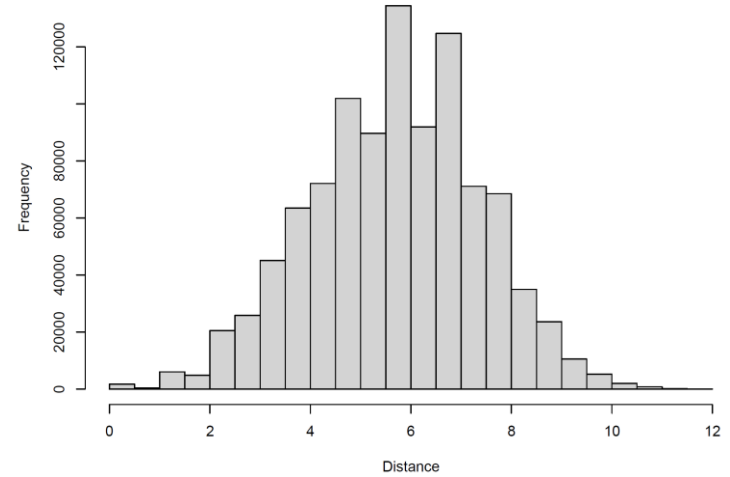
実験２結果（世帯収入データ，$k = 2, \ldots, 7$）

世帯収入データを完全k-concealment化したときの加工コスト

k	手法2 （くじ引き法）	手法2 ＋クラスタリング	提案手法 （TSP解法）	提案手法 ＋クラスタリング
2	5535.23	1819.31	1683.76	1718.13
3	11066.13	4083.07	3367.52	3443.10
4	16600.02	7413.35	5416.70	5656.75
5	22153.13	10041.37	7465.88	7858.74
6	27682.26	12741.39	9718.28	10364.58
7	33205.50	15973.04	11970.68	12826.53



疑似人流データの距離分布



世帯収入データの距離分布

世帯収入データの場合，提案手法3のみで加工したとき
最小コストの完全k-concealment化ができた

↓

個人間距離分布の違いによるものか？

まとめ

- Tamirはk-匿名化の加工の無駄を指摘し，それを改善するための指標k-concealmentを提案した.

- データを低コストで完全k-concealment化するためには，辺が重複しないように辺の総和が小さい完全マッチングをk種類作成すればよい.

- データを低コストで完全k-concealment化する手法として，TSP解法やクラスタリングを用いたアルゴリズムを提案した.

- 世帯収入データや疑似人流データを用いて評価実験を行った結果，データによって有効な手法が異なった.