表データの最適セル秘匿処理における非決定論的手法の検討

南　和宏　　データ科学研究系 教授

表データのセル秘匿処理

- 表データには、機密情報の漏洩リスクの高いセル値が含まれるため、それらのセル値を秘匿する一次秘匿処理を実施
- ただし、表データは行計、列計に関する線形一次式を内包するため、追加の2次秘匿処理が必要

度数分布表

	P_1	P_2	P_3	合計
M_1	20	24	28	72
M_2	38	38	[illegible]	79
M_3	40	39	42	121
合計	98	101	73	272

1次秘匿 →

	P_1	P_2	P_3	合計
M_1	20	24	28	72
M_2	38	38	NA	79
M_3	40	39	42	121
合計	98	101	73	272

2次秘匿 ↓

	P_1	P_2	P_3	合計
M_1	NA	24	NA	72
M_2	NA	38	NA	79
M_3	40	39	42	121
合計	98	101	73	272

最小度数ルール (e.g., $x_i > 10$) を侵害

2次秘匿処理は秘匿セル数の最小化問題

- 2次秘匿処理では、1次秘匿セルの機密性を拘束条件として、情報損失（秘匿するセルの数）の最小化を目指す
  - 秘匿パターン　$y_i \in \{0,1\} \quad i = 1, \ldots, n$

10	NA
5	80

$\leftrightarrow$ (0,1,0,0)

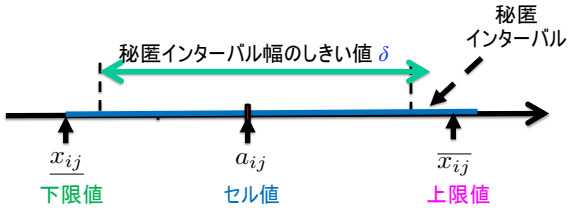
  - 目的関数: 秘匿セル数

$$\sum_{i=1}^{n} y_i$$

  - 拘束条件：各1次秘匿セル値の機密性保護

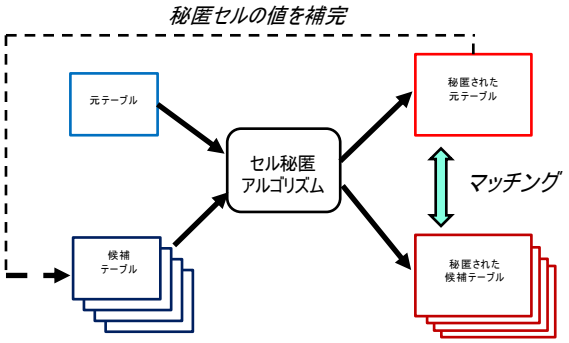
秘匿インターバルの要件

- 行計、列計の線形式を満足する可能な値の幅が指定したしきい値より大きいことが安全性の条件



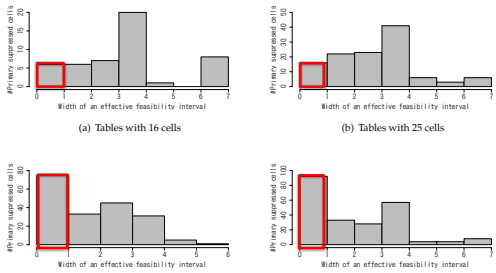
秘匿パターンのマッチングによる候補テーブルの絞り込み攻撃

- 2次秘匿表の秘匿セルに行計、列計の関係式を満足する候補値で補完し、秘匿処理の再計算を行う
- 秘匿パターンが再現できた候補テーブルの値のみが真の候補値となる



評価実験

- 4種類のサイズの表データをランダムに生成し、マッチング攻撃による実際の秘匿インターバルを評価



(a) Tables with 16 cells　(b) Tables with 25 cells

(c) Tables with 36 cells　(d) Tables with 49 cells

秘匿インターバル幅の頻度分布：赤枠で囲った部分は一意にセル値が特定されたセル数を示す

非決定論的アルゴリズムの安全性評価

- ランダム要素を加えた非決定論的セル秘匿アルゴリズムにより，同じ表から異なる秘匿表が出力され得る場合，マッチング攻撃の防止は可能
- マッチング攻撃を拡張した新規の攻撃手法が依然として存在しており、その安全性評価を検討