Vol.2019-DPS-178 No.34
Vol.2019-CSEC-84 No.34
2019/3/5

k-匿名化されたデータの安全性評価

山田　道洋[1,a)]　菊池　浩明[2]

概要：
匿名化手法の 1 つとして，k 匿名化が広く知られている．k 匿名化されたデータでは，少なくとも k 人は同一のレコードを持つために，k が大きくなればなるほど正しく k 人を再識別することが困難になる．本稿では，完全 k 匿名化されたデータから任意の m 個のレコードを選び，q 個正しく識別する確率分布を検討する．その結果を一般化し，任意の k 匿名化されたデータにおいて，不特定の人物が再識別される確率を求める．これらの結果に基づき，k 匿名化されたデータが用いられた匿名加工再識別コンテスト PWSCUP2018 の再識別率における最適な戦略を明らかにし，理想的な閾値についての検討も行う．

キーワード：匿名加工，k-匿名性

1. はじめに

2017 年 5 月 30 日に全面施行された改正個人情報保護法によって，匿名加工情報という新たな情報の類型が定義されたことにより，一定の条件の下で，本人の同意がなくても第三者提供や目的外利用が可能となった．そこで，保持するデータを匿名加工して第三者に提供するという，新しい情報流通方式が徐々に定着し始めている．しかしながら，データの特徴に応じて最適な匿名加工を行う技術は，必ずしも自明ではない．匿名加工の手法や安全性指標は，数多く存在しており，個人情報保護委員会が作成したガイドライン等に示された情報だけでは不十分である．

より良い匿名加工データを作成するには，個々のデータ加工技術者によって，有用性を維持する加工と，安全性指標の選択・検証作業が重要となる．しかし，匿名加工情報の加工技術の公開は，既に流通している情報の安全性を損ねる可能性があるため，安全に管理することが定められており，匿名加工の技術を高める機会は限られていることが課題であった．このような状況において，情報処理学会コンピュータセキュリティ研究会 (CSEC) は，匿名加工技術の発展のために，匿名加工・再識別コンテスト (PWS Cup) を開催している．これまでに，毎回異なる有用性と安全性の指標を提案し，参加者間での技術の向上と知識の共有を促してきた．PWS Cup 2018 では，加工方法としての頻繁に利用されてきた「一般化」の手法に焦点を当て，データの任意の部分集合を選んでの再識別を許すなど，いくつもの特徴的な試みをしている．中でも，あらかじめ決められた閾値を超えた再識別が行われたデータを「撃墜」と定義し，それを防止するように加工することを求めたルールは独創的であり，コンテスト参加者の動機付けを高めた．

その一方，コンテストでは，ほとんどのチームがリスクを評価した上で，$k = 2$ の 2-匿名化が最適であることを選んで競われた．また，「撃墜」をするための再識別仮名数も $m = 7$ (「撃墜」が生じる最小値) を選択するチームが上位を占めてしまうなど，コンテストが本来求めている多様性や加工されたデータの安全性についての議論や改善が必要であることが明らかになった．そこで，本稿では，PWS Cup 2018 の再識別の平均確率を詳細に分析し，そこにおける課題を明らかにする．それらを解決するために，k-匿名化されたデータの再識別に対して，クラスタ内での再識別確率の解析を行い，クラスタ間での合計再識別数の評価を試みる．

2. PWS Cup 2018 における安全性指標

2.1 「撃墜」の定義

PWS Cup 2018 においては，安全性の条件を次のように設定していた [2]．

定義 2.1 (安全性の条件 H_0) 任意の再識別アルゴリズム L，任意の A' の仮顧客の部分集合 S について，L によって S のすべての仮顧客が対応するトランザクション T の

[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
[2] 明治大学　総合数理学部　先端メディアサイエンス学科
Department of Frontier Media Science, School of Interdisciplinary Mathematical Sciences，Meiji University
[a)] cs172001@meiji.ac.jp

顧客へと正しく推定される確率が $p^{|S|}$ 以下である．

例えば，適切に 7-匿名化された A' があるとすると，同値類が特定されてしまったとしても **7 名全員を正しく当てる確率**は $1/7! = 1/5040$ であり，$p = 1/3$ とすると，$p^7 = 1/2187 > 1/5040$ なので，この条件を満たす．

コンテストでは，H_0 を帰無仮説として，有意水準 α について，

$$Pr(m \text{ 人中 } s \text{ 人以上が再識別される}) < \alpha \qquad (1)$$

であるときに, その匿名加工トランザクションが有意水準 α で棄却され，匿名化トランザクション T は安全でないという仮説（「撃墜」と呼ばれていた）が採択される．

実際には，式 (1) を厳密に求める代わりに，その十分条件である，

$$\sum_{k=s}^{m} \binom{m}{k} p^k$$

が用いられていた．極力，多様な加工が提出されることを期待して，$p = 1/3$, $\alpha = 0.01/20$ (有意水準 0.01 に 20 回検定が繰り返される（チーム数）想定での Bonferroni 補正) が用いられていた．

2.2 「撃墜」の確率解析

しかしながら，実際のコンテストでは，ほとんどのチームがリスクを評価した上で，$k = 2$ の 2-匿名化が最適であることを選んで競われた．また,「撃墜」をするための再識別仮名数も $m = 7$ (「撃墜」が生じる最小値) を選択するチームが上位を占めてしまった．

ここに次の課題があると考える．

- k-匿名化されたデータのリスクを「(そのクラスタを) 全員正しく当てる確率」で見積もっているが，この条件は厳しすぎる．$k!$ 個の順列を完全に識別しなくても，部分的に識別される仮名が多く無視できない．しかし，k-匿名化されたレコードの集合を再識別するリスクを厳密に解析するのはそれほど自明ではない（3 節）
- 安全性の条件 H_0 は任意の部分集合 S について定めているが，この帰無仮定が強すぎる．(4 節)
- $k = 2$ 以上の k-匿名性を想定した「撃墜」の閾値はいくらか？ (5 節)

例えば，k-匿名化の同値類の元が特定されることは「最悪のケース」と想定されているが，最大知識攻撃者モデルでは容易に起こりうるケースである．そもそも，擬似識別子はその属性の組で個人を特定するリスクがある属性と定義されているので，最大知識を仮定しなくても考慮しなくてはならないリスクだと言える．

図 1 に，再識別施行数 m についての再識別率の期待値の分布を示す．CUP 2018 のパラメータについて厳密に算出したところ，図の様な不連続な周期性を持ち，$m^* = 24$ の時，極大値を取っている．

再識別挑戦人数を a，a に対する安全性基準人数を b，再識別挑戦ペア数を A ペア，再識別しなければならない最低ペア数を B ペアとする．A ペア中 B ペア再識別に成功する通り数は，正解ペアと不正解ペアの重複のない組み合わせ $({}_AC_i)$ から求められる．k 匿名の 1 ペアをランダムに回答し正解する確率は $\frac{1}{k!}$，不正解の確率は $1 - \frac{1}{k!}$ なので，A ペア中 B ペア正解する確率は，$(\frac{1}{k!})^B(1 - \frac{1}{k!})^{A-B}$ となる．また，安全性基準以上の人数を再識別とみなされるため，PWSCUP2018 にて，$k = 2$ の匿名化データの再識別率は次の式で計算できる．

$$\text{再識別率} = \sum_{i=B}^{A} (\frac{1}{k!})^i (1 - \frac{1}{k!})^{A-i} {}_AC_i \qquad (2)$$

なお，A と B は，

$$A = \begin{cases} \lceil a/k \rceil & (a \bmod k = k-1 \ \& \ b \bmod k = k-1\ ) \\ \lfloor a/k \rfloor & (\text{otherwise}) \end{cases}$$

$$B = \begin{cases} \lfloor b/k \rfloor & (b \bmod k = 0) \\ \lceil b/k \rceil & (b \bmod k = k-1) \end{cases}$$

で定められる．例えば，24 人再識別（$a = 24$）に挑戦する場合，安全性基準人数は 18 人である（$b = 18$）．ここで，再識別挑戦ペア数は $12(\lfloor 24/2 \rfloor)$，再識別しなければならない最低ペア数は $9(\lfloor 18/2 \rfloor)$ となり，再識別率は 0.627 となる．また，$a = 22$, $b = 17$ の時，2 匿名においてランダムに回答する場合，ペア内で 1 人だけ不正解 $(b \bmod k = k-1)$ することはなく，2 人正解か，0 人正解かである．そのため，$b \bmod k = k-1$ の場合，必要な正解ペア数は $\lceil b/k \rceil$ となり，$A = 11$，$B = 9$ から，再識別率は 0.351 となる．

3. 同値類の元の再識別率

（問題）k 個の要素からなる仮名化データがある．この時，正しく再識別される仮名の平均数を求めよ．

仮名の集合は k-匿名化されたレコードのモデルである．例えば，$k = 3$ の時，正しい仮名を $(1,2,3)$ とした時，再識別 $(1,2,3)$ は，3 個の仮名を正しく識別し，$(1,3,2),(2,1,3),(3,2,1)$ は 1 個を正答し，$(2,3,1),(3,1,2)$ は全て外れている．

この問題は，不動点をちょうど x 個持つ k 次の置換の数 $f_k(x)$ を求める問題と言い換えることができる [3]. 置換の数は，すぐに $f_k(k) = 1$, $f_k(k-1) = 0$ であり，$x < k-1$ の時，

$$f_k(x) = \binom{k}{x} a_{k-x} \qquad (3)$$

である．ここで，a_k は k 個の置換で不動点が 0 のものの数であり，攪乱順列と呼ばれる．

$$a_k = (k-1)(a_{k-1} + a_{k-2})$$

Vol.2019-DPS-178 No.34
Vol.2019-CSEC-84 No.34
2019/3/5

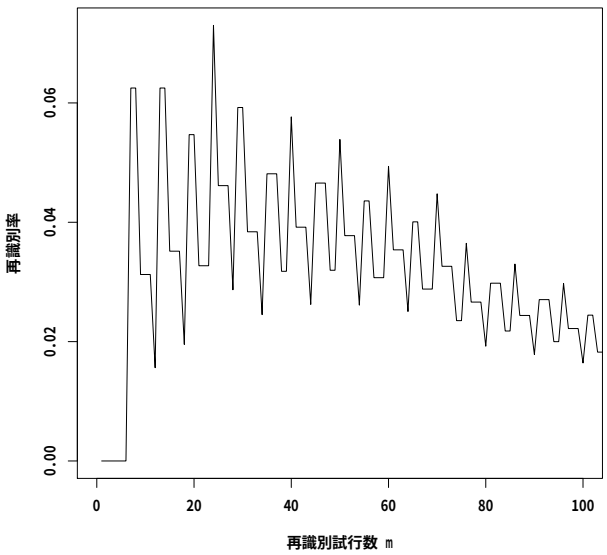

図 1 試行数 m に関する期待値の分布

表 1 攪乱順列の総数

k	a_k
1	0
2	1
3	2
4	9
5	44
6	265
7	1854
8	14833
9	133496
10	1334961
11	14684570
12	176214841

と定義され，

$$a_k = \sum_{i=2}^{k} \frac{(-1)^i k!}{i!}$$

と一般化されることが知られている．表 1 に，$k \leq 12$ までの攪乱順列の総数 a_k を示す．$k = 3$ の時，$a_3 = 2$ であり，これは先の例の，全て外れている置換 $(2,3,1),(3,1,2)$ の数を表している．$k = 2$ の時は，$(2,1)$ の一通り．

攪乱順列の総数 a_k を用いて算出した長さ k の順列の不動点数 $f_k(x)$ を表 2 に示す．興味深いことに，$x = k-1$ の時は不動点が常に 0 であり，単調性もなく，不自然な分布をしている．例えば，図 2 と図 3 に $k = 4, 6$ の場合を示す．

しかし，これらの分布の積分を取ると常に $k!$ になっていることに注意しよう．表 2 における，sum の行に該当する．これは，モンモール問題として知られている性質 [3]，

$$\sum_{x=0}^{k} x \cdot f_k(x) = k! \qquad (4)$$

表 2 長さ k の順列の不動点数 $f_k(x)$

$x \backslash k$	2	3	4	5	6	7
7						1
6					1	0
5				1	0	21
4			1	0	15	70
3		1	0	10	40	315
2	1	0	6	20	135	924
1	0	3	8	45	264	1855
0	1	2	9	44	265	1854
sum	2	6	24	120	720	5040

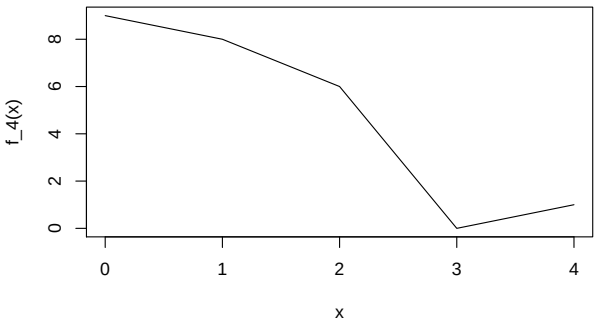

図 2 不動点数 $f_4(x)$ の分布 $(k = 4)$

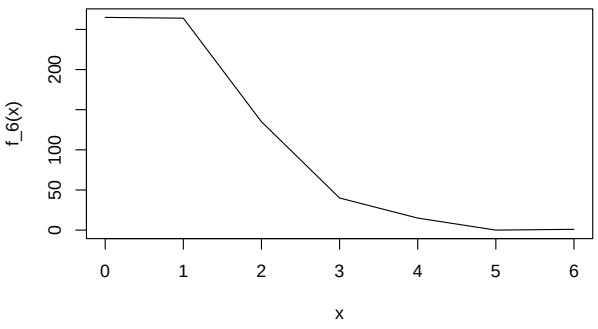

図 3 不動点数 $f_6(x)$ の分布 $(k = 6)$

である．式を

$$\sum_{x=0}^{k} x \cdot \frac{f_k(x)}{k!} = \sum_{x=0}^{k} x \cdot Pr(\text{不動点が } x \text{ 個}) = 1 \qquad (5)$$

と変形すると分かりやすい．すなわち，大きさ k の同値類の元（クラスタ）が再識別される個数 x の期待値が 1 になることを表している．クラスタには k 個の要素があるので，平均再識別率は常に $1/k$ であり，これは k-匿名化されたデータにおいて成立すると広く知られている性質に他ならない．（既知の事実かも知れないけど，私は初めて知った）

Vol.2019-DPS-178 No.34
Vol.2019-CSEC-84 No.34
2019/3/5

4. 再識別に関する二項検定

(問題) n 個の仮名から成る（完全）k-匿名化されたデータから m 仮名選んだら q 個正しく再識別された．この時，匿名化は安全と言えるか？

統計的仮説検定（二項検定）によってこの問題を検討する．まず，完全に k-匿名化されているので，データ全体は n/k 個の同値類の元に分割されている．擬似識別子の必要性から，m 個の仮名については，攻撃者にその（真の）識別子の集合が分かっているものとする．すなわち，m/k 個の同値類のそれぞれの元について，再識別を試みた和が q である．

帰無仮説 H_0: k-匿名化されたデータは安全である．（再識別は偶然に生じた）

対立仮説 H_1: k-匿名化されたデータは安全でない．（有効な再識別が行われた）

ここで，H_1 より，片側検定である．有意水準は $\alpha = 0.05$ とする．帰無仮説のもとで，検定統計量 q の確率分布を考える．m/k 個の同値類の元があり，各元について，式 (5 より，平均 k 個の要素から 1 個が再識別されるので，正しく再識別される仮名の数 q は $p = 1/k$ でサイズ m/k の二項分布 $B(m/k, 1/k)$ に従う．すなわち，

$$Pr(q|m,k) = \binom{m/k}{q/k} (\frac{1}{k})^{\frac{q}{k}} \left(\frac{k-1}{k}\right)^{\frac{m-q}{k}} \qquad (6)$$

である．二項分布の性質より，平均 $E[q] = (m/k)p = m/k^2$, 分散 $Var[q] = (m/k)p(1-p) = m(k-1)/k^3$ である．図 4 に，$m = 12$, $k = 2$ の時の確率密度関数を示す．$\mu = m/k^2 = 12/4 = 3$ が平均である．

ここで，再識別数を表す確率変数 Q が

$$Pr(Q > \theta|m,k) > \alpha$$

となる閾値は，$\theta = 9$ であり，P 値は

$$Pr(q > \theta) = 0.019 < \alpha$$

である．図にこの棄却域を示す．

こうして再識別する仮名数 m について求めた棄却域 $theta$ を図 5 に示す．また，検定を r 回繰り返す（r チームが再識別を試行する）時，少なくとも 1 つ以上が第 1 種の誤りを犯す確率は，

$$1 - (1-\alpha)^r \approx \alpha r$$

とみなして，有意水準を $\alpha' = \alpha/r$ とする Bonferroni 補正を $r = 13, 30$ について適用する（PWS Cup 2018 のチーム参加者数 14 より）．ほぼ m に対して線形に見えるが，

$$\mu + \sigma = \frac{m}{k^2} + \sqrt{\frac{k-1}{k^2} m}$$

に比例して分布している．

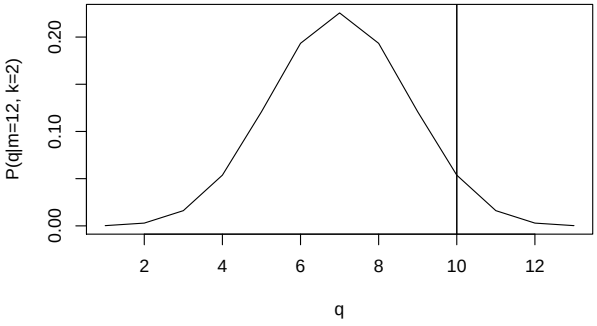

図 4 $Pr(q|k=2)$ の確率密度関数

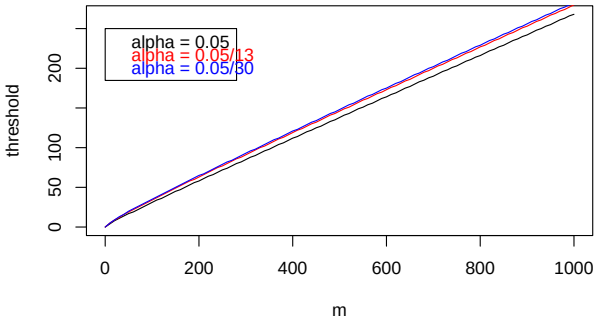

図 5 m についての棄却域 θ の分布 $(k=2)$

5. k-匿名性と安全性

式 (6) に基づいて，$k = 2, 3, 4$ のそれぞれについて求めた確率密度関数を図 6 に示す．k を大きくすることにより，再識別される仮名の平均値は，k^2 に反比例して小さくなる．k-匿名化によって，加工されたデータの再識別が著しく困難になることを表している．

k を与えた時に，平均再識別率を算出したが，逆に，q 仮名が正しく再識別された時，加工されたデータの $\hat{k}$ 推定値を算出できると興味深い．また，本稿では完全 k-匿名化を仮定したが，いくつかの異なる k の値が混在している様なデータの安全性を評価することも今後の課題である．

6. おわりに

k-匿名化されたデータに対する再識別の確率解析を行い，統計的確率検定に基づいて安全性の棄却域を算出した．棄却域 θ は，再識別試行数 m に比例して拡大する．k-匿名化された同値類の元については，k に依らずに，再識別される仮名の期待値が 1 になることを示した．

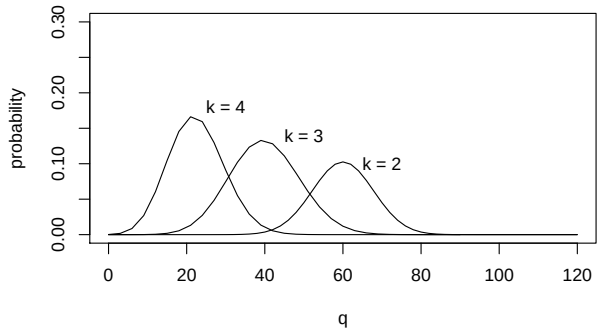

図 6 $k = 2, 3, 4$ による確率密度関数の違い

参考文献

[1] 濱田，他， “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 ～”, コンピュータセキュリティシンポジウム CSS, 2018.
[2] 濱田， “PWS Cup 2018 安全性評価の詳細”, PWS Cup 2018 コンテスト文書.
[3] 「攪乱順列の公式」, 高校数学の美しい物語, (http://https://mathtrain.jp/montmort, 2019 年 2 月参照).