6 再識別リスク
―匿名化の再識別リスクの考え方の一例―

野島　良 | 国立研究開発法人 情報通信研究機構

仮名化，一般化の再識別リスク

秘密にすべきこと

仮名化や一般化に対する再識別リスクを検討したいと思う．そこでまず次のような簡単なデータを取り上げる：

氏名	年齢	症状
井上太郎	67	腰痛
鈴木心	44	胃痛
佐藤花子	32	頭痛
田中京介	27	腹痛

本データに対する加工手法として，単純に「氏名」のみを取り除いたものを考える：

年齢	症状
67	腰痛
44	胃痛
32	頭痛
27	腹痛

このデータの再識別リスクを考えるため，攻撃者は次の情報を保有しているものとする：

（ア）井上太郎さんが本加工データに含まれていること

（イ）井上太郎さんが 67 歳であること

攻撃者は，（イ）を加工データとマッチングさせることにより，「67, 腰痛」の行（レコード）が井上太郎さんであることを知ることができる．ところが，前提条件（ア）が成立しなければ，話がまったく違ってくる．すなわち，攻撃者が，

Sample ＝｛井上太郎, 鈴木心, 佐藤花子, 田中京介｝

を知らなければ，67 歳は国内に数百万人[1]いることから，攻撃者が本加工データから得られる井上太郎さんの情報の価値は著しく低くなる．逆に

$$井上太郎 \in Sample$$

を知っているならば，本加工は無防備となる．したがって，再識別リスクを考える際は，Sample が攻撃者から見てどの程度秘密になっているかを考える必要がある．

どうしても秘密にできないこと

実は Sample を攻撃者に知られないよう努力していても，井上太郎∈ Sample であることがばれることがある．これを見ていくために，井上太郎さんが 116 歳である場合を考える．すると，彼の加工データは次のようになる：

116	腰痛

この加工データは，Sample を秘密にしていても再識別リスクが高い．なぜならば，一般に 116 歳は国内最高齢であり，公知である可能性が高いためである．この場合に再識別リスクを抑えるためには，たとえば，

90 歳以上	腰痛

とすればよい．これは 90 歳以上の人が国内に多数在住することから，安全であるとみなすことに不安がないことを根拠としている．文献 1）によると，実際に 90 歳以上は 2016 年 11 月時点で 200 万人を超えている．したがって，本加工データの公開によ

り，200 万人超の中の誰かが「腰痛持ち」であるということが知られるのみであり，井上太郎さんも公開に対して抵抗感はないはずである．

それ以外の加工に対する再識別リスク

仮名化と一般化に対する再識別リスクにおいては，基本的に攻撃手法としてマッチングのみを考えておけばよい．しかし，トランザクションデータのようなスパースなデータの場合は仮名化や一般化のみで有用性と再識別リスクを同時に保証することは難しく，さまざまな加工手法を組み込む必要がある．その場合は，攻撃手法としてマッチングのみを考えておけばよいという状況ではなくなり，複数の異なる性格の攻撃手法を考える必要性がでてくる．これを具体的に見ていくために下記データに対する加工を考える：

氏名	症状	診療日
井上太郎	腰痛	1/7
鈴木心	胃痛	1/8
佐藤花子	頭痛	1/12
井上太郎	腰痛	1/20

次の手法で加工することを考える：
①氏名の仮名化
②ノイズとして前後 3 日からランダムに選び，診療日に加える
この手法により次のような加工データが得られる：

仮名	症状	診療日
A	腰痛	1/10 (+3)
B	胃痛	1/11 (+3)
C	頭痛	1/11 (-1)
A	腰痛	1/19 (-1)

このデータの再識別リスクを考えていくため，攻撃者が次の情報を保有しているものとする：
（ウ）加工アルゴリズム
（エ）井上太郎さんが 1/7 に病院に行ったこと
この加工データに対しては少なくとも 2 通りの攻撃手法が存在する：

- **（再識別アルゴリズム 1）** ノイズが前後 3 日であることから，井上太郎さんのデータ

井上太郎	腰痛	1/7

を加工したデータは，

?	腰痛	D

である．ただし，D は，1/4 ～ 1/10 のいずれかである．したがって，攻撃者は，井上太郎さんが仮名 A であることを知ることができる
- **（再識別アルゴリズム 2）** 1/7 に最も近いデータを井上太郎さんのデータとする．この場合は，

A	腰痛	1/10

が最も 1/7 に近いため，A が井上太郎さんであると推測する

このように仮名化と一般化以外の加工手法に対しては，性格の異なる複数の再識別アルゴリズムが存在し得る．リスク評価を行うためには，1 つ 1 つの加工手法（とデータ）に対して，最も強力な再識別アルゴリズムを考えなければならず，基準となる再識別リスク評価手法の確立に対する障害となっている．

参考文献
1）総務省統計局公表データより
http://www.stat.go.jp/data/jinsui/2.htm#monthly

（2018 年 1 月 31 日受付）

■**野島　良（正会員）** ryo-no@nict.go.jp
国立研究開発法人 情報通信研究機構 研究マネージャー．研究テーマ：セキュリティ，暗号理論．