動的データセットにおけるプライバシ保護の厳密な安全性評価と妥当な安全性評価について

坂田奈々子† 上土井陽子†† 村上 頼太† 若林 真一††

† 広島市立大学 情報科学部 〒 731–3194 広島市安佐南区大塚東三丁目 4-1
†† 広島市立大学 大学院情報科学研究科 〒 731–3194 広島市安佐南区大塚東三丁目 4-1

あらまし 挿入や削除が行われる動的データセットの再公開においてプライバシを保護する方法として先行研究では m-不変性が提案され，その安全性を評価する方法として全射関数を用いて確率的に評価する方法が提案されている．しかし，この安全性評価方法では，攻撃者に再公開表を作成したときの情報を与えることで，安全性評価値が上昇することがあることがわかった．本研究では m-不変性を満たす時の動的データセットの再公開における安全性を厳密な確率によって評価する方法について考察し，その妥当性について検討する．

キーワード m-不変性, 確率的リスク, 動的データセット, 再公開

1. はじめに

挿入や削除が行われる動的データセットの再公開においてプライバシを保護する方法として先行研究 [1] では m-不変性が提案され，その安全性を評価する方法として全射関数を用いて確率的に評価する方法が提案されている．しかし，この安全性評価方法では，攻撃者に再公開表を作成したときの情報を与えている．

本研究では m-不変性を満たす時の動的データセットの再公開における安全性を厳密な確率によって評価する方法について考察する．厳密な安全性評価と従来評価手法を比較したとき，m-不変性を満たす再公開であっても m-不変性を満たすように作成されたという情報を攻撃者に利用させなければ目標となる評価値にならない場合が存在することが分った．

2. 準　　備

本研究において必要な基礎定義を述べるにあたり，元データの列を次の (1)∼(3) に分類する．(以下，T は公開者により維持されている元データを示すとする．)

(1) T は識別属性 A^{id} をもつ．例えば，A^{id} は名前の集合を表す．

(2) T は d 個の準識別子（QI）属性 A_i^{qi} をもつ．例えば，A_1^{qi} は年齢の集合，A_2^{qi} はジップコードの集合である．

(3) T は機密データ A^S をもつ．例えば，A^S は病名の集合である．

また，各タプル t に対し，$t[A]$ は属性 A における t の値を示し，j 番目の表 T を $T(j)$ とし時刻 j における T のスナップショットとする．$T^*(j)$ は $T(j)$ を一般化した表を示し，$T(j)$ のある統計値を提供する補助表を $R(j)$ とする．公開者は $T^*(j)$, $R(j)$ の組を公開し，その際にデータは後で述べる偽造を含む一般化条件を満たす状態で公開される場合があるとする．以上を踏まえ，基礎定義を述べる．

[定義 2.1]（QI グループと分割） QI グループとはタプルの部分集合であり，$T(j)$ に対しての分割は QI グループの和集合が $T(j)$ と等しくなる互いに素な QI グループで構成されている．各 QI グループは分割内で，一意の ID を割り当てられる．また，$T(j)$ のタプル t に対し，$t.QI(j)$ は t を含む QI グループの *hosting* グループと呼ばれる．

[定義 2.2]（偽造を含む一般化） 公開表 $T^*(j)$ は $T(j)$ に基づいて作成され，以下の特性を持つ．

（1） $T^*(j)$ は ”グループ ID ”と呼ばれる A^g と，A^{id} を除いた $T(j)$ の全ての属性値を含む．

（2） 同じグループ ID A^g を持つ $T^*(j)$ の全てのタプルは全て同じ QI 属性値を持つ．これらのタプルは $T^*(j)$ で 1 つの QI グループを形成し，グループ内の A^g の値を ID として持つ．一般化されたタプルは元のタプルと等しい機密属性値を持ち，一般化された QI 属性値の範囲は元のタプルの属性値を含む範囲である．

（3） $T(j)$ の各 QI グループに対して，$T^*(j)$ は偽造されたタプル t_c^* を任意の個数含み，その機密属性値 $t_c^*[A^S]$ は A^S の領域の任意の値であり，グループ ID $t_c^*[A^g]$ は QI の ID と同じ値になる．

$t \in T(j)$ に対して，$t.QI^*(j)$ は t の一般化されたタプルを含む $T^*(j)$ での QI グループを示し，$T^*(j)$ において t の一般化された *hosting* グループとして参照される．

[定義 2.3]（補助表） 補助表 $R(j)$ は”グループ ID”と”Count”と呼ばれる 2 つの列を持つ．少なくとも 1 つの偽造を含む公開表 $T^*(j)$ での $R(j)$ において行 $\langle g, c\rangle$ が存在する．ここで g は偽造タプルが挿入されている QI グループ QI^* の ID を，c は QI^* にある偽造タプルの数を示す．$T^*(j)$ において偽造タプルが存在しない場合，$R(j)$ は空である．$R(j)$ により偽造の情報を公開することは，公開表を用いた統計的な解析において偽造タプル挿入による情報歪曲の影響を勘案するための助けとなる．

[定義 2.4]（一般化基準） 一般化基準は $T^*(1), \cdots, T^*(n)$ での QI グループによって満たされるべき制限の集合であり，k-匿名性と l-多様性は一般化基準のうちの 2 つである．k-匿名性は，

表 1　1 回目の元データ

Name	Age	Zip.	Disease
Bob	21	12k	消化不良
Alice	22	14k	気管支炎
Andy	24	18k	インフル
David	23	25k	胃炎
Gray	41	20k	インフル
Helen	36	27k	胃炎
Jane	37	33k	消化不良
Ken	40	35k	インフル
Linda	43	26k	胃炎
Paul	52	33k	消化不良
Steve	56	34k	胃炎

表 2　表 1 を一般化した表

G. ID	Age	Zip.	Disease
1	[21,22]	[12k,14k]	消化不良
1	[21,22]	[12k,14k]	気管支炎
2	[23,24]	[18k,25k]	インフル
2	[23,24]	[18k,25k]	胃炎
3	[36,41]	[20k,27k]	インフル
3	[36,41]	[20k,27k]	胃炎
4	[37,43]	[26k,35k]	消化不良
4	[37,43]	[26k,35k]	インフル
4	[37,43]	[26k,35k]	胃炎
5	[52,56]	[33k,34k]	消化不良
5	[52,56]	[33k,34k]	胃炎

表 3　2 回目の元データ

Name	Age	Zip.	Disease
Bob	21	12k	消化不良
David	23	25k	胃炎
Emily	25	21k	インフル
Jane	37	33k	消化不良
Linda	43	26k	胃炎
Gray	41	20k	インフル
Mary	46	30k	胃炎
Ray	54	31k	消化不良
Steve	56	34k	胃炎
Tom	60	44k	胃炎
Vince	65	36k	インフル

表 4　表 3 を一般化した表

G. ID	Age	Zip.	Disease
1	[21,23]	[12k,25k]	消化不良
1	[21,23]	[12k,25k]	胃炎
2	[25,43]	[21k,33k]	インフル
2	[25,43]	[21k,33k]	消化不良
2	[25,43]	[20k,30k]	胃炎
3	[41,46]	[20k,30k]	インフル
3	[41,46]	[20k,30k]	胃炎
4	[54,56]	[31k,34k]	消化不良
4	[54,56]	[31k,34k]	消化不良
5	[60,65]	[36k,44k]	消化不良
5	[60,65]	[36k,44k]	インフル

表 5　表 3 を一般化した表（偽造を含む）

Name	Age	Zip.	Disease
Bob	[21,22]	[12k,14k]	消化不良
C1	[21,22]	[12k,14k]	気管支炎
David	[23,25]	[21k,25k]	胃炎
Emily	[23,25]	[21k,25k]	インフル
Jane	[37,43]	[26k,33k]	消化不良
C2	[37,43]	[26k,33k]	インフル
Linda	[37,43]	[26k,33k]	胃炎
Gray	[41,46]	[20k,30k]	インフル
Mary	[41,46]	[20k,30k]	胃炎
Ray	[54,56]	[31k,34k]	消化不良
Steve	[54,56]	[31k,34k]	胃炎
Tom	[60,65]	[36k,44k]	胃炎
Vince	[60,65]	[36k,44k]	インフル

表 6　補　助　表

G. ID	Count
1	1
3	1

表 7　背景知識表の一部

G. ID	Name	Age	Zip.	lifespan
*	Bob	21	12k	[1,2]
*	Alice	22	14k	[1,2]
…	…	…	…	…
*	Helen	36	27k	[1,1]
*	David	23	25k	[1,2]
…	…	…	…	…
1	C1	0	0	[2,2]
3	C2	0	0	[2,2]

$T^*(j)$ の各 QI グループが少なくとも k タプルを持たなければならないという制約を課し，l-多様性は各 QI グループが，同じ機密な値を高々(QI グループの大きさ)$\times\frac{1}{l}$ 個持つという制約を課している．

[定義 2.5]（履歴の統合） $n \geq 1$ の時，時系列で変化する元データの和集合を $U(n)$ で表す．$U(n)$ はタイムスタンプ 1,2,...,n の順に T にある全てのタプルを含んでおり

$$U(n) = \bigcup_{j=1}^{n} T(j)$$

で表現される．また，$t \in U(n)$ の各タプルは $t.Lifespan = [x, y]$ という項をもつことによりタイムスタンプ以外の情報の重複をまとめて，集約した形で表現される．ここで，x は t が表 $T(j)$ に追加された時のタイムスタンプの値 j，y は t が $T(j)$ から削除される 1 つ前のタイムスタンプの値 j を示す．

以降，攻撃者がもつ背景知識を定義する．

[定義 2.6]（背景知識） 時刻 n において，相手は予め

（1） 利用される一般化基準

（2） "グループ ID"と呼ばれる A^g，そして元データセットの和集合 $U(n)$ から機密属性 A^S を除いた $U(n)$ の全ての属性を持つ背景知識表 $B(n)$（表 7）

の 2 つを持つ．このとき，$B(n)$ には (i)A^S を除いた $U(n)$ の全て，(ii)$U(n)$ の各タプルの $Lifespan$，(iii) 全ての偽造タプルの公開された詳細情報が組み込まれている．

[定義 2.7]（プライバシ漏洩） もし，攻撃者が $T^*(1), \cdots, T^*(n)$ と $B(n)$ を利用して，任意のタプル $t \in U(n)$ の機密な値を正しく見つけ出すことができた場合，プライバシ漏洩が発生する．

[定義 2.8]（再公開） 公開者がマイクロデータ T の $n-1$ 個の異なるバージョンの匿名化された表 $\{T^*(1), R(1)\}, \cdots, \{T^*(n-1), R(n-1)\}$ (n は 1 以上の整数) を公開するとする．ここで，$\{T^*(j), R(j)\}(1 \leqq j \leqq n-1)$ は定義 2.2 と定義 2.3 で示された表の組とする．プライバシを保護した再公開の目的は，プライバシ漏洩のリスクを最小にしつつ，できる限り元データに近い情報を公開するような $\{T^*(n), R(n)\}$ の組を計算することである．

[定義 2.9]（$signatue$） 任意の $j \in [1, n]$ において $T^*(j)$ の QI グループを QI^* とする．QI^* の $signature$ は QI^* グループに含まれる機密属性値の集合である．

[定義 2.10]（m-不変性） $T^*(j)$ の各 QI グループが少なくとも m 個のタプルを含み，各グループ内の全てのタプルが異なる機密属性値を持つならば，一般化された表 $T^*(j)(i \leqq j \leqq n)$ は m-ユニークであるという．この時，以下の条件が成り立つならば，一連の公開された表 $T^*(1), \cdots, T^*(n)(n \geqq 1)$ は m-不変性を満たすという．

1） 全ての $j \in [1, n]$ において $T^*(j)$ は m-ユニークである．

2） 任意のタプル $t \in U(n)$ において，t.Lifespan が $[x, y]$ であるとき $t.QI^*(x), t.QI^*(x+1), \cdots, t.QI^*(y)$ は同じ $signature$ を持つ．この時 $t.QI^*(j)$ は時間 $j \in [x, y]$ で t の一般化されたホスティンググループである (定義 2.2 参照)．

m-ユニークは，全ての QI グループで各機密属性値は高々一度しか現れないことを示している．一般的に m-ユニークは l-多

様性 ($l = m$) を暗示している．例としては，表 2 と表 4 が m-不変性を満たした公開表列である．

3. 厳密な確率によるリスク評価

参考文献 [1] では安全性を評価する基準としてプライバシ漏洩のリスクが評価されている．本節では，参考文献 [1] のリスク評価の為の定義を基礎として，文献 [1] で対象としている攻撃者よりも自由な攻撃者に対するリスクを厳密に評価する方法に従来リスク評価定義を変形する．本稿でも文献 [1] と同様に，公開表を統合した表 $U^*(n)(n \geqq 1)$ と背景知識表 $B(n)(n \geqq 1)$ を 2 つの集合とみなし，$U^*(n)$ から $B(n)$ への全射関数を考えマイクロデータの再構築を行う．また，その全射関数の中でも元のマイクロデータを矛盾なく再構築する全射を理にかなった全射とする．ここで，文献 [1] での理にかなった全射の定義を区別する為，本研究では制約なしの理にかなった全射と呼ぶ．制約なしの理にかなった全射の数を数えることにより，プライバシ漏洩の厳密な確率的なリスクを計算する．

次に，全射関数と理にかなった全射関数の定義を以下に示す．

［定義 3.1］（全射関数） 以下の条件を満たす場合，
$U^*(n) \rightarrow B(n)$ のマッピング f は全射関数である．

1. 各タプル $t^* \in U^*(n)$ を，任意のタプル $b \in B(n)$ にマッピングする関数は $f(t^*) = b$ で表現される．
2. 任意のタプル $b \in B(n)$ において，$f(t^*) = b$ となるようなタプル $t^* \in U^*(n)$ が少なくとも 1 つ存在する．
3. $f(t^*) = b$ の時，
   - 3.1 $b[A^l]$ は $t^*[A^{tm}]$ を含む
   - 3.2 $t^*[A^g] = b[A^g]$ である ($b[A^g] =^*$ の時，この等価性は常に成り立つ)
   - 3.3 $t^*[A_i^{qi}]$ は各 QI 特性 A_i^{qi} に従い $b[A_i^{qi}]$ を含むような範囲である ($b[A_i^{qi}] = 0$ の時，包含関係は常に成り立つ)

［定義 3.2］（制約のない理にかなった全射） 以下の条件を満たす場合，定義 3.1 における関数 f は理にかなっている．

1. 任意のタプル $b \in B(n)$ において，$f(t^*) = b$ を満たすようなタプル $t^* \in U^*(n)$ の集合 $f^{-1}(b)$ が与えられた時，
   - 1 $f^{-1}(b)$ の全てのタプルは同じ機密な値をもつ
   - 2 b のライフスパン $b[A^l]$ 内の任意のタイムスタンプ j において，$t^*[A^{tm}] = j$ となるような唯一のタプル $t^* \in f^{-1}(b)$ が存在する

文献 [1] の理にかなった全射関数の定義では，上記の 1 の条件に加え，(2. f^{-1} は利用される一般化基準と一致するような一般化の候補を決める) という条件が存在している．本稿では，条件 2 を削除した定義を厳密で理にかなった全射関数を定義するものとして，制約のない理にかなった全射関数と呼ぶこととする．

定義 3.2 の制約のない理にかなった全射関数の数を用いてプライバシ漏洩のリスクを計算する．

［定義 3.3］（プライバシ漏洩の厳密な確率的なリスク） プライバシ漏洩のリスクとは，攻撃者がある個人 t に関するプライバシを破ることができる可能性を示しており，以下の式で定義される．t を履歴の統合表 $U(n)$ のタプルとすると，t のプライバシ漏洩のリスク $risk_{strict}(t)$ は

$$risk_{strict}(t) = \frac{n_{breach.strict}(t)}{n_{total.strict}}$$

で示される．ここで $n_{total.strict}$ は制約のない理にかなった全射関数の数で，$n_{breach.strict}$ は制約のない理にかなった全射関数のうち t がある機密な値に結び付けられることができる全射関数の数である．

ある制約のない理にかなった全射関数は，あるタプルに関しては正確に再構築を行うかもしれないが，その他のタプルに関しては正確に再構築を行わない場合がある．よって，$U^*(n)$ の各タプルはそれぞれ異なるプライバシ漏洩のリスクを持っているといえる．

特に $n_{breach.strict}(t) = n_{total.strict}$ である時 (すなわち $risk_{strict}(t) = 1$ の時)，背景知識 $B(n)$ を持つ攻撃者は，100%の確率でタプル t の機密な値を見つけることができる．

4. m-不変性保持による厳密な確率的リスクの制御効果

文献 [1] では，m-不変性を保持させるように公開表の列を作成することで，プライバシ漏洩のリスクを $1/m$ 以下に抑えることができると主張されている．本節では，この主張を厳密な確率的リスクの観点から検討する．

4.1 2-不変性を満たす公開表列の厳密な確率的リスク評価例

表 8 と表 9 を第 1 回目，第 2 回目の公開表とし，表 8，9 を統合した表を表 10 とする．また，攻撃者が知り得る最大の知識表を背景知識として表 11 に示す．表 8，9 の公開表の列は m-不変性の定義を満たしており，2-不変性を満たす例である．

表 8　1 回目の公開表 $T^*(1)$

Name	G. ID	Age	Zip.	Disease
Bob	1	[21,22]	[12k,14k]	消化不良
Alice	1	[21,22]	[12k,14k]	気管支炎
Andy	2	[23,24]	[18k,25k]	消化不良
David	2	[23,24]	[18k,25k]	胃炎

表 9　2 回目の公開表 $T^*(2)$

Name	G. ID	Age	Zip.	Disease
Bob	1	[21,22]	[12k,14k]	消化不良
C1	1	[21,22]	[12k,14k]	気管支炎
David	2	[21,23]	[12k,25k]	胃炎
Emily	2	[21,23]	[12k,25k]	消化不良

表 10 から表 11 への理にかなった全射関数の全てを表 12 に示す．表 12 の全射関数の数から定義 3.3 の risk を計算すると，表 13 のように各個人がそれぞれの病気になる確率を出すことができる．一方，参考文献 [1] では m-不変性を満たす時のリスクについての補題と，その証明が示されている．補題と証明を以下に示す．

表 10　表 8 と表 9 の統合表 $U^*(2)$

No	G. ID	Age	Zip.	Disease	timestamp
1.1	1	[21,22]	[12k,14k]	消化不良	1
1.2	1	[21,22]	[12k,14k]	気管支炎	1
1.3	2	[23,24]	[18k,25k]	消化不良	1
1.4	2	[23,24]	[18k,25k]	胃炎	1
2.1	1	[21,22]	[12k,14k]	消化不良	2
2.2	1	[21,22]	[12k,14k]	気管支炎	2
2.3	2	[21,23]	[12k,25k]	胃炎	2
2.4	2	[23,23]	[12k,25k]	消化不良	2

表 11　背景知識表 $B(2)$

G. ID	Name	Age	Zip.	lifespan
*	Bob	21	12000	[1,2]
*	Alice	22	14000	[1,1]
*	Andy	24	18000	[1,1]
*	David	23	25000	[1,2]
1	C1	ϕ	ϕ	[2,2]
*	Emily	21	12000	[2,2]

［補題 4.1］（文献 1）　もし，$T*(1), \cdots, T*(n)$ が m-不変性を満たすならば，その時の任意の t $\in U(n)$ において，

$$risk(t) \leqq \frac{1}{m}$$

を満たす．

ここで，表 13 で示した risk の計算結果を見てみると，Bob が消化不良の時，Alice が気管支炎の時，Andy が消化不良の時，David が消化不良の時，C1 が気管支炎の時，Emily が消化不良の時は risk が 2/3 になっており，上記の補題 4.1 の式を満たしていない．従って表 8 と 9 は 2-不変性を満たす例であったが，厳密な確率的リスク評価では，リスクを 1/2 以下に抑制することができていないことが分かった．

表 12　全ての全射関数を示す表

1.1(消)	1.2(気)	1.3(消)	1.4(胃)	2.1(消)	2.2(気)	2.3(胃)	2.4(消)
Bob	Alice	Andy	David	Bob	C1	David	Emily
				C1	Emily	David	Bob
		Andy	David	Emily	C1	David	Bob
		David	Andy	Bob	C1	Emily	David
Alice	Bob	Andy	David	C1	Bob	David	Emily
		David	Andy	C1	Bob	Emily	David

表 13　表 12 を用いて risk を計算した表

	消化不良	気管支炎	胃炎
Bob	$\frac{4}{6}=\frac{2}{3}$	$\frac{2}{6}=\frac{1}{3}$	0
Alice	$\frac{2}{6}=\frac{1}{3}$	$\frac{4}{6}=\frac{2}{3}$	0
Andy	$\frac{4}{6}=\frac{2}{3}$	0	$\frac{2}{6}=\frac{1}{3}$
David	$\frac{4}{6}=\frac{2}{3}$	$\frac{2}{6}=\frac{1}{3}$	0
C1	$\frac{2}{6}=\frac{1}{3}$	$\frac{4}{6}=\frac{2}{3}$	0
Emily	$\frac{4}{6}=\frac{2}{3}$	0	$\frac{2}{6}=\frac{1}{3}$

4.2　情報提供による確率的なリスクの制御

前小節の結果は，文献 [1] の補題 4.1 と矛盾しているように見えるが，このような関係は，理にかなった全射の定義を厳密な確率的リスクを算出する為に変更したことによって引き起こされていると考えられる．そこで補題 4.1 の証明を考察することで，文献 [1] ではどのように確率的なリスクを制御しようとしているのか検討する．

4.3　補題 4.1 の証明　(文献 1)

t は $U(n)$ の任意のタプルとし，b は t によって生成された $B(n)$ の行とする．任意の理にかなった全射関数 $f : U*(n) \longrightarrow B(n)$ を与え，$AQ(b, f)$ を $f^{-1}(b)$ に属する少なくとも 1 つのタプルを含む $T*(1), \cdots, T^*(n)$ の QI グループの集合と定義する．

n_{total} 個ある $U^*(n)$ から $B(n)$ への理にかなった全射を考える．任意の理にかなった全射 f と f' が同じグループになるならば，$AQ(b, f) = AQ(b, f')$ になるようにそれらを分割する．n_{group} を結果として生じるグループの総数とする．i 番目 $(1 \leqq i \leqq n_{group})$ のグループ F_i において，v として t の機密な値を再構築する F_i に属する全射の数を表す為に $cnt(F_i, v)$ を使用する．最終的に $cnt(F_i, t[A^s]) \leqq \dfrac{|F_i|}{m}$ を示すことができれば，

$$risk(t) = \frac{\Sigma_{i=1}^{n^{group}} cnt(F_i,\, t[A^s])}{n_{total}} \leqq \frac{\Sigma_{i=1}^{n^{group}} |F_i|}{m \cdot n_{total}} = \frac{1}{m}.$$

が成り立ち，補題を証明することができる．

任意の理にかなった全射関数 $f \in F_i$ が与えられると，理にかなった全射は m-不変性を満たすように見えなければならない為，$AQ(b, f)$ での全ての QI グループは同じ signature を持つ．一般性を失わずに，$AQ(b, f)$ の各 QI グループが x 個の機密な値 $v_1, v_2, \cdots, v_x$ を持つと仮定する．m-不変性を満たす為には x 個の値は少なくとも m 種類ある．

v_1 として t の機密な値を再構築する F_i に属する任意の全射を f_1 とする．また，もう一つの全射 $f_2 : U^*(n) \longrightarrow B(n)$ を設計する．$f_2(t^*)$ は任意のタプル $t^* \in U^*(n)$ において未定義である．次に各 $T^*(j)(1 \leqq j \leqq n)$ で QI グループを調べる．もし，QI グループが 2 つのタプル t_1^* と t_2^* を $t_1^*[A^s] = v_1$ と $t_2^*[A^s] = v_2$ で満たすならば，$f_2(t_1^*) = f_1(t_2^*)$，$f_2(t_2^*) = f_1(t_1^*)$ でセットする．全ての QI グループを調べた後に，$f_2(t^*)$ が未定義のままな任意のタプル $t^* \in U^*(n)$ において，$f_2(t^*) = f_1(t^*)$ を満たす．従って，f_2 は F_i に属している．

f_2 は t の機密な値を v_2 として再構築する．任意の f_1 においての f_2 が存在するので，$cnt(F_i, v_1) \leqq cnt(F_i, v_2)$ を意味する．対照的に，$cnt(F_i, v_2) \leqq cnt(F_i, v_1)$ から $cnt(F_i, v_2) = cnt(F_i, v_1)$ を得ることができる．分析を $v_1, \cdots, v_x$ の全ての要素まで広げると $cnt(F_i, v_1) = cnt(F_i, v_2) = \cdots cnt(F_i, v_x) = \dfrac{|F_i|}{m}$ といえる．　(証明終)

補題 4.1 では，理にかなった全射を m-不変性を保持できるものだけに限定することにより確率的リスクを抑制しようとしていることが分かった．また，m-不変性を保持できる全射のみを考える為には理にかなった全射の定義にある「f^{-1} は利用される一般化基準と一致するような一般化の候補を決める」という条件が必要である．

これらのことを 4.1 と同じ表 8 から表 11 を用いて再考すると，厳密な確率的リスクを制御しようとしている為，表 10 から表 11 への全射関数を考える際にも m-不変性を満たしていなければならないとして全射関数を考えると，表 10 から表 11 への理にかなった全射関数の全てを表 14 に示す．この時，表 12

と表 14 を比べると表 14 では表 12 における，上から 2 番目の行と 3 番目の行が削除されている．なぜなら，m-不変性を満たすには signature が同じでなければならないが，表 12 の 2 行目と 3 行目の Bob に注目すると，1.1 から 1.4 の signature は {消化不良, 気管支炎} であるのに，2.1 から 2.4 の signature は {胃炎, 消化不良} となっているからである．これは m-不変性を満たせていない．従って，表 14 では表 12 から 2 つの行が削除されている．表 14 の全射関数の数から定義 3.3 の risk を計算すると，表 15 のように各個人がそれぞれの病気になる確率を出すことができる．

表 15 をみると，全てのリスクが 1/2 以下になっているので，補題 4.1 を満たしていることがわかる．

表 14　全ての全射関数を示す表

1.1(消)	1.2(気)	1.3(消)	1.4(胃)	2.1(消)	2.2(気)	2.3(胃)	2.4(消)
Bob	Alice	Andy	David	Bob	C1	David	Emily
		David	Andy	Bob	C1	Emily	David
Alice	Bob	Andy	David	C1	Bob	David	Emily
		David	Andy	C1	Bob	Emily	David

表 15　表 14 を用いて risk を計算した表

	消化不良	気管支炎	胃炎
Bob	$\frac{2}{4}=\frac{1}{2}$	$\frac{2}{4}=\frac{1}{2}$	0
Alice	$\frac{2}{4}=\frac{1}{2}$	$\frac{2}{4}=\frac{1}{2}$	0
Andy	$\frac{2}{4}=\frac{1}{2}$	0	$\frac{2}{4}=\frac{1}{2}$
David	$\frac{2}{4}=\frac{1}{2}$	0	$\frac{2}{4}=\frac{1}{2}$
C1	$\frac{2}{4}=\frac{1}{2}$	$\frac{2}{4}=\frac{1}{2}$	0
Emily	$\frac{2}{4}=\frac{1}{2}$	0	$\frac{2}{4}=\frac{1}{2}$

5. 確率的リスク評価の妥当性の検討

前節の結果では，動的データの再公開の確率的なリスク評価では攻撃者に公開表列の作成に利用した一般化基準等の情報を背景知識に加えることで，背景知識だけを攻撃者が持つ場合のプライバシ漏洩の確率的リスクと比べて低いリスクを得られる場合が存在することを示した．本節では，動的データの再公開の確率的なリスク評価の妥当性についてさらに検討する．

5.1 一般化による確率的なリスク評価値の変化

表 16 を 1 回目の公開表 $T^*(1)$，表 17 を 2 回目の公開表 $T^*(2)$，表 18 を統合表 $U^*(2)$，表 19 を背景知識表 $B(2)$ とする．これらの表を用いて表 18 から表 19 への理にかなった全射関数の数を算出し，その結果を表 20 に示す．ここで，定義 3.3 の方法で risk を計算し，厳密なリスク評価を行うと，表 21 のような結果が得られた．

上記の例は，全ての risk が 1/2 以下に抑えられている．ここで，表 17 をさらに一般化することで，risk がどのように変化するのか確かめる．

表 17 をさらに一般化した表を表 22 に示し，それに伴って表 16 と表 22 の統合表を表 23 に示す．表 22 で行ったさらなる一般化は，タプル 2.1, 2.2 の郵便番号の範囲を [12k,14k] から [12k,21k] に，タプル 2.3, 2.4 の郵便番号の範囲を [21k,25k] から [12k,25k] に広げるというものである．これらの表を用いて表 23 から表 19 への理にかなった全射関数の数を算出し，その

表 16　1 回目の公開表 $T^*(1)$

Name	G. ID	Age	Zip.	Disease
Bob	1	[23,25]	[12k,14k]	消化不良
Alice	1	[23,25]	[12k,14k]	気管支炎
David	2	[22,24]	[18k,25k]	インフル
Andy	2	[22,24]	[18k,25k]	気管支炎

表 17　2 回目の公開表 $T^*(2)$

Name	G. ID	Age	Zip.	Disease
Bob	1	[21,25]	[12k,14k]	消化不良
C1	1	[21,25]	[12k,14k]	気管支炎
Emily	2	[22,25]	[21k,25k]	気管支炎
David	2	[22,25]	[21k,25k]	インフル

表 18　表 16 と表 17 の統合表 $U^*(2)$

No	G. ID	Age	Zip.	Disease	timestamp
1.1	1	[23,25]	[12k,14k]	消化不良	1
1.2	1	[23,25]	[12k,14k]	気管支炎	1
1.3	2	[22,24]	[18k,25k]	インフル	1
1.4	2	[22,24]	[18k,25k]	気管支炎	1
2.1	1	[21,25]	[12k,14k]	消化不良	2
2.2	1	[21,25]	[12k,14k]	気管支炎	2
2.3	2	[22,25]	[21k,25k]	気管支炎	2
2.4	2	[22,25]	[21k,25k]	インフル	2

表 19　背景知識表 $B(2)$

G. ID	Name	Age	Zip.	lifespan
*	Bob	25	12000	[1,2]
*	Alice	23	14000	[1,1]
*	David	22	25000	[1,2]
*	Andy	24	18000	[1,1]
1	C1	ϕ	ϕ	[2,2]
*	Emily	25	21000	[2,2]

表 20　全ての全射関数を示す表

1.1(消)	1.2(気)	1.3(イ)	1.4(気)	2.1(消)	2.2(気)	2.3(気)	2.4(イ)
Bob	Alice	David	Andy	Bob	C1	Emily	David
		Andy	David	Bob	C1	David	Emily
Alice	Bob	David	Andy	C1	Bob	Emily	David
		Andy	David	C1	Bob	David	Emily

表 21　表 20 を用いて risk を計算した表

	消化不良	気管支炎	インフル
Bob	$\frac{2}{4}=\frac{1}{2}$	$\frac{2}{4}=\frac{1}{2}$	0
Alice	$\frac{2}{4}=\frac{1}{2}$	$\frac{2}{4}=\frac{1}{2}$	0
David	0	$\frac{2}{4}=\frac{1}{2}$	$\frac{2}{4}=\frac{1}{2}$
Andy	0	$\frac{2}{4}=\frac{1}{2}$	$\frac{2}{4}=\frac{1}{2}$
C1	$\frac{2}{4}=\frac{1}{2}$	$\frac{2}{4}=\frac{1}{2}$	0
Emily	0	$\frac{2}{4}=\frac{1}{2}$	$\frac{2}{4}=\frac{1}{2}$

表 22　表 17 をさらに一般化した表

Name	G. ID	Age	Zip.	Disease
Bob	1	[21,25]	[12k,21k]	消化不良
C1	1	[21,25]	[12k,21k]	気管支炎
Emily	2	[22,25]	[12k,25k]	気管支炎
David	2	[22,25]	[12k,25k]	インフル

表 23　表 16 と表 22 の統合表

No	G. ID	Age	Zip.	Disease	timestamp
1.1	1	[23,25]	[12k,14k]	消化不良	1
1.2	1	[23,25]	[12k,14k]	気管支炎	1
1.3	2	[22,24]	[18k,25k]	インフル	1
1.4	2	[22,24]	[18k,25k]	気管支炎	1
2.1	1	[21,25]	[12k,21k]	消化不良	2
2.2	1	[21,25]	[12k,21k]	気管支炎	2
2.3	2	[22,25]	[12k,25k]	気管支炎	2
2.4	2	[22,25]	[12k,25k]	インフル	2

表 24　全ての全射関数を示す表

1.1(消)	1.2(気)	1.3(イ)	1.4(気)	2.1(消)	2.2(気)	2.2(気)	2.4(イ)
Bob	Alice	David	Andy	Bob	C1	Emily	David
		Andy	David	Bob	C1	David	Emily
Alice	Bob	David	Andy	C1	Bob	Emily	David
		David	Andy	C1	Emily	Bob	David
		David	Andy	Emily	C1	Bob	David
		Andy	David	C1	Bob	David	Emily

表 25　表 24 を用いて risk を計算した表

	消化不良	気管支炎	インフル
Bob	$\frac{2}{6} = \frac{1}{3}$	$\frac{4}{6} = \frac{2}{3}$	0
Alice	$\frac{4}{6} = \frac{2}{3}$	$\frac{2}{6} = \frac{1}{3}$	0
David	0	$\frac{2}{6} = \frac{1}{3}$	$\frac{4}{6} = \frac{2}{3}$
Andy	0	$\frac{4}{6} = \frac{2}{3}$	$\frac{2}{6} = \frac{1}{3}$
C1	$\frac{3}{6} = \frac{1}{2}$	$\frac{3}{6} = \frac{1}{2}$	0
Emily	$\frac{1}{6}$	$\frac{3}{6} = \frac{1}{2}$	$\frac{2}{6} = \frac{1}{3}$

結果を表 24 に示す．ここで，定義 3.3 の方法で risk を計算すると，表 25 のような結果が得られた．

表 17 をさらに一般化した後に計算した risk では，Bob が気管支炎の時，Alice が消化不良の時，David がインフルの時，Andy が気管支炎の時において，risk が 2/3 になる．表 17 の時には，全ての risk が 1/2 以下に抑えられていたが，表 17 をさらに一般化して計算すると一部の risk が 2/3 となり，一般化する前の risk より値が上がってしまった．通常，攻撃者が個人のプライバシを破りにくくする為に一般化を行うにも関わらず，上記の例では，一般化を行って risk を高くしてしまっている．

5.2　確率的リスク評価の問題点

先行研究 [1] で提案された一般化基準に従い，攻撃者に最大 $1/m$ の確率でしか病名を特定されないようにすることを目的に m-不変性を満たすよう作成された公開表列に対して，厳密な確率的リスク評価を行った場合 risk が $1/m$ より大きくなる例が存在する．それらの risk を $1/m$ 以下に抑えるには，公開表列の統合表と背景知識表を用いて，データを再構築する際にも m-不変性を満たすような全射のみを考える必要がある．

制約がなければ，表 4.6 の結果のように risk が $1/m$ より大きくなってしまう．公開者が攻撃者に制約を与えることで目標の評価値に抑制しているということは，攻撃者により多くの情報を与えることでリスクを下げているということであり，一般的な考え方とは矛盾している．なぜなら，一般的な考え方では攻撃者に情報を多く与えればプライバシ漏洩のリスクは高くなり，情報が少ないほどリスクは低くなると考えるからだ．
これは，情報が少ないのにもかかわらず元データに近い情報を得やすくなる為攻撃者にとっては好都合であるが，公開者にとってはより多くの情報を与えなければリスクを抑制することができない為不都合である．

また，5.1 節で行ったように元の一般化よりも QI 属性値の一般化の範囲を広げた場合，厳密な確率的リスク評価を行った場合 risk は元の一般化の時の risk よりも大きくなる例が存在する．一般的なリスクの考え方では，一般化の範囲を広げれば攻撃者は個人がかかるであろうと予測される病名数が多くなるので，プライバシはより保護され，プライバシ漏洩のリスクは低くなるか，元の一般化の時と変わらないかのどちらかであると考えられる．しかし，この場合でも，攻撃者への制約がなければ表 25 の結果のように元の一般化の時の risk よりも値が高くなってしまい，一般的な考え方と矛盾してしまう．

このように，確率的リスク評価には公開者が攻撃者に情報を与えることでリスクを抑制し目標の評価値になるようにしていることや，一般化を行いリスクを下げようとしているにもかかわらず，m-不変性を保持し続けなければリスクが高くなってしまう場合があるという一般的な直観と矛盾するという問題点が存在することがわかった．本来，攻撃者がどのように全射を再構築するかを公開者が強制することはできない．
したがって，これらのことを踏まえて考えると，今後は動的データセットのプライバシ漏洩のリスクを評価する指標として確率的リスクを直観と矛盾しない定義に変形し，その弱点を理解したうえで利用するという選択と，厳密な確率的リスクを制御できるよう従来の一般化をより強化したプライバシ保護方法を考案するという選択の両面からの検討が必要である．

5.3　確率的リスク評価の妥当性

本部分節では risk が 1 になる (100%の確率で病名を特定できる) 場合について考え，確率的リスク評価の妥当性について検討する．

表 26 を 1 回目の公開表 $T^*(1)$，表 27 を 2 回目の公開表 $T^*(2)$，表 28 を統合表 $U^*(2)$，表 29 を背景知識表 $B(2)$ とする．

攻撃者に制約を与えない厳密なリスク評価を行うため，表 29 から表 28 への制約なしの理にかなった全射を考える．表 29 から表 28 への制約なしの理にかなった全射関数の全てを表 30 に示す．表 30 の全射関数の数から定義 15 の risk を計算すると，表 31 のように各個人がそれぞれの病気になる確率を出すことができる．

上記の例で，Bob，Alice，C1 の 3 人は厳密な確率的リスク評価を行った場合 risk が 1 になっているので 3 人は病名が特定されている．このような場合を考えた時，一般的には以下のことが言える．

[定理 5.1]　厳密な確率的リスク評価法で攻撃者に病名を特定される確率が 100%ならば，従来のリスク評価法でも 100%の

表 26　1 回目の公開表 $T^*(1)$

Name	G.ID	Age	Zip.	Disease
Bob	1	[21,22]	[12k,14k]	消化不良
Alice	1	[21,22]	[12k,14k]	消化不良
Andy	2	[23,24]	[18k,25k]	インフル
David	2	[23,24]	[18k,25k]	消化不良

表 27　2 回目の公開表 $T^*(2)$

Name	G.ID	Age	Zip.	Disease
Bob	1	[21,22]	[12k,14k]	消化不良
C1	1	[21,22]	[12k,14k]	消化不良
David	2	[23,25]	[21k,25k]	インフル
Emily	2	[23,25]	[21k,25k]	消化不良

表 28　表 26 と表 27 の統合表 $U^*(2)$

No	G.ID	Age	Zip.	Disease	timestamp
1.1	1	[21,22]	[12k,14k]	消化不良	1
1.2	1	[21,22]	[12k,14k]	消化不良	1
1.3	2	[23,24]	[18k,25k]	インフル	1
1.4	2	[23,24]	[18k,25k]	消化不良	1
2.1	1	[21,22]	[12k,14k]	消化不良	2
2.2	1	[21,22]	[12k,14k]	消化不良	2
2.3	2	[23,25]	[21k,25k]	インフル	2
2.4	2	[23,25]	[21k,25k]	消化不良	2

表 29　背景知識表 $B(2)$

G.ID	Name	Age	Zip.	lifespan
*	Bob	21	12000	[1,2]
*	Alice	22	14000	[1,1]
*	Andy	24	18000	[1,2]
*	David	23	25000	[1,1]
1	C1	ϕ	ϕ	[2,2]
*	Emily	25	21000	[2,2]

表 30　全ての全射関数を示す表

1.1(消)	1.2(消)	1.3(イ)	1.4(消)	2.1(消)	2.2(消)	2.2(イ)	2.4(消)
Bob	Alice	Andy	David	Bob	C1	Emily	David
				C1	Bob	Emily	David
Bob	Alice	David	Andy	Bob	C1	David	Emily
				C1	Bob	David	Emily
Alice	Bob	Andy	David	Bob	C1	Emily	David
				C1	Bob	Emily	David
Alice	Bob	David	Andy	Bob	C1	David	Emily
				C1	Bob	David	Emily

表 31　表 30 を用いて risk を計算した厳密な確率的リスク評価結果

	消化不良	インフル
Bob	1	0
Alice	1	0
Andy	$\frac{4}{8}=\frac{1}{2}$	$\frac{4}{8}=\frac{1}{2}$
David	$\frac{4}{8}=\frac{1}{2}$	$\frac{4}{8}=\frac{1}{2}$
C1	1	0
Emily	$\frac{4}{8}=\frac{1}{2}$	$\frac{4}{8}=\frac{1}{2}$

確率で病名が特定される．

[証明] 厳密な確率的リスク評価を行なった時に risk が 1 になった場合において，そこから従来の評価法を算出する為に攻撃者に制約を与える理にかなった全射を算出することを考える．理にかなった全射を算出する為には，“ 公開者が与えた一般化基準を満たしていないものを制約なしの理にかなった全射から削除する ”ことになるが，risk が 1 になる場合はどの全射を削除することになっても，$n_{total.strict}$ の数が減れば，同じように $n_{breach.strict}$ の数も減るので結果的に risk は 1 になる．

定理 5.1 の対偶をとると以下が言える．

［系 5.1］ 従来のリスク評価法で攻撃者に病名を特定される確率が 100%でなければ，厳密な確率的リスク評価法でも病名を特定される確率は 100%でない．

系 5.1 より，従来の評価法で risk を計算した場合に risk が 1 でなく，100%の確率で病名が特定されなければ，厳密な確率的リスク評価法を用いても 100%病名が特定されることはないことがわかった．したがって，従来の評価法を用いて計算した risk と，厳密な確率的リスク評価法を用いて計算した risk を比較したとき，厳密な確率的リスク評価値の方が高くなってしまうかもしれないが，従来の評価法のリスク評価値をみることでなければ，厳密な確率的リスク評価値が 100%であるか否かを決定することはできる．

5.2 節では従来の評価法は公開者が攻撃者に一般化基準の情報を与えて攻撃者を制限しなければ一般的な直観と矛盾が生じてしまうなどの問題点を指摘したが，本節で述べたように従来の評価法で確実に病名が特定されなかったなら，厳密な確率的リスク評価法でも特定されることはないので，厳密な確率的リスク評価を無理に用いるよりも，現段階では確立された従来の評価法を用いる方が現実的であると言える．

6. ま と め

本研究では，m-不変性を用いた公開表列において，m-不変性を保持するという制限を攻撃者に与えて評価する従来の評価法と，攻撃者に対する制限をなくして評価する厳密なリスク評価を比較し，確率的リスク評価の妥当性について考察を行った．実際に比較してみると従来の評価法では目標の評価値になっていたものも，厳密なリスク評価を行うと目標の評価値にならない場合が存在し確実にリスクを正確に抑制することができなかった．従って，m-不変性を保持するという制限を攻撃者に与えなければ評価値までリスクを正確には抑制できない場合があることをがわかった．

実際に比較を行うと，攻撃者に情報を与えて評価値を抑制することは本来，攻撃者の行動は公開者が制限できないことと両立せず，また，情報を攻撃者に多く与えているという点では弱い部分もあるが，現段階では現実的な評価法だということがわかった．

今後の課題としては，動的データの公開表列を作成する上で，良い指標となる確率的リスク評価にかわる別のリスク評価方法

を提案し，プライバシデータ情報の種類によってリスク評価や一般化を使い分ける方法などを考える．さらに，文献 [2] で提案され，文献 [3] で厳密な確率的リスク評価方法との関係を考察した m-不変性を満たさない公開表列の安全性評価方法の妥当性を今回得られた確率的リスク評価の妥当性と比較しながら検討する．

文　　献

[1] Xiaokui Xiao and Yufei Tao“ m-Invariance : towards privacy preserving re-publication of dynamic datasets , ”Proc. of SIGMOD'07, pp.689-700 (2007).

[2] 沖田利絵子, 上土井陽子, 若林真一, “　動的データセットの再公開における精確な安全性の計算方法の提案　”, 第 7 回データ工学と情報マネージメントに関する論文集 (DEIM2015), G2-3 (2015).

[3] 上土井陽子, 堀内淳史, 沖田利絵子, 若林真一, “　動的データのプライバシ保護再公開における精確な安全性の評価について　”, 2016 暗号と情報セキュリティシンポジウム (SCIS2016), 2C1-4 (2016).