プライバシ保護データ公開

近年の発展のサーベイ

Privacy-preserving data publishing: A survey of recent developments

Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu

資料スライド：神嶌 敏弘

この資料について

この資料は次の『プライバシ保護データ公開 (Privacy-preserving Data Publication)』に関するサーベイの勉強会資料です

Privacy-Preserving Data Publishing:
A Survey of Recent Developments

Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu
ACM Computing Surveys, Volume 42 Issue 4 (2010)

※ この資料は概要のみですので, 詳細はこの原著を参照してください

目次

1. はじめに

Introduction

プライバシ保護データ公開

プライバシ保護データ公開

privacy-preserving data publishing (PPDP)

敵対的な環境で，データの有用性を保ったままデータを公開する手法

methods and tools for publishing data in a more hostile environment, so that the published data remains practically useful while individual privacy is preserved

背景：医療情報を交換するなど組織間でデータをやりとりする必要生じるが，これらは個人情報を含むのでプライバシ的に問題

データの収集と公開のシナリオ

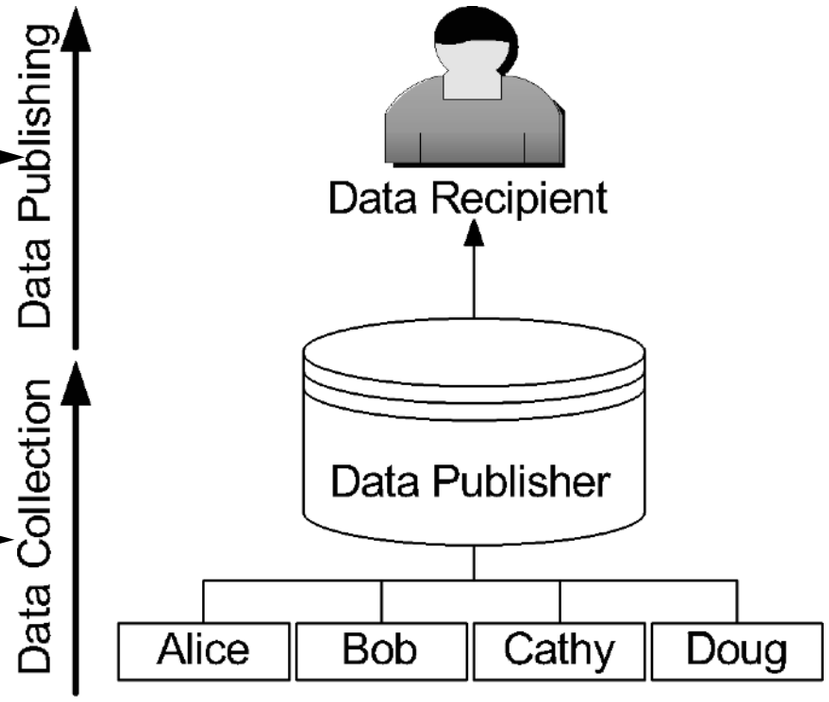
データ公開段階 (data publishing)

データ公開者が収集したデータを，実際にデータを分析する**データ受領者** (data recipient) に送る

データ収集段階 (data collection)

データ公開者 (data publisher) は，レコードが対象とする個人である**レコード所有者** (record owner) からデータを収集

※ データを集めた人が所有者ではない



データ公開者のモデル

データ公開者の二つのモデル

- 信頼できない (untrusted)

  信頼できない公開者は，センシティブ情報 (sensitive information) をレコード所有者から得ようとする

- 信頼できる (trusted)

  レコード所有者は公開者を信頼しており，公開者には個人情報を開示するが，信頼していないデータ受領者には開示したくない

このサーベイでは信頼できるデータ公開者のモデルを前提

※ 信頼できない公開者の場合は，暗号学的な対処，匿名通信，統計的手法などで匿名でレコードを収集する

公開者・受領者についての前提

- **データ公開者は非専門家**：公開者は，受領者のデータ分析や分析時期についての知識はなく，任意の分析を受領者ができるようにする．公開者が分析をプライバシを保護しつつ受領者に分析をアウトソーシングする場合も．
- **データ受領者は攻撃者となりうる**：このことは，受領者が攻撃者とならない暗号通信とは異なる点．プライバシと情報の効用を両立するのがPPDPの課題．
- **分析結果ではなくデータを公開**：分析結果の公開よりデータの公開の方が，データ操作・分析の自由度が大きいので，問題としては厳しい．しかし，分析結果を公開するには公開者が分析の専門家でなければならない．

レコードの真実性

レコードの真実性 (truthfulness at the record level)

データ表中のレコードが実在の個人と対応していることで，次のような操作で失われる

- 元データと同じ統計的性質をもつデータを人工的に生成
- 無作為にデータを摂動・ノイズを付加した場合

- 薬の副作用を調べる場合などは，真実性がなく実在しないレコードがあると，副作用の事例としては意味をなさない
- 暗号化による公開では，属性や値が具体的に何を意味するのかという，データのセマンティクスが分からない

データ表の構成要素

- **明示的識別子** (explicit identifier)：米の社会保障番号などレコード所有者を明示的に特定できる属性
- **擬識別子** (quasi identifier; QID)：潜在的にレコード所有者を特定できる属性の集合
  - **擬識別子グループ** (qid group)：擬識別子を組み合わせた値が同じ値 qid になる，データ表中の全てのレコードを集めたグループ
- **センシティブ属性** (sensitive attribute)：病歴，収入，障害などの配慮が必要な個人についての情報を含む属性
- **非センシティブ属性** (non-sensitive attribute)：上記のどれにも当てはまらない属性

※ これらは互いに疎で，各レコードはそれぞれ別の所有者のものであるとするのが一般的.

匿名化

[Cox 1980; Dalenius 1986]

匿名化 (anonymization)

特定可能性とセンシティブ属値の両方，もしくはいずれかを，センシティブ属性を分析するために保持しつつ秘匿する

明示的識別子は削除するが，それでも他のデータと併せると個人情報を，擬識別子を利用して知ることができる事例 [Sweeney 2002a]

- 公開医療データと公開投票者リストの生年月日・性別・住所郵便番号から前マサチューセッツ州知事 William Weld の個人情報が分かった
- 生年月日・性別・住所郵便番号の3属性を擬識別子とすると，米国人の87%は一意に特定可能

その他の用語

- **被攻撃者** (victim)：擬識別子が知られていて，攻撃者が情報を得ようとしているレコード所有者
- **匿名化操作** (anonymization operation)：元の擬識別子 QID の，レコードが特定されないような対応する擬識別子 QID' への変換，元データの統計量に基づくデータの生成，そして元データに対するノイズ付加など
- **匿名化問題** (anonymization problem)：プライバシモデルに基づきそのプライバシ要求を満たしつつ，効用を最大化するように変換したデータ表 T を生成する問題
- **情報計量** (information metric)：匿名化した表の効用を測る計量

本サーベイの対象範囲

- **プライバシ保護データマイニング**：PPDPと異なり，手法は特定のマイニングタスクと密接に結びついており，レコードの真実性もない.
- **統計的開示制御 (statistical disclosure control)**：前者のみ扱う
  - **非対話的クエリモデル**：受領者のシステムへのクエリは一つだけ. 受領者が自身のタスクを満たすクエリを作るのは難しい
  - **対話的クエリモデル**：受領者や攻撃者は，それまでに得たクエリへの応答結果をふまえて，一連のクエリを投入可能. ただし，プライバシ保護を考えるときは，クエリ数は，データに対して線形よりも少ない（でなければ，$1-o(1)$ のデータが復元できてしまう）

※ PPDPはSDCと較べて，背景攻撃，センシティブ属性の推定，一般化，多様な効用尺度を扱う.

2. 攻撃モデルとプライバシモデル

Attack Models and Privacy Models

プライバシ保護とは？

[Dalenius 1977] の定義

- 公開データを参照できたとき，たとえ他の情報源から得た背景知識を攻撃者が知っていたとしても，データを参照出来なかった場合と比べて被攻撃者について攻撃者はいかなる追加情報も得ることはない

[Dwork 2006] の否定的な結果

- このような絶対的なプライバシ保護は背景知識が存在すると不可

例：被攻撃者の年齢はアメリカ女性の平均年齢より5歳若いという知識などがあると年齢が分かってしまう

PPDPでは，攻撃者のレコード所有者に対する背景知識を制限した，より現実的なプライバシ保護を想定

攻撃モデル

被攻撃者の擬識別子値を知っている攻撃者が，その被攻撃者を，公開データのレコードと結び付ける

- **レコードリンク (record linkage)**：公開データのレコードと結び付ける
- **属性リンク (attribute linkage)**：センシティブ属性値と結び付ける
- **データ表リンク (table linkage)**：公開データ表に被攻撃者が存在するかどうかを明らかにする

背景知識と比べて公開データから攻撃者があまり情報を得られないようにするとという無情報価値原理 (uninformative principle) を満たす

- **確率的攻撃 (probabilistic attack)**：事前と事後の信念分布に大きな変化が生じる．QID や センシティブ属性の区別はしないのが一般的

攻撃モデル・プライバシモデル一覧

プライバシモデル	攻撃モデル			
	レコードリンク	属性リンク	データ表リンク	確率的攻撃
k 匿名性	●			
マルチ k 匿名性	●			
l 多様性	●	●		
確信度限定		●		
(α, k)多様性	●	●		
(X, Y)プライバシ	●	●		
(k, e)匿名性		●		
(ε, m)匿名性		●		
個人化プライバシ		●		
t 近接性		●		●
δ 存在性			●	
(c, t)分離性	●			●
ε 差分プライバシ			●	●
(d, γ)プライバシ			●	●
分布プライバシ			●	●

レコードリンク攻撃

レコードリンク攻撃 (record linkage attack)

- 公開データ表 T の少数のレコード群（対象グループ）が，擬識別子のある値 qid によって特定されること
- 擬識別子が qid である被攻撃者は，この対象グループと関連付けられ，攻撃者は被攻撃者の追加情報を得て，被攻撃者のレコードを一意に特定する可能性が生じる.

レコードリンク攻撃

病院の元データ

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	30	Flu
Writer	Female	30	HIV
Dancer	Female	30	HIV
Dancer	Female	30	HIV

攻撃者所有の外部データ

Name	Job	Sex	Age
Alice	Writer	Female	30
Bob	Engineer	Male	35
Cathy	Writer	Female	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Engineer	Male	38
Gladys	Dancer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32

他に同じ組み合わせはない

- 病院（公開者）は元データを公開したい
- 攻撃者は〈Job, Sex, Age〉の擬識別子が共通の外部データを所持
- 〈Job, Sex, Age〉＝〈Lawyer, Male, 38〉のレコードは他にはないので，Doug は HIV と分かる

k 匿名性

[Samarati & Sweeney 1998]

k 匿名性 (k-anonymity)

- 擬識別子によるレコードリンクを防ぐためのプライバシモデル
- 擬識別子の値が同じレコードが少なくとも k–1 個存在することがプライバシ要求

- 被攻撃者がある特定のレコードと結び付けられる確率はたかだか $1/k$

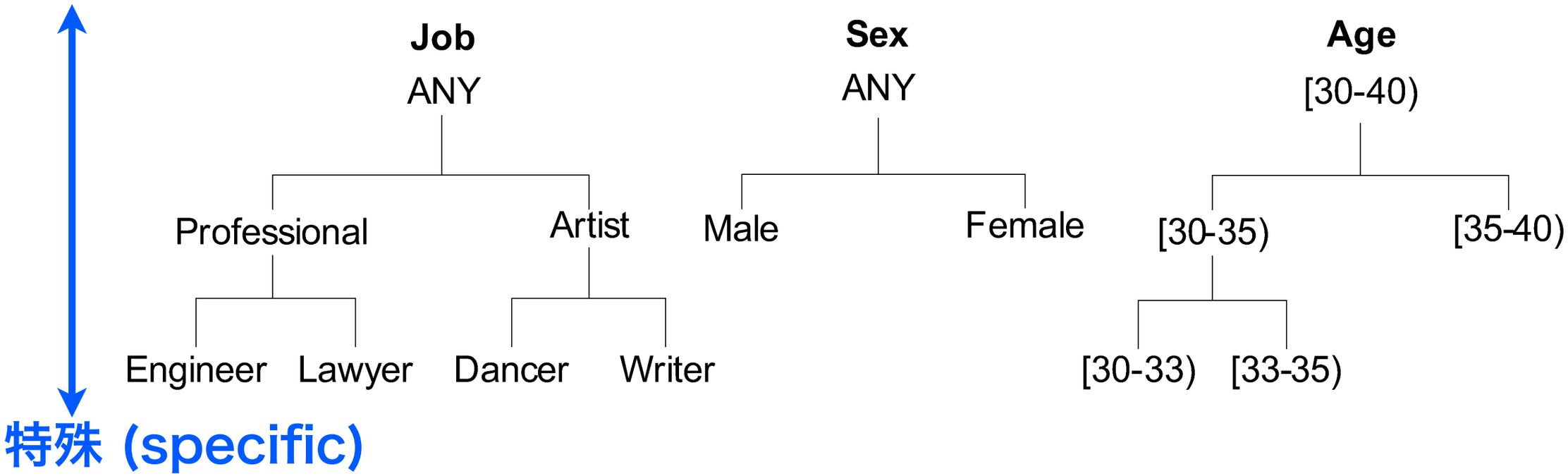
  なぜならあるレコードは他の k–1 個のレコードと擬識別子によっては区別できない

- 属性リンクに対してはk匿名性は無力

  擬識別子のみを考慮し，センシティブ属性は考慮していない

k 匿名性

分類木 (taxonomy tree)



- 分類木 (taxonomy tree) はより，一般的で抽象的な値から，最も特殊で具体的な値までを，木の形式にまとめたもの
- この例では，名義変数の Job には，Engineer のより一般的な値として Professional があり，数値変数の Age では一般・特殊関係は区間の包含関係になっている

k 匿名性

3匿名化した公開データ

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

3レコードうちどれが対応するか不明

攻撃者所有の外部データ

Name	Job	Sex	Age
Alice	Writer	Female	30
Bob	Engineer	Male	35
Cathy	Writer	Female	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Engineer	Male	38
Gladys	Dancer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32

- 擬識別子〈Job, Sex, Age〉の値が同一のレコードが3個以上の3匿名
- 3匿名化したことで，Dougの擬識別子の値〈Lawyer, Male, 38〉は〈Professional, Male, [35-40)〉と一般的な値と置き換えられた
- 分類木に基づいて値を対応付けても，Doug の擬識別子値は三個のレコードと対応するので，この3人の所有者のだれが Doug か不明に

k 匿名性

- **効用とプライバシのトレードオフ**
  より多くの属性を潜在的な擬識別子とみなせばプライバシは向上するが，k 匿名化にはより多くの情報を秘匿して効用を下げる必要
- 複数の集合が擬識別子である場合には，部分集合性が冗長な属性の排除に有用
  - **部分集合性 (subset property)**：$QID' \subseteq QID$ について，データ表 T が QID に関して k 匿名であるなら，T は QID' に対して k 匿名
- **同一所有者のレコードが複数ある場合は無力**
  例：このとき，QID が同じであるグループは k 人より少なくなり，保護されなくなる

例：QID に病名が含まれる場合，複数の病気にかかっている人は複数のレコードがある可能性がある

(X,Y)匿名性

[Wang+ 2006]

(X,Y)匿名性 ((X,Y)-anonymity)

- 互いに疎な属性集合 X と Y に基づく概念
- データ表 T が (X,Y)匿名であるとは，ある整数 k について

$$a_{Y(X)} = \min\{a_{Y(x)} \mid x \in X\} \leq k$$

- ただし，X のある値 x と Y について x の匿名度 $\mathrm{a}_{Y(x)}$ とは，x と同時に生じうる Y の値の種類数

- 属性 X の値は，少なくとも k 種の Y の属性値と結びついている
- X が擬識別子で〈Job, Sex, Age〉Y がセンシティブ属性〈Disease〉とすると，擬識別子の値からセンシティブ属性値を推論しにくくなる
- k 匿名は同じ所有者のレコードが複数あると問題だが，X を擬識別子，Y を明示的識別子とし(X,Y)匿名にすれば問題なくなる

マルチ関係 k 匿名性

[Nergiz+ 2007]

マルチ関係 k 匿名性 (multi-relational k-anonymity)

- ここまではデータ表が一つだったが，複数のデータ表を含むマルチ関係DBの方が一般的

  ➡この状況を考慮したプライバシモデル

- 明示的識別子を含むデータ表に加え，これと関係のある複数のデータ表に擬識別子やセンシティブ属性など秘匿すべき情報がある
- 明示的識別子を含むデータ表と，これと関係がある複数のデータ表全てjoinして作ったデータ表でのk匿名性

- ある所有者のレコードは PT 中では一つだが，join したデータ表で複数になる

  ➡ join 後のデータ表で，擬識別子が同じレコードが k 個以上ではなく，擬識別子が同じレコードのグループ中に k 人以上のレコード所有者が含まれるようにする

擬識別子選択のジレンマ

擬識別子選択のジレンマ

どの属性を，擬識別子，センシティブ属性，非センシティブ属性にするかは，プライバシと効用のトレードオフがあり困難な問題

- 擬識別子＝攻撃者が他で入手できそうな情報
  センシティブ・非センシティブ属性＝その他
  - 擬識別子とすべきものを他の属性にすると，リンク攻撃を受ける
  - センシティブ属性を擬識別子とすると，他の擬識別子からセンシティブ情報が推論できたり，不要な情報損失を生じる [Aggarwal 2005]

- ほとんどのレコードを一意に特定できるような属性の最小集合を選ぶという擬識別子の選択手法 [Motwani+ 2007]
  - この方法では攻撃者の知識が不要になるが，そのためにリンク攻撃の対象となる属性を擬識別子としない可能性は残る

レコードリンク用手法の限界

レコードリンク攻撃を抑制できてもセンシティブ属性値が漏洩する場合

3匿名性が既に達成されているデータ表だが…

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

擬識別子＝⟨Artist, Female, [30-35)⟩

センシティブ属性値の3/4はHIV

レコード所有者が特定されなくてもそのセンシティブ属性値が推定可

↓

属性リンク攻撃

属性リンク攻撃

属性リンク攻撃 (attribute linkage attack)

- 攻撃者は，被攻撃者のレコードを厳密には特定できないが，その被攻撃者が所属するグループのセンシティブ属性値から，被攻撃者のセンシティブ属性値を推定
- [Clifton 2000] では，公開データの大きさを制限する
  ➡ 効用は低下
- 以下に示す方法は，主に擬識別子とセンシティブ属性の値の相関を減らす

属性リンク攻撃

病院の公開データ

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	30	Flu
Writer	Female	30	HIV
Dancer	Female	30	HIV
Dancer	Female	30	HIV

攻撃者所有の外部データ

Name	Job	Sex	Age
Alice	Writer	Female	30
Bob	Engineer	Male	35
Cathy	Writer	Female	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Engineer	Male	38
Gladys	Dancer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32

- 攻撃者は，公開データから擬識別子＝〈Dancer, Female, 30〉であればDiseaseが100%の確率でHIVと分かる
- 攻撃者が，Emily のレコードが公開データに含まれることを知っていていれば，自身がもつデータとの擬識別子の一致により，Emilyの病名がHIVであると分かる

l 多様性

[Machanavajjhala 2006,2007, (Ohrn+ 1999)]

l 多様性 (l-diversity)

- 擬識別子の値が同じになる各擬識別子グループ中では，センシティブ属性は少なくとも l 種類の well-represented な値を含んでいる
- “well-represented” の定義によって幾つもの l 多様性に分かれる

相違 l 多様性 (distinct l-diversity)
(a.k.a. p-sensitive k-anonymity [Truta+ 2006])

- well-represented を，l 種の異なる値であることとした『普通の』l 多様性
- $k=l$ のとき，すなわち k多様性は自動的に成り立つが，グループ内であるセンシティブ属性値が頻出すると確率的推論を防げない

エントロピー l 多様性

エントロピー l 多様性 (entropy l-diversity)

- 擬識別子グループ内で，センシティブ属性値の分布のエントロピーが $\log l$ 以上となるようにする，すなわち，分布を一様に近づける

$$-\sum_{s \in S} \Pr(qid, s) \log \Pr(qid, s) \leq \log l$$

ただし，$\Pr(qid, s)$ は，擬識別子値が qid の擬識別子グループ内で，センシティブ属性値が s である割合

- 擬識別子の値から，センシティブ属性値が確率的に推論されないようにすることをより厳密に保証
- エントロピーの数値からは漏洩リスクの確率が直感的に分かりにくく，機密性や値の分布に基づいて保護水準 l を決めるのが困難

再帰的 (c, l) 多様性

再帰的 (c, l) 多様性 (recursive (c, l)-diversity)

- 高頻度の値がさらに高頻度になりすぎないように，低頻度の値がさらに低頻度になりすぎないように保証
- データ表中の，各擬識別子グループについて擬識別子グループ内で再頻出のセンシティブ属性値の個数が，頻度が最下位の $m - l + 1$ 個のセンシティブ属性値の個数の総和より小さい

$$f_1 < c \sum_{i=l}^{m} f_i$$

※ f_i はグループ内で第 i 番目に頻出するセンシティブ属性値の個数で，m はセンシティブ属性値の種類数

- 攻撃者が外部知識によって候補属性値をいくつか除外することに成功しても，残りの候補値の推定は困難なまま

属性リンク攻撃での背景知識

- positive disclosure-recursive (c,l)-diversity と negative / positive disclosure-recursive (c,l)-diversity [Machanavajjhala 2006,2007]
  - 被攻撃者を含む擬識別子グループのセンシティブ属性値が {Flu, Cancer, HIV} であるとき，被攻撃者にインフルエンザの症状がなければ Flu を除外できる
- k 単位の情報：最も高確率で判明するデータ表中のデータ所有者のセンシティブ属性値 [Martin+ 2007]
  - 1単位の情報：4人が同居しているとき他の3人が Flu なら，残りのもう一人も，センシティブ属性値の候補のうち高確率で Flu と予測可能

l 多様性の限界

- l 多様性は暗黙的にセンシティブ属性値が一様分布していると仮定
- この仮定が成立しないときには効用が低下する

例 [Domingo-Ferrer+ 2008]

- レコード数が1000件のデータ表
- 二値のセンシティブ属性 HIV があり，HIV＝Yes のレコードが5件だけ

⬇

- $k = l$ の緩い条件で k 匿名性と l 多様性を同時に満たす
- どの擬識別子グループにも必ず HIV=Yes のレコードが必要

⬇

- 全体で擬識別子グループ数はたかだか5個までとなり，効用の損失が大きい

確信度限定

[Wang+ 2005, 2007]

確信度限定 (confidence bounding)

- QID：擬識別子，s：センシティブ属性値，h：しきい値
- $\mathrm{conf}(qid \to s)$ はある擬識別子値が qid であるときにセンシティブ属性値が s になる割合で，その最大値を $\mathrm{Conf}(QID \to s)$
- **プライバシ・テンプレート** $\langle QID \to s, h \rangle$ を満たす：

$$\mathrm{Conf}(QID \to s) \leq h$$

- (c, l)多様性やエントロピー l 多様性より，確信度の限界は直感的
- いろいろな擬識別子の組み合わせそれぞれに対して異なるしきい値を設定可能
- (c, l)多様性のような，背景知識があるときのリンク攻撃を防ぐ効果はない

確信度限定

[Wang+ 2005, 2007]

- 擬識別子：QID＝{Job, Sex, Age} の場合
- プライバシ・テンプレート〈QID→HIV, 10%〉：どの擬識別子グループに所属していても，そのことをもってセンシティブ属性値が HIV であることが推論できる確率はたかだか10%

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

4レコードの
うち3件

- 擬識別子が〈Artist, Female, [30-35)〉のとき HIV である確信度は75%なのでこのプライバシ・テンプレートを満たさない

(X, Y)プライバシと(α, k)匿名性

[Wang+ 2006]

(X, Y)プライバシ ((X, Y)-privacy)

- (X, Y)匿名性と確信度限定の組み合わせ
- X の各グループ $\mathbf{x}$ について，Y の値は少なくとも k 種類で，かつ $\mathrm{conf}(\mathbf{x} \to \mathbf{y}) \leq h, \mathbf{y} \in Y$ が成立

[Wong+ 2006]

(α, k)匿名性 ((α, k)-anonymity)

- (X, Y)プライバシと類似
- 全擬識別子が qid のグループは k 個以上のレコードがあり，かつ $\mathrm{conf}(qid \to s) \leqq \alpha$ を満たす

- (X, Y)プライバシも(α, k)匿名性も，Y や擬識別子グループに，k 個レコード以上としてリンク攻撃を防ぐとと共に
  - ➡ Y やセンシティブ属性の値に偏りがあるために生じる漏洩を防ぐ

センシティブ属性が数値変数の場合

[Zhang+ 2007]

(k, e)匿名性 ((k, e)-anonymity)

- データ表 T を，センシティブ属性の範囲が e より大きな k 種類の値をとるようなグループに分ける

近接性攻撃 (proximity attack) [Li+ 2008]

センシティブ属性がある狭い範囲にあるという背景知識から近似的値を得る

- 擬識別子グループ10レコードを含む
- 7種のセンシティブ属性値，9レコードは [30-35] で残りは 80
  範囲は 80 − 30 = 50 と広いが，90%の確信度で [30-35] と分かる

[Li+ 2008]

(ε, m)匿名性 ((ε, m)-anonymity)

- 近接性攻撃を防ぐため，センシティブ属性値がある値 s の近傍 $[s-\varepsilon, s+\varepsilon]$ にある確率をたかだか $1 / m$ とする

歪み攻撃

[Li+ 2007]

歪み攻撃 (skewness attack)

- 属性リンク攻撃の一種で，l 多様性では対処できない
- 全体でのセンシティブ属性値の分布と，擬識別子グループでの分布に大きな歪みがあるため，その値が予測できる

- データ表のセンシティブ属性値の95%は Flu で，残り5%が HIV
- ある擬識別子グループ50%のセンシティブ属性値は Flu で，残り50%がHIV
  ➡ 2 多様性の条件を満たす
- この擬識別子グループのレコード所有者は，全体の5%に比べて50%と相対的に高い確率で HIV であると分かる

t 近接性

[Li+ 2007]

t 近接性 (t-closeness)

- 歪み攻撃に対する対処
- 任意の擬識別子グループについて，全体のセンシティブ属性の分布とそのグループ内での分布との差を Earth Mover距離 で測ったとき，その差がたかだか t であること

Earth Mover距離：一方の分布をもう一方の分布に変化させるために，必要な確率質量と移動距離の総和

t 近接性の制限

- センシティブ属性ごとに保護水準を設定できない
- 全擬識別子グループでセンシティブ属性の分布を等しくするので，擬識別子とセンシティブ属性の相関関係の情報が失われる
  - この制限への対処は，リスクを代償としてしきい値を変えるか

    [Domingo-Ferrer+ 2008]，確率的プライバシモデルに変えるかしかない

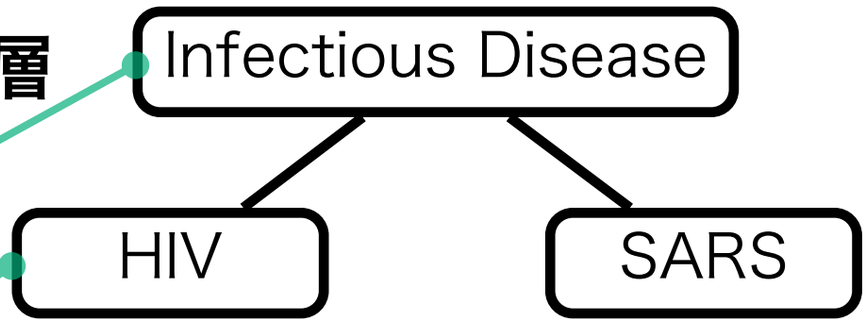
個人化プライバシ

[Xiao+ 2006]

個人化プライバシ (personalized privacy)

- レコード所有者が個別にプライバシ水準を決定可能
- センシティブ属性の値に階層構造があり，どの部分木の値まで公開してよいかを個別に指定
- 指定した部分木以下のどの値であるか分かる確率をしきい値以下に

センシティブ属性の階層



- **Alice**：Disease の値が Infectious Disease であることは明かしてもよいが，具体的な病名は公開したくない
- **Bob**：具体的な病名を公開してもよい

個人化プライバシ

- 確信度限定も個人化プライバシも，擬識別子グループのセンシティブ属性値の推定確率を制限する点は同じ
  - 確信度限定では全体で一律の確率しきい値を，個人化プライバシでは所有者ごとに変更可能 ➡ より高いデータの効用を達成可能

各所有者が適切な保護水準を選択するのは実際には難しい

- 適切な水準は，センシティブ属性の分布に依存するが，データの公開前ではそうした情報は得られない

  例：ありふれた病気であれば保護水準を下げてもリスクは増えない
- こうした不確実な状態では保護水準は高めに設定されやすく，そのためデータの効用が下がってしまう

FF 匿名性

[Wang+ 2009]

FF匿名性 (FF anonymity)

- 擬識別子とセンシティブ属性に重複がある場合に対処
- 任意の属性値集合から，センシティブ属性値を推定するときの確信度をしきい値以下にする

freeform 攻撃

- X の値を知ることで，高い確率で s の値を知ることができる状況が存在する
  - X：データ表中の任意の属性集合 X
  - s：X に含まれているセンシティブ属性

データ表リンク攻撃

データ表リンク攻撃 (table linkage attack)

- レコードリンクや属性リンクでは，被攻撃者が所有するレコードがデータ表中に存在することを，攻撃者が知っている前提
- この存在するかどうかの情報自体を得ることが，データ表リンク攻撃の目的

データ表リンク攻撃

匿名化して公開したデータ

Job	Sex	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

外部公開データ

Name	Job	Sex	Age
Alice	Artist	Female	[30-35)
Bob	Professional	Male	[35-40)
Cathy	Artist	Female	[30-35)
Doug	Professional	Male	[35-40)
Emily	Artist	Female	[30-35)
Fred	Professional	Male	[35-40)
Gladys	Artist	Female	[30-35)
Henry	Professional	Male	[35-40)
Irene	Artist	Female	[30-35)

- 公開者は，3匿名化処理をしてデータを公開した
- 攻撃者は外部公開データから被攻撃者の情報を得る
- 匿名化公開データのレコードは，外部公開データに存在するという情報は攻撃者は知っている
- 〈Artist, Female, [30-35)〉のレコードは，匿名化公開データでは4件，外部公開データでは5件なので，Alice の存在確率は 0.8

δ 存在性

[Nergiz + 2007]

δ 存在性 (δ-presence)

- 潜在的被攻撃者がデータ表に含まれる確率を $\delta=(\delta_{\min}, \delta_{\max})$ に限定
- 外部公開データ表 E と非公開のデータ表 T があり，$T \subseteq E$ の関係
- 匿名化して公開したデータ表 T' が$(\delta_{\min}, \delta_{\max})$存在性を満たす条件

$$\delta_{\min} \leq \Pr[t \in T \mid T'] \leq \delta_{\max}, \forall \in E$$

（E 中のレコードの所有者が T 中に存在する確率を限定）

- δ存在性は間接的にレコードリンクや属性リンクも抑制できる

  ⬅ δ%しか存在すると確信できないなら，レコードリンクや属性リンクの成功確率もたかだかδ%に制限される
- δ 存在性は比較的漏洩リスクのないモデルだが，攻撃者が知ることのできる公開データ表 E に関する知識の仮定は非現実的

確率的攻撃

確率的攻撃 (probabilistic attack)

- レコード，属性，データ表を厳密にリンクするではなく，公開データを参照することで被攻撃者のセンシティブ情報についての確率的信念がどう変化するかを扱う
- 確率的攻撃プライバシモデルでは，公開データ参照の事前と事後の信念の差が小さいという無情報原理 (uninformative principle) [Machanavajjhala+ 2006] を達成することが目的

(c, t)分離性

[Chawla+ 2005]

(c, t)分離性 ((c, t)-isolation)

- データを公開することで，攻撃者がいずれかのレコード所有者を分離しやすくすることを防ぐ
- 超球 $\mathrm{B}(\mathbf{p}, c\,\delta_p)$ 内にデータ表の点がたかだか t 個
  - $\mathbf{p}$：統計的DB中の被攻撃者のデータ点
  - $\mathbf{q}$：背景知識と公開データから攻撃者が推定した被攻撃者のデータ点
  - δ_p：$\mathbf{p}$ と $\mathbf{q}$ の間の距離
  - $\mathrm{B}(\mathbf{x}, r)$：$\mathbf{x}$ を中心とする半径 r の超球

- (c, t)分離性はレコードリンクを防いでいるとみなせる
- 距離が容易に定義できるので数値属性に適す

ε差分プライバシ

[Dwork 2006]

ε差分プライバシ (ε-differential privacy)

- レコード所有者がデータ表に加わってもリスクは大きくは増加しないことを保証する
- 単一レコードの削除や追加が，分析の出力結果に大きな影響を与えないことを保証

➡ あるデータ所有者が，データ公開者に情報を提供しなくても，出力結果に大きくは影響しない

- レコードの有無自体を保護しているので，攻撃者が任意の背景知識を使っても安全
- レコード数 n に対してクエリ数が sublinear ならノイズの大きさはたかだか $O(\sqrt{n})$

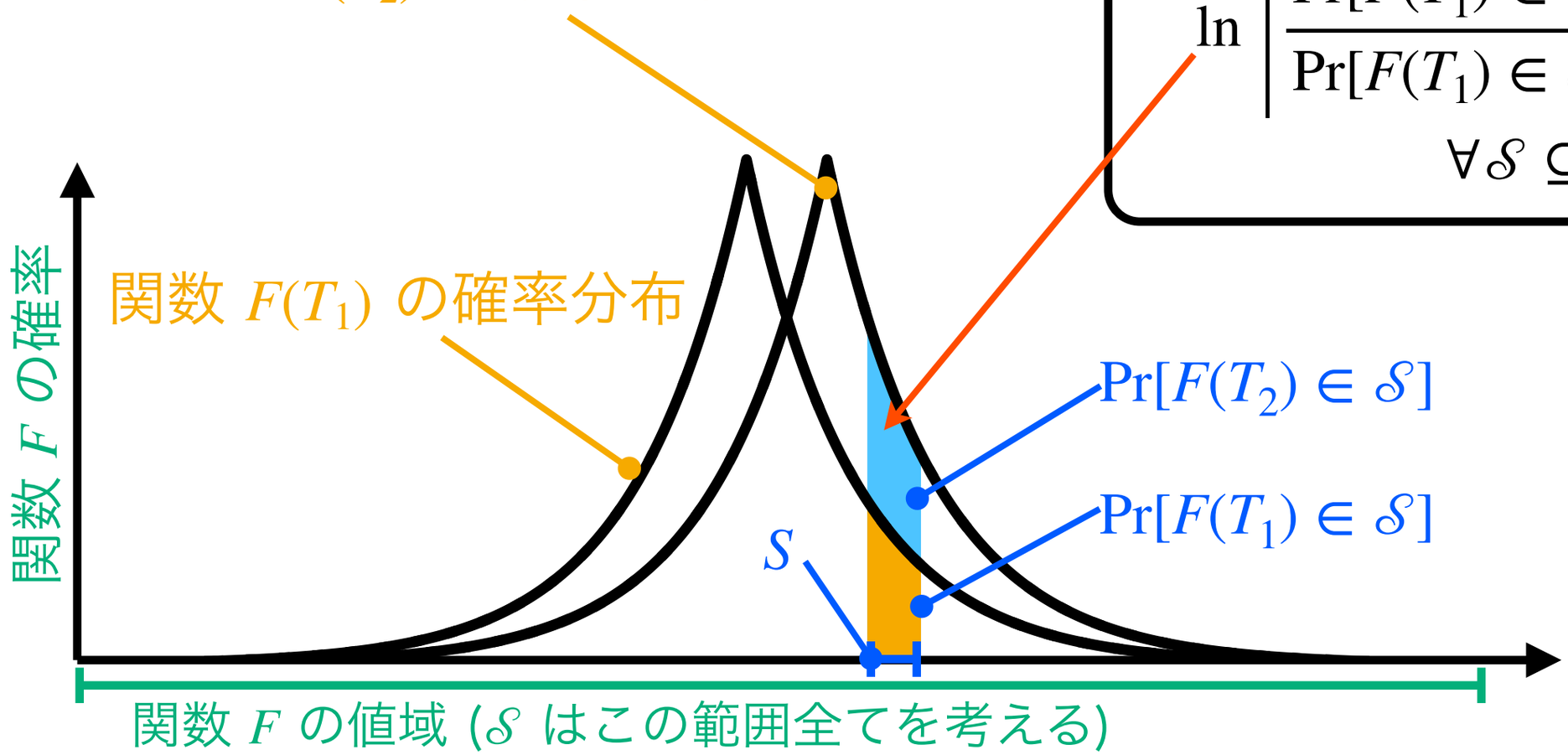
ε差分プライバシ

- 関数 $F(T)$：データ表 T の確率的関数
  データ表 T の統計量を計算して，それに無作為なノイズを加える
- T_1，T_2：1レコードを除いては全く同一のデータ表

ε差分プライバシの条件

$$\ln \left| \frac{\Pr[F(T_1) \in \mathcal{S}]}{\Pr[F(T_1) \in \mathcal{S}]} \right| \leq \varepsilon$$

$$\forall \mathcal{S} \subseteq \mathrm{Range}(F)$$



ε差分プライバシ：メカニズム

[McSherry&Talwar “Mechanism Design via Differential Privacy” FOCS2007]

- $q(T, S)$：データ表 T とパラメータ S の任意の実数関数

  S は分布 $\Pr[S]$ に従い，その地域は離散でも連続でもよい

- Δq：任意の T_1 と T_2 の対の集合上の，関数 q の差の絶対値の最大値

指数メカニズム

次の確率で S を出力すると，ε差分プライバシが達成される

$$\Pr[S] \propto \exp(\varepsilon\, q(T, S) /\ 2\, \Delta q)$$

ラプラスメカニズム

$q(T, S) = -|F(T) - S|$ とおくと，ラプラス分布に従うノイズを加えた，

$\varepsilon/2$ 差分プライバシを達成するラプラスメカニズムになる

$$\mathrm{F}(T) + \mathrm{Laplace}(0, 2\, \Delta q / \varepsilon)$$

(d, γ)プライバシ

[Rastogi+ 2007]

(d, γ)プライバシ ((d, γ)-privacy)

- d 独立な攻撃者にとっての，被攻撃者の存在に関する事後分布は γ で抑えられ，かつ事前確率より事後確率が大きくは減らない
- **d 独立 (d-independent)**：公開データを観測する前に，被攻撃者がデータ表に存在するかどうかについての，攻撃者の信念が d 以下

- 攻撃者の能力をモデル化し，プライバシ保護と効用の理論的な上界を示している
- プライバシと効用の相応なトレードオフが可能なのは，事前確率が小さいときに限られる
- 差分プライバシとは異なり，レコード間の独立性や，攻撃者の事前の信念について仮定がある

分布プライバシ

[Blum+ 2008]

- 学習理論に基づいたプライバシモデル
  - データ表は，ある分布からの標本と考え，その標本からクエリ値を予測する関数を学習する問題とみなす
  - 多項式個の標本で学習できるなら十分な効用があるデータと考え，この条件の下で差分プライバシの条件を満たすようにする
- 離散値のクエリでは，誤差はクエリのVC次元に線形で増加
- 連続値のクエリでは，最悪の場合には不可能

※ サーベイの記述では分からなかったので，原著の論文のアブストの概要より

3. 匿名化操作

Anonymization Operations

匿名化操作

匿名化操作 (anonymization operation)

一連の匿名化操作を適用してプライバシ要求を満たす

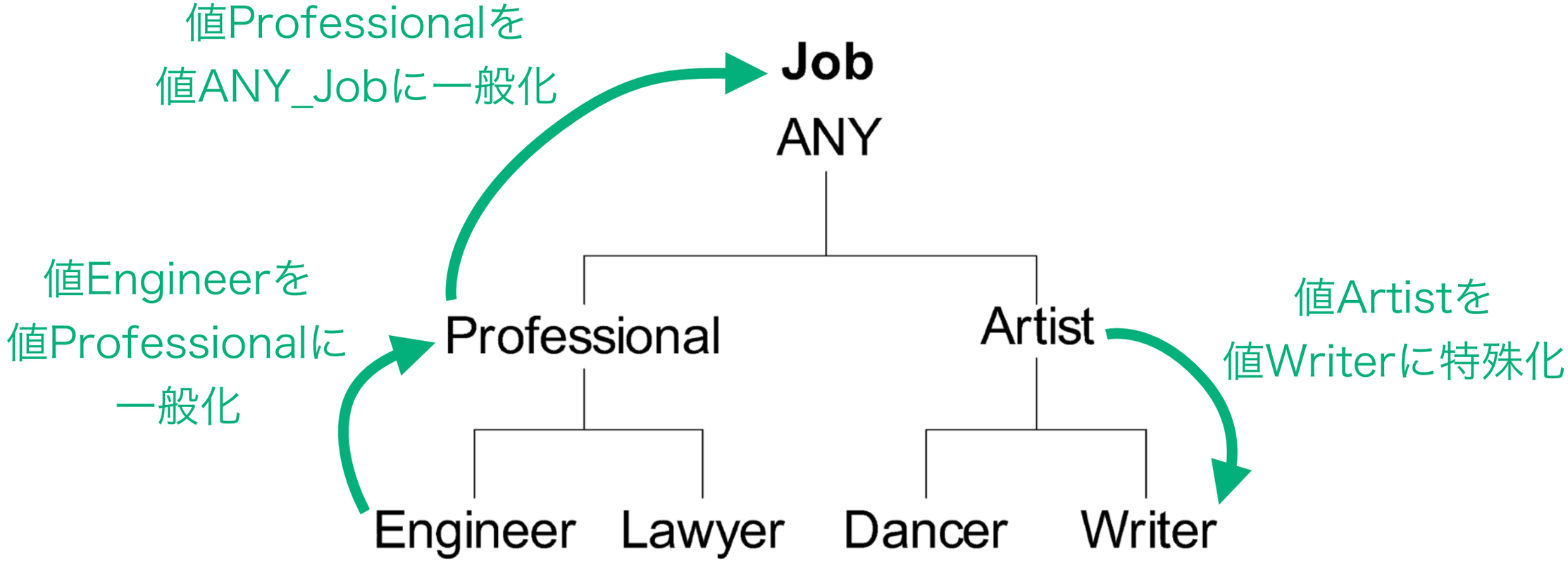
- **一般化 (generalization)** と **抑制 (suppression)**： QID などの特徴を別の値と置換する
- **分解 (anatomization)** と **順列置換 (permutation)**： グループへの分解と要素の入れ替えである順列置換によって QID とセンシティブ特徴の間の相関を消す
- **摂動 (perturbation)**：元データの統計量を保存するように，ノイズ付加，値の集約，値の置換，値の人工生成を行う

一般化

一般化 (generalization)

- カテゴリ値の場合：値の階層でより一般的な値と置換する
- 連続値の場合：その値を含むより幅の広い区間と置換したりする

※ 逆の演算は **特殊化 (specialization)**



全体一般化

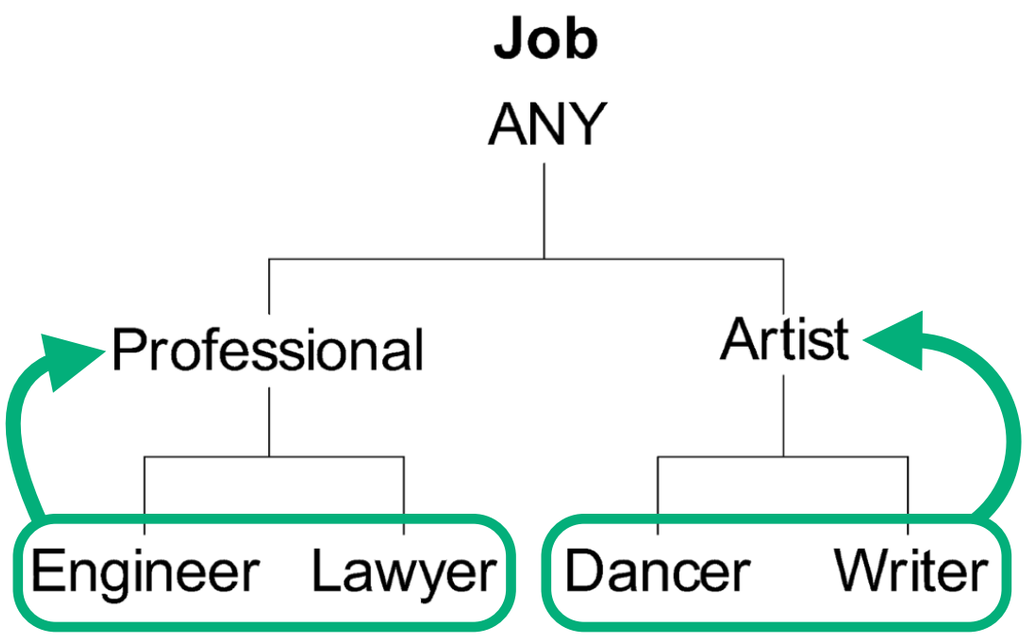
[LeFevre+ 2005, Samarati 2001, Sweeney 2002b]

全体一般化 (full-domain generalization)

属性を全て同じ水準に揃えて一般化する

- 実行に必要な探索空間は小さい
- 分類の粒度が大きいので，データの歪みは大きくなる

例：Engineer や Lawyer を Professional に一般化するのなら
Dancer や Writer も必ず同じ水準の Artist に一般化する



部分木一般化

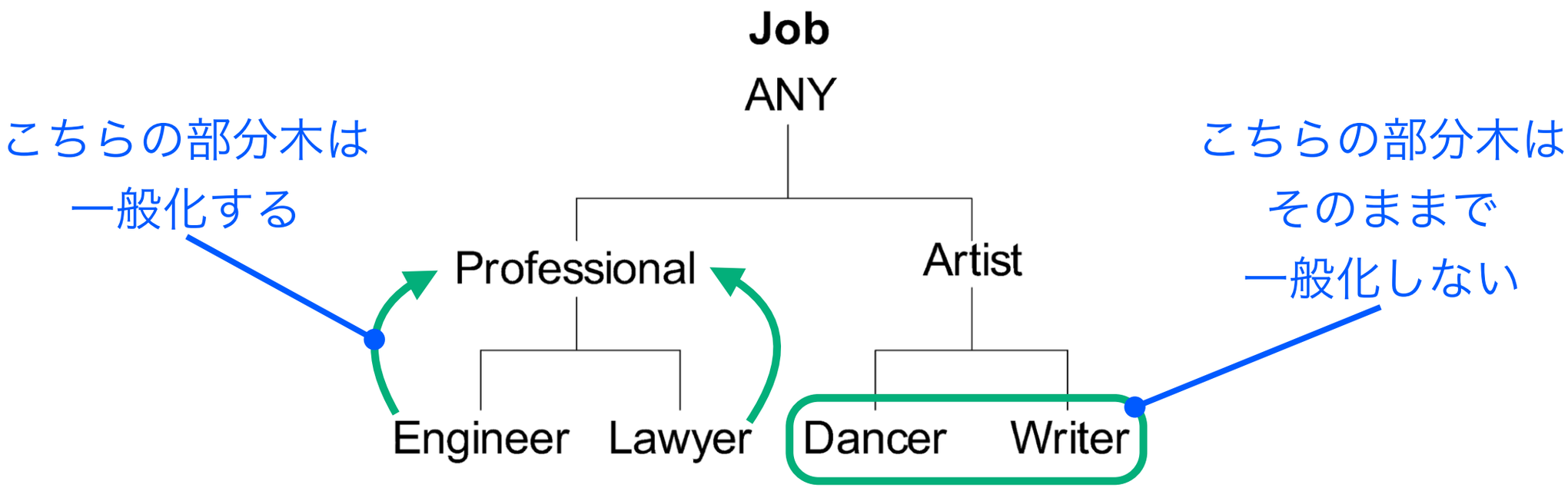
[Bayardo+ 2005; Fung+ 2005,2007; Iyengar 2002; LeFevre+ 2005]

部分木一般化 (subtree generalization)

分類木の非終端ノードの値は，全ての子ノードの値を一般化するか，全く一般化しないかのいずれか

- 直感的には，分類木のある部分木の下を切り取るイメージ

例：Engineer を Professional に一般化するのなら同じ部分木の Lawyer も Professional にする必要があるが，Artist の部分木の Dancer や Writer は一般化する必要はない



子孫一般化

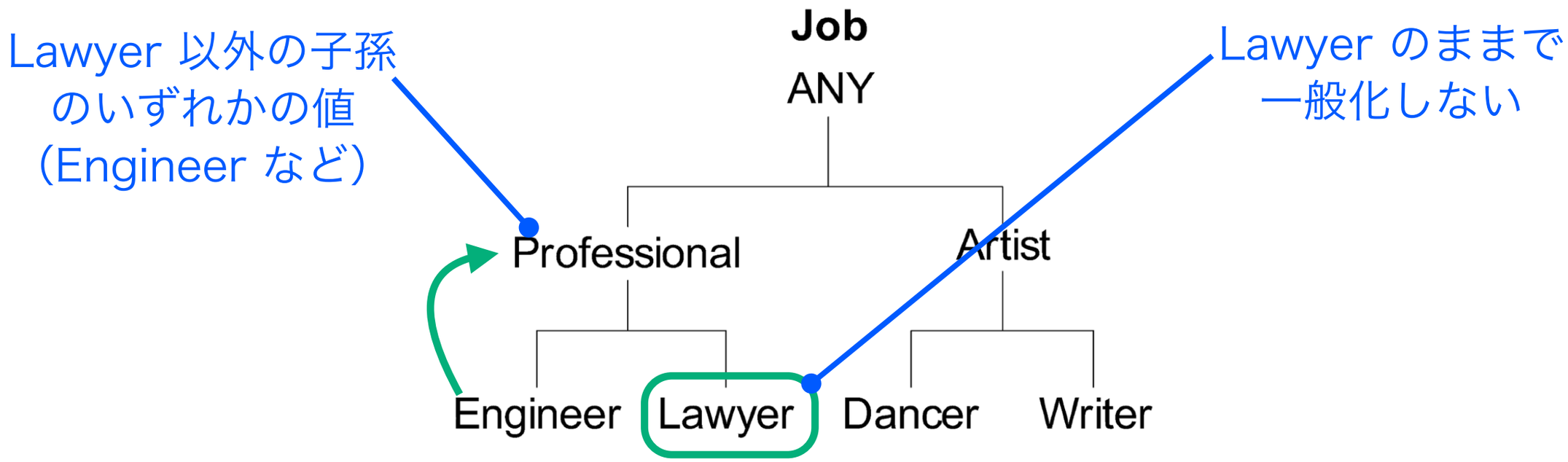
[LeFevre+ 2005]

子孫一般化 (sibling generalization)

部分木一般化と似ているが，一般化しない子孫があり，親ノードの値を欠損している子ノードの値として利用

- プライバシ要件を満たさない子ノードのみを置換するので，歪みは小さい

例：Engineer を Professional に一般化しても，Lawyer を Professional にする必要はない．値 Professional は部分木中の Lawyer 以外のいずれかの値と解釈する．



セル一般化

[LeFevre+ 2005; Wong+ 2006; Xu+ 2006]

セル一般化 (cell generalization)

- **大域再符号化 (global recoding)**：全体，部分木，子孫一般化は，データ表中の同じ具現値は同じ値に一般化する
- **局所再符号化 (local recoding)**：セル一般化では，データ表中の同じ具現値でも一般化するものと，しないものの混在を許す
  - 柔軟な再符号化なので，歪みを小さくできる
  - データ探索問題：標準的なマイニング手法では，実際は同じでもEngineer とProfessional を異なる値として扱ってしまう

値 Professional に一般化

値 Engineer のまま変更しない

病院の元データ

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	30	Flu
Writer	Female	30	HIV
Dancer	Female	30	HIV
Dancer	Female	30	HIV

多次元一般化

[LeFevre+ 2006a, 2006b]

多次元一般化 (multidimensional generalization)

個々の値を一般化するのではなく，qid グループをまとめて一般化

- *qid* グループが異なると，同じ値を一般化しても違う値なることも
- 問題のある *qid* グループだけを一般化できるので，全体や部分木一般化より歪みは小さい
- 局所再符号化ではないが，データ探索問題を生じる
- クラスタリングによるクラスタごとに一般化する方法も [Nergiz+ 2007]

例：〈Engineer, Male〉→〈Engineer, ANY_Sex〉と
〈Engineer, Female〉→〈Engineer, Female〉のような一般化が可能

多次元一般化

[LeFevre+ 2006a, 2006b]

1次元一般化 全体・部分木一般化など

$$f_i : D_{A_i} \to D', \text{ for } A_i \in QID$$

多次元一般化 各属性ごとに値を一般化

$$f : D_{A_1} \times D_{A_2} \times \cdots \times D_{A_n} \to D'$$

$qid = \langle v_1, v_2, \ldots, v_n \rangle$ をまとめて $qid' = \langle u_1, u_2, \ldots, u_n \rangle$ と一般化

※ D_{A_i} は属性 A_i の定義域

抑制

抑制 (suppression)

非公開であることを示す特殊な値と置換する

(逆の演算は **開示 (disclosure)**)

抑制の種類

- **レコード抑制 (record suppression)**

  レコード全体をまとめて抑制する

  [Bayardo+ 2005; Iyengar 2002; LeFevre+ 2005; Samarati 2001]

- **値抑制 (value suppression)**

  データ表中のある値を，全ての具現値について抑制

  [Wang+ 2005, 2007]

- **セル抑制 (cell suppression)** / **局所抑制 (local suppression)**

  データ表中のある値を，一部の具現値について抑制

  [Cox 1980; Meyerson+ 2004]

匿名化操作の選択

匿名化操作は，探索空間とデータの歪みのトレードオフを考慮

全体符号化	⇔	**局所再符号化**
探索空間小・歪み大		探索空間大・歪み小

- **極小匿名 (minimally anonymous)**：あるプライバシ要求を満たしているが，一連の匿名化操作をそれ以上減らすと満たさなくなる状態
  - 一般的には効率面からみて実用的
- **最適匿名 (optimally anonymous)**：あるプライバシ要求を満たし，かつある尺度で元の表の情報を最も保存している状態
  - いろいろなプライバシ要求と一般化・抑制の組み合わせについて最適匿名の探索がNP困難であることが示されている

分解

[Xiao+ 2006a]

分解 (anatomization)

擬識別子とセンシティブ属性を，擬識別子表とセンシテイブ表の異なるデータ表に分解してその関連性を消す

- **擬識別子表** (QIT; quasi-identifier table) と**センシティブ表** (ST; sensitive table)
- **グループID**：グループごとに両方の表で同じ値をとる共通属性
  - グループ内のセンシティブ属性値が l 種類で，一度ずつ出現するなら，擬識別子とセンシティブ属性が結びつく確率は $1/l$

分解の例

センシティブ属性 Disease, QID＝{Age, Sex}

元データ

Age	Sex	Disease (sensitive)
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	Flu
38	Female	Flu
38	Female	Heart
38	Female	Heart

2多様な QID グループ

Age	Sex	Disease (sensitive)
[30–35)	Male	Hepatitis
[30–35)	Male	Hepatitis
[30–35)	Male	HIV
[30–35)	Male	Hepatitis
[30–35)	Male	HIV
[30–35)	Male	HIV
[35–40)	Female	Flu
[35–40)	Female	Flu
[35–40)	Female	Heart
[35–40)	Female	Heart

グループ生成時には値を一般化

擬識別子表

Age	Sex	GroupID
30	Male	1
30	Male	1
30	Male	1
32	Male	1
32	Male	1
32	Male	1
36	Female	2
38	Female	2
38	Female	2
38	Female	2

分割後の表では元の値

グループID

センシティブ表

GroupID	Disease (sensitive)	Count
1	Hepatitis	3
1	HIV	3
2	Flu	2
2	Heart	2

レコード数

分解の長所・短所

- 擬識別子表もセンシティブ表もデータを修正する必要がない
  - QIDやセンシティブ属性の値を含む集約演算で，値の分布が分からなくなる一般化よりも正しい解を得られる
- 分割した表に対応した専用アルゴリズムが必要
- 連続値の公開には向かない

例：38歳の心臓病は何人？

元データ

Age	Sex	Disease (sensitive)
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	Flu
38	Female	Flu
38	Female	Heart
38	Female	Heart

正解=2

分解

Age	Sex	GroupID
30	Male	1
30	Male	1
30	Male	1
32	Male	1
32	Male	1
32	Male	1
36	Female	2
38	Female	2
38	Female	2
38	Female	2

GroupID	Disease (sensitive)	Count
1	Hepatitis	3
1	HIV	3
2	Flu	2
2	Heart	2

同じグループID

3人×(心臓病の2人/同一IDの4人)=1.5人

一般化

Age	Sex	Disease (sensitive)
[30–35)	Male	Hepatitis
[30–35)	Male	Hepatitis
[30–35)	Male	HIV
[30–35)	Male	Hepatitis
[30–35)	Male	HIV
[30–35)	Male	HIV
[35–40)	Female	Flu
[35–40)	Female	Flu
[35–40)	Female	Heart
[35–40)	Female	Heart

この区間は35〜39なので
38歳の可能性は1/5

2人×1/5=0.4人

順列置換

[Zhang+ 2007]

順列置換 (permutation)

レコードをグループに分けて，グループ内でセンシティブ属性値を入れ替えることで，擬識別子と数値センシティブ属性の関係を断ち切る

摂動

[Adam+ 1989]

摂動 (perturbation)

元の値から計算した値とあまり違わないような
統計量が得られる人工の値でデータを置き換える

- 統計的開示制御の分野で長く研究されてきた
- 元データにはあった実世界のレコード所有者との対応はなくなり（レコード真実性はなくなる）公開レコードは合成されたものとなる
  - データ表を人間が見ることに意義はなく，選択統計量のみを公開しているようなもの
  - 汎化や抑制は真実性を保証するが，統計量は摂動より不正確になる

加法的ノイズ

加法的ノイズ (additive noise)

元の値 s を，ある分布に従う乱数 r を加えた値 $s + r$ と置き換える

- 統計的開示制御で，数値変数を保護するためによく使われる
- プライバシは公開値から元の値がどれだけ推定されにくいかで測る
- 平均や相関などの単純な統計量に加え，より複雑なデータマイニングの結果もは保存できる
- 値の相関が大きく，ノイズが小さいときには元の値を復元できる

  [Kargupta+ 2003]，これに対する対抗策 [Huang+ 2005]

データ交換

データ交換 (data swapping)

低次の頻度や周辺頻度は保存されるように
センシティブ属性の値を，レコード間で交換

- 数値属性も[Reiss et al. 1982]カテゴリ属性 [Reiss 1984] も保存できる

[Domingo-Ferrer+ 2002]

順位交換 (rank swapping)

ある属性の値で昇順で整列し，
その系列中で p% 以内の近辺にある値と交換する

- 通常の交換よりよく統計量を保存する

人工データ生成

人工データ生成 (synthetic data generation)

統計量を保つように生成した人工の値と置換する

- **サンプリング**：統計モデルを生成し，そのモデルからサンプリングした値を使う
- **凝縮 (condensation)**：レコードを複数のグループにまとめ，グループごとに総和や共分散などの値を求め，これらのグループの統計量を保存するように人工データを生成 [Aggarwal & Yu 2008a, 2008b]

4. 情報計量

Information Metrics

情報計量

情報計量 (information metric)

プライバシの保護の度合いと共に重要なデータの有用性を測る計量

使用時期による分類

- **データ計量 (data metric)**：元のデータ表に対する，匿名化した表データ全体の品質を測るために用いる
- **探索計量 (search metric)**：匿名化の探索アルゴリズムの各ステップで，情報の保存の最大化や歪みの最小化をした匿名化を実現するために用いる

使用目的による分類

- 一般用途 (general purpose)，特殊用途 (special purpose)，そしてトレードオフ用途 (trade-off purpose)

一般用途計量

一般用途計量 (general purpose metric)

利用目的で適切な情報計量は異なるが,
データ公開者が公開データの用途を知らない場合に用いる

最小歪み原理 (principle of minimal distortion; MD原理)

利用目的が不明の場合に
元データと匿名化データの間の『類似性』で有用性を評価する

- 一般化や抑制には罰則を与える

  例：10個の事例を1階層一般化すると罰則10, さらに上位階層に一般化すると罰則10を加える [Samarati 2001]

ILoss

[Xiao+ 2006b]

ILoss

ある値を一般的な値 v_g に変えたときの情報損失を次式で測る

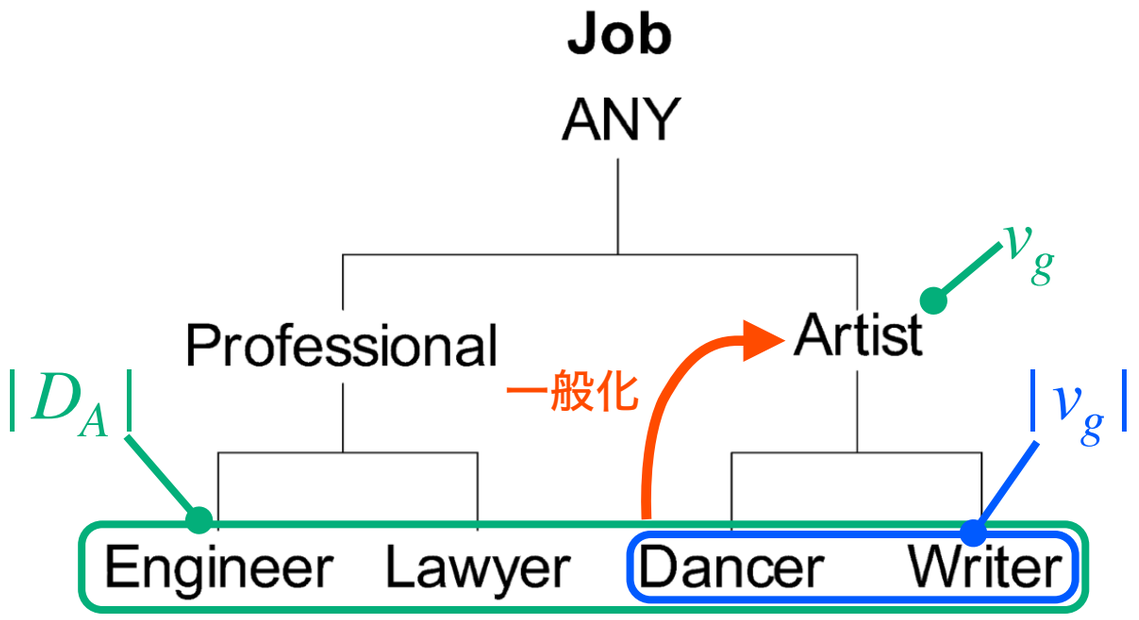
$$ILoss = (|v_g| - 1)/|D_A|$$

$|v_g|$ は値 v_g の子孫である値の，$|D_A|$ は v_g の属性 A の終端値の種類数

- 元データ中の値が全て終端値である場合にしか使えない
- v_g が元データと同じ値であるなら，$ILoss(v_g)=0$ となる

例：Dancer を Artist に一般化

$$ILoss(\text{Artist}) = \frac{|v_g| - 1}{|D_A|} = \frac{2-1}{4} = 0.25$$



ILoss

[Xiao+ 2006b]

レコード r での損失

$$ILoss(r) = \sum_{v_g \in r} w_i ILoss(v_g)$$

ただし，w_{i} は，属性 A_{i} の重み

データ表 T 全体での損失

$$ILoss(r) = \sum_{r \in T} ILoss(r)$$

discernibility計量

[Skowron+ 1992]

discernibility計量 (DM)

各レコードごとに QID で区別不可能になると罰則を与える
レコードが大きさ s のグループに属するようになるとき罰則は s

99件のレコードは
他の99件と
区別不可能

それぞれ
50件が他の50件と
区別不可能

Engineer と Lawyer を
Professional に一般化

Engineer: 99
Lawyer: 1

Professional: 100

Engineer: 50
Lawyer: 50

Professional: 100

100件が他の100件と区別不可能

新たに識別不能になるものが多い

distinctive属性

[Sweeney 1998]

distinctive属性 (DA)

全体一般化のとき，匿名化最小にするための探索計量で，
データ中で異なる値の数が最も多い属性を一般化する対象に選ぶ

- このヒューリスティックは探索には使えるが，匿名化表の評価には使えない

専門用途計量

専門用途計量 (special purpose metric)

公開データをどう分析するか分かっているときに
その分析に必要な情報を保存するようにして公開する

- 分析結果自体を公開することはしないのは，データ公開者に分析能力がなかったり，分析アルゴリズムを変更できるようにするため

例：肺癌データの分類

- 年齢は肺癌かどうかの分類に有用だが，誕生日は役にたたない
  - ➡ 誕生日を一般化して公開しても分類精度に影響しない

⬇

- データの歪みを一般の歪みではなく，分類誤差で測る
- テスト集合ではなく，訓練集合の誤差を使って測っている
- 汎化や抑制によって，ノイズがなくなり分類精度が向上する場合も

分類計量

[Iyengar 2002]

分類計量 (classification measure; CM)

公開データ自体を訓練データとし，訓練データの誤差で歪みを評価

- グループ内で少数派クラスのレコードを一般化・抑制する
  ⬅ もともと少数派で誤分類されやすいので，影響が小さくてすむ

- CMはデータ計量なので，歪み自体に罰則を与える
  ➡ 汎化によってノイズが分類に影響する情報になったりして困る
  ➡ 分類への影響の少ない匿名化操作を順位付けするような探索計量が望ましい
- 訓練データでの誤差を評価しているので，匿名化したデータから実際に分類してみないと汎化誤差は評価できない

トレードオフ計量

トレードオフ計量 (trade-off metric)

公開したデータの有用性とプライバシの保護の度合いの
トレードオフ関係を考慮した計量

- 有用性とプライバシ保護の最適トレードオフを達成するための計量

情報/プライバシのトレードオフ原理

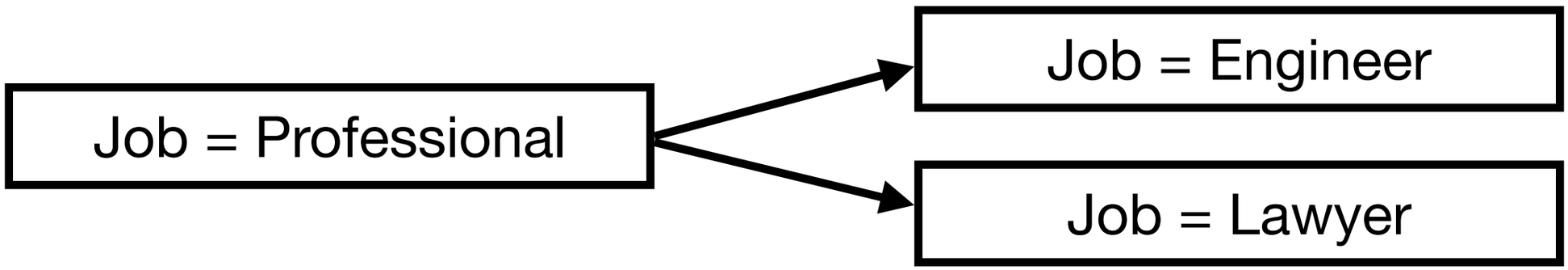
[Fung+ 2005,2007]

情報/プライバシのトレードオフ原理
(principle of information/privacy trade-off)

トレードオフを考慮して探索尺度を設計するため規準

例：Job の値が Professional であるグループは，特殊化操作 s により，Engineer と Lawyer の値をとるグループに分かれる

特殊化操作 s



特殊化操作 s により情報利得 $IG(s)$ を得て，プライバシ $PL(s)$ を失う

情報/プライバシのトレードオフ原理

[Fung+ 2005,2007]

情報/プライバシのトレードオフを次式で評価し
これを最大にする特殊化操作 s を選ぶ

$$IGPL(s) = \frac{IG(s)}{PL(s)+1}$$

$IG(s)$ の例

分類問題のとき，決定木ID3で用いる情報量の利得やMD歪みの減少量

$PL(s)$ の例

特殊化した属性 s を含むQIDグループの匿名度の減少量の平均

$$PL(s) = \mathrm{avg}_j\{A(QID_j) - A_s(QID_j)\}$$

ただし，$A(QID_j)$ と $A_s(QID_j)$ はレコード j の特殊化前後の匿名度

※ 特殊化する代わりに，一般化して情報損失とプライバシ利得を考えた $ILPG$ 尺度もある

5. 匿名化アルゴリズム

Anonymization Algorithms

匿名化アルゴリズムの特徴

レコードリンク

アルゴリズム	操作	計量	最適性
二分探索 [Samarati 2001]	FG, RS	MD	最適
MinGen [Sweeney 2002b]	FG, RS	MD	最適
Incognito [LeFevre+ 2005]	FG, RS	MD	最適
K-Optimize [Bayardo+ 2005]	SG, RS	DM, CM	最適
μ-argus [Hundepool+ 1996]	SG, CS	MD	最小
Datafly [Sweeney 1998]	FG, RS	DA	最小
遺伝アルゴリズム [Iyengar 2002]	SG, RS	CM	最小
ボトムアップ一般化 [Wang+ 2004]	SG	ILPG	最小
トップダウン特殊化 [Fung+ 2005,2007]	SG, VS	IGPL	最小
クラスタ分析用トップダウン特殊化 [Fung+ 2009]	SG, VS	IGPL	最小
モンドリアン多次元 [LeFevre+ 2006a]	MG	DM	最小
ボトムアップ・トップダウン貪欲法 [Xu+ 2006]	CG	DM	最小
TDS2P [Wang+ 2005; Mohammed+ 2009]	SG	IGPL	最小
濃密化 [Aggarwal+ 2008a, 2008b]	CD	heuristics	最小
r-Gatherクラスタリング [Aggarwal+ 2006]	CL	heuristics	最小

※ 一般化（FG=全体一般化，SG=部分木一般化, CG=セル一般化，MG=多次元一般化）抑制（RS=レコード抑制，VS=値抑制，CS=セル抑制）AM=分解，PM=順列置換，AN=加法的ノイズ，SP=サンプリング，CD=凝縮，CL＝クラスタリング

匿名化アルゴリズムの特徴

属性リンク

アルゴリズム	操作	計量	最適性
トップダウン開示 [Wang+ 2005,2007]	VS	IGPL	最小
Progressive局所再符号化 [Wong+ 2006]	CG	MD	最小
*l*多様性Incognito [Machanavajjhala+ 2007]	FG, RS	MD, DM	最適
InfoGainモンドリアン [LeFevre+ 2006b]	MG	IG	最小
Anatomy [Xiao+ 2006a]	AM	heuristics	最小
(k, e)匿名順列化 [Zhang+ 2007]	PM	最小誤り	最適
貪欲個人化 [Xiao+ 2006b]	SG, CG	lLoss	最小
*t*近接性Incognito [Li+ 2007]	FG, RS	DM	最適

※ 一般化（FG=全体一般化， SG=部分木一般化, CG=セル一般化， MG=多次元一般化）抑制（RS=レコード抑制， VS=値抑制，CS=セル抑制）AM=分解， PM=順列置換， AN=加法的ノイズ， SP=サンプリング， CD=凝縮， CL＝クラスタリング

匿名化アルゴリズムの特徴

データ表リンク

アルゴリズム	操作	計量	最適性
SPALM [Nergiz+ 2007]	FG	DM	最適
MPALM [Nergiz+ 2007]	MG	heuristics	最小

確率的攻撃

アルゴリズム	操作	計量	最適性
Cross-Training Round Sanitization [Chawla+ 2005]	AN	統計的	N/A
ε差分プライバシ加法的ノイズ [Dwork 2006]	AN	統計的	N/A
$\alpha\beta$アルゴリズム [Rastogi+ 2007]	AN, SP	統計的	N/A

※ 一般化（FG=全体一般化，SG=部分木一般化, CG=セル一般化，MG=多次元一般化）抑制（RS=レコード抑制，VS=値抑制，CS=セル抑制）AM=分解，PM=順列置換，AN=加法的ノイズ，SP=サンプリング，CD=凝縮，CL＝クラスタリング

レコードリンクモデル用最適匿名化アルゴリズム

最適匿名化アルゴリズム (optimal anonymization algorithm)

- 全体一般化とレコード抑制によって，あるデータ尺度での最適な k 匿名化を実現する
- 全体一般化の探索空間は小さく実行可能だが，あまり大規模化はできない

MinGen [Sweeney 2002b]

- 全体最適化を網羅探索し，最小歪みの最適化をする
- 大きなデータには向かない

二分探索 [Samarati 2001]

- 全ての最小汎化を特定した後，二分探索で最小歪みのものを探索
- 全ての最小汎化の列挙のコストのため大規模化は無理

レコードリンクモデル用最適匿名化アルゴリズム

Incognito [LeFevre+ 2005]

- ボトムアップ型の一般化アルゴリズムで，値域一般化の k 匿名化を全探索する

観測5.1 (rollup特性：qid グループの大きさを計算するのに利用)

qid が $\{qid_1, \ldots, qid_c\}$ の一般化であるなら $|qid| = \sum_i^c |qid_i|$

観測5.2 (一般化特性：探索の枝刈りに利用)

QID 中の全属性についてデータ表 T より特殊ではないデータ表を T' とする．QID 上で T が k 匿名であるなら，T' も QID 上で k 匿名である

- データ表の qid について，もし qid' が qid の一般化であり，かつ $|qid| \geqq k$ であるなら，$|qid'| \geqq k$ を満たす

  ➡ T が k 匿名なら，それ以上一般化する必要はない

- このため MinGen や Samarati の方法より性能はよいが，QID の大きさに対してどのアルゴリズムも指数的に複雑度は増加

レコードリンクモデル用最適匿名化アルゴリズム

K-Optimize [Bayardo and Agrawal 2005]

- ノードはある匿名化操作をした状態に相当する 木構造を使った効率的な枝刈り
- 属性にある全順序を仮定し，最も一般的な表から，識別尺度 DM と分類尺度 CM に基づいて，その子孫が大域最適にはなりえないものを枝刈りする
- 他のアルゴリズムと異なり，部分木一般化とレコード抑制を採用

レコードリンクモデル用
極小匿名化アルゴリズム

極小匿名化アルゴリズム (minimal anonymization algorithm)

- 欲張りアルゴリズムで効率的に，極小の匿名化を達成するデータ表を探索するアルゴリズム
- 極小の解しか発見できないが，最適匿名化アルゴリズムより大規模化が可能

μ-argus [Hundepool+ 1996]

- 3属性の値の組合せの頻度を計算し，部分木一般化とセル抑制を欲張り的に適用して探索
- 3属性より多い属性に対しては k 匿名性を満たさない可能性

レコードリンクモデル用
極小匿名化アルゴリズム

Datafly [Sweeney 1998]

- 実用的規模のk匿名化が可能な最初のアルゴリズム
- Datafly は全体一般化で，レコード抑制を用いる
- qidグループ数の大きさの配列を確保し，異なる属性値数が最大になる属性を選ぶという DA尺度ヒューリスティック探索に基づいて k 回未満の組合せを欲張り法で一般化する

Genetic [Iyengar 2002]

- CM尺度と部分木一般化を用いて分類情報を保存するように k 匿名化を達成する
- 一般化の状態を遺伝子として符号化し，遺伝的アルゴリズムを利用
- 汎用の手法で匿名化した場合より分類に関しては予測性能はよい
- 遺伝アルゴリズムは大規模データに対しては非効率的

レコードリンクモデル用極小匿名化アルゴリズム

ボトムアップ一般化 [Wang+ 2004]

- 情報損失とプライバシ利得のトレードオフを表す$ILPG$尺度を利用
- ボトムアップに一般化して探索する，分類用の効率的な極小 k 匿名化手法
- qid グループの大きさはrollup特性により一般化で増加するが，その増加が極小である critical generalization を適用候補にする．もし，それがない場合には他の一般化を使う

レコードリンクモデル用
極小匿名化アルゴリズム

トップダウン特殊化 [Fung+ 2005, 2007]

- 情報利得とプライバシ損失のトレードオフを表す$ILPG$尺度を利用
- トップダウンに，観測5.2の一般化特性を利用しつつ特殊化し，k 匿名化を満たさなくなる直前で停止
- カテゴリ・数値属性の両方に対応

クラスタリング用の k 匿名化 [Fung+ 2008,2009]

- クラスラベルがないので，最初にクラスタリングしておき，その分類結果を保存するようにトップダウン特殊化を適用

レコードリンクモデル用極小匿名化アルゴリズム

ボトムアップに対するトップダウンの利点

- 途中でアルゴリズムを停止してもk匿名性が達成されている
- 単一高次元の QID では歪みが大きくなりがちだが，複数 QID を扱える
- トップダウンだと，ある qid グループがそれ以上特殊化できないと分かれば，元データを捨てられるが，ボトムアップだとアルゴリズム終了時まで全データを保持する必要
- トップダウンでは QID 選択のジレンマ（2.1節）が生じる

レコードリンクモデル用極小匿名化アルゴリズム

モンドリアン多次元法 [LeFevre+ 2006a]

- トップダウンの特殊化で，多次元一般化を欲張り法で探索
- モンドリアン多次元法では，特殊化したものがそれぞれ k レコード以上になっている qid グループのみを特殊化する
  - なお，通常ののトップダウン特殊化では，ある値 v を含む全ての qid グループを同時に特殊化，すなわち，k レコード以上のグループだけを特殊化
- この緩和により歪みは小さくなるが，探索空間は広くなる
  - 文献 [Xu+ 2006] はさらにセル一般化も導入

レコードリンクモデル用
摂動アルゴリズム

摂動アルゴリズム (perturbation algorithm)

- 統計情報を保存しつつ，被攻撃者とレコードの関連リンクを摂動によって切断するアルゴリズム

凝縮 (condensation) [Aggarwal+ 2008a,2008b]

- レコードを，k 以上の大きさの重複のないグループに割当て，各グループ内での，レコード中の属性の平均や相関などの統計量を保存するデータを生成
- 分類木が不要
- データストリームにも，新規データを最も近いグループに分類して，統計量を再計算すれば対応可能．その際，大きすぎるグループは適宜分割する．

レコードリンクモデル用
摂動アルゴリズム

- ***r*-gatherクラスタリング** [Aggarwal+ 2006]
  - データ点を大きさが r 以上のクラスタに分類し，クラスタをその中心，大きさ，半径，センシティブな値と共に公開
  - はずれ値の影響を減らすため，割合 ε のデータをはずれ値として公開データからは削除する拡張も

レコードリンクモデル用摂動アルゴリズム

- 中心が q で半径 cδp の球中に t 未満のデータしかない q(c, t) 分離性（2.1節）を達成する匿名化法を2種類提案 [Chawla+ 2005]
  - **recursive histogram sanitization**：局所密度が 2t を超えなくなるまで再帰的に元データを部分矩形領域に分割し，それらの領域の境界と領域中のデータ数を公開．高次元に用の拡張も提案
  - **density-based perturbation**：[Agrawal+ 2000] の拡張で，摂動するデータ点の近くの密度に応じた加法的ノイズを付加
- **cross-training round sanitization**：これらはデータ点自体は保護できるが，t 近隣の密度は球の半径から漏洩

  ➡ 両者を組み合わせた拡張：データを無作為に二つのグループに分割し，一方は recursive histograom sanitization を適用し，このヒストグラムに応じた，ガウスノイズをもう一方のグループのデータに加える

属性リンクモデル用アルゴリズム

l 多様性プライバシモデルを用いて属性リンク攻撃を防ぐ手法

l多様性Incognito [Machanavajjhala 2006,2007]

- レコードリンク用のボトムアップIncognitoを l多様性用に修正
- l多様性は一般化に関して非減少であるという一般化特性に基づく
- 全体一般化と部分木一般化を採用

InfoGain Mondrian [LeFevre+ 2006b]

- 分類やクエリ応答に特化したk多様性やエントロピーl多様性を満たす欲張りアルゴリズム
- 多次元一般化を採用

属性リンクモデル用アルゴリズム

トップダウン開示 (Top-Down Disclosure) [Wang+ 2005,2007]

- プライバシテンプレート〈$QID \rightarrow s, h$〉：QIDグループについて s に関する推論の確信度は h を超えない
- 全ての属性値が抑制された状態から初めて，各反復で$IGPL$を最大にする属性値を公開し，プライバシテンプレート集合のいずれかに違反するようになったら停止
- 次に観測に基づき，極小抑制データ表が獲得できる

  観測 5.3（公開特性）：プライバシテンプレート〈$QID \rightarrow s, h$〉で，あるデータ表がこれに違反するとき，その表で抑制された値を公開して得られるデータ表はどれも違反することになる.
- 全体，部分木，子孫一般化にも拡張可能で，セル一般化に拡張し(α, k)匿名化を満たす拡張も [Wang+ 2006]

属性リンクモデル用アルゴリズム

(k, e)匿名化順列 ((k, e)-anonymity permutation) [Zhang+ 2007]

- 誤差 E の総和を最小化するようにレコードをグループ化する最適順列手法
  - 誤差 E には各グループ内のセンシティブ属性の範囲などを利用
- 時間・空間複雑度は共に $\mathrm{O}(n^2)$
- 工程管理や公的統計での range coding (coarse coding) と関連
- 数値を範囲に分け，範囲の境界を保持するような一般化
- グループ境界も集約クエリの効率的計算に活用

属性リンクモデル用アルゴリズム

個人化プライバシ [Xiao+ 2006b]

- データ所有者ごとに異なるプライバシ要求を満たす欲張り法：
  - 最初に全 QID 属性を最も一般的にし，センシティブ属性は一般化しない
  - 各反復で，トップダウンに，各 qid グループごとに QID 属性を特殊化し，センシティブ属性にセル一般化を適用して個人の要求を満たすようにする
  - センシティブ属性の一般化に関して漏洩確率は非増加で，かつ最も一般的な状態にまでセンシティブ属性は一般化できるので，プライバシ要求を満たすことは可能になる
  - $ILoss$ による情報の損失が少ないものを欲張り的に採用し，$ILoss$ が改善できなくなったら停止

データ表リンクモデル用アルゴリズム

データ表リンク：公開データ表にある被攻撃者が含まれるかを識別

存在性アルゴリズム (presence algorithm) [Nergiz+ 2007]

- δ存在性：外部データ表 E に関して，　一般化したデータ表 T' が $\delta_{\min} \leq P(t \in T|T') \leq \delta_{\max}, \forall t \in E$ を満たす
  - **SPALM (Single-Dimensional Presence Algorithm)**
    - 全定義域の1次元一般化を採用
    - 全定義域一般化に対するδ存在性の反単調性，すなわち，データ表 T が δ 存在なら，その一般化である T' も δ 存在であることを利用した枝刈りを実行.
  - **MPALM (Multi-Dimensional Presence Algorithm)**
    - 多次元一般化を採用した極小匿名化アルゴリズム
    - 柔軟な一般化を使っているのでSPALMより情報の損失は小さい
    - 計算量は $O(|C| \, |E| \log_2 |E|)$（Cはデータ表の個人属性）

匿名データの最低限性攻撃

[Wong+ 2007]

最低限性攻撃 (minimality attack)

匿名化アルゴリズムの特性を利用した攻撃

- 想定している目標被攻撃者の QID やプライバシ要求は状況から攻撃者は推定でき，そこから匿名化操作やアルゴリズムも推定できる

最低限性原理 (minimality principle)：

- ボトムアップに汎化したときは，k 匿名を一度達成した時点でそれ以上は汎化しない
- 最低限性攻撃はこの原理を利用し，不可能な状況を避けつつ匿名化操作を逆に適用するもの

匿名データの最低限性攻撃

[Wong+ 2007]

例：確信度の限界〈{Job, Sex}→HIV, 60%〉は攻撃者に既知

元の非公開の表

Job	Sex	Disease
Engineer	Male	HIV
Engineer	Male	HIV
Lawyer	Male	Flu
Lawyer	Male	Flu
Lawyer	Male	Flu
Lawyer	Male	Flu
Lawyer	Male	Flu

匿名化して公開した表

Job	Sex	Disease
Professional	Male	HIV
Professional	Male	Flu
Professional	Male	Flu
Professional	Male	Flu
Professional	Male	Flu
Professional	Male	Flu
Professional	Male	HIV

外部の表

Name	Job	Sex
Andy	Engineer	Male
Calvin	Lawyer	Male
Bob	Engineer	Male
Doug	Lawyer	Male
Eddy	Lawyer	Male
Fred	Lawyer	Male
Gabriel	Lawyer	Male

- 外部の表から〈Lawyer, Male〉が5人，〈Engineer, Male〉は2人
  ➡ 一般化して公開されているので，元の表は確信度限界を満たす
- HIVが2人とも〈Lawyer, Male〉と仮定 ➡ 確信度 2/5≦60% ➡ 否定
- 部分木一般化が適用されている
  ➡ 確信度限界を満たすのは〈Engineer, Male〉の2人がHIVの場合のみ

⬇

Bob と Andy がHIVであることが漏洩

m-confidentiality

[Wang+ 2007]

m-confidentiality

レコード所有者からのセンシティブ属性値と繋がる確率を制限して，最低限性攻撃を回避

l多様性に対する最低限性攻撃の回避

1. 最初に k 匿名化
2. 匿名化後の qid グループで l 多様性を満たさないものを，センシティブ属性値が l 多様性を満たすように

- **適用可能な最低限性攻撃**：一般化，抑制，分解，順列置換操作をする最小限匿名化と最適匿名化の両方を用いるもの
- **対象プライバシモデル**：l 多様性，(α, k) 匿名性，(k, e) 匿名性，分解，t 近接性，m 不変性，(X, Y) プライバシなどのモデル

確率的攻撃モデル用のアルゴリズム

確率的攻撃

摂動で防御 ➡ 最低限性攻撃は利用できない

非決定的変換を用いている ➡ 元データに逆変換できない

摂動の分類

大域的な摂動：信頼できるデータ公開者がまとめて摂動

- このサーベイの主な方法

局所的な摂動：データ所有者は自身のみを信頼し，摂動してからデータ公開者に送信 [Agrawal+ 2005]

- 大域的な摂動にも利用できるが非効率的 [Rastogi+ 2007，Dwork 2007]

クラス情報を保存する摂動

クラス情報を保存する摂動

- 分類器を学習するために，クラス情報を保存するようにデータに摂動を加える
- 摂動を加えたデータを集約することで，元データの分布を再構成
- 通常の分類器は利用できず，専用の分類アルゴリズムが必要

決定木 [Agrawal+ 2000]，単純ベイズ [Zhang+ 2005]

クエリ結果の集約

- クエリに応じて公開した統計量を集約して，センシティブ情報を得る行為を抑止
- 統計分野で，統計情報の開示抑制や，クエリ結果の集約として研究
  [Cox 1980, Chawla+ 2005, Duncan+ 1998, Matloff 1988, Ozsoyoglu+ 1990]

k%-dominance rule [Cox 1980]：2個か3個かの要素が，総和の統計量の中で k% 以上を占めるようであれば，その公開を抑制

- その他，クエリの数や重複の制限，集約，データの集約，データの交換などの方法
  - 実装が難しい [Farkas and Jajodia 2003]，あまり生じない問題を扱う
- この分野のサーベイ文献：[Adam+ 1989; Domingo-Ferrer 2001; Moore 1996; Zayatz 2007]

ε差分加法的ノイズ

[Dwork 2008]

差分プライバシ [Dwork 2006]

- の標準偏差が $\sigma \geqq \varepsilon / \Delta f$ の指数分布に従う外乱を各軸に加える
- ただし，Δf は，レコード一つの追加・削除で生じるクエリ f への応答の差の最大値

確率的差分プライバシ [Machanavajjhala+ 2008]

- 元データからモデルを構築し，そのモデルから得たデータで，元データを置換
- 代表的ではないデータの除去や，定義域の縮小を行う

$\alpha\beta$アルゴリズム

[Rastogi+ 2007]

$\alpha\,\beta$アルゴリズム

- (d, γ) プライバシ用のアルゴリズム
- (d, γ) プライバシ：被攻撃者のレコードが元データ表 D にある確率の，公開データ表 T の観測の前後での差を抑える
- 元データ表のレコードの部分集合を確率 $\alpha + \beta$ で選び，それを公開データ表に挿入
- 全属性の定義域上から偽のレコードを生成し，それが元データ表 D 中になければ，確率 β でデータ表 T に挿入

※ 公開データは元データと偽データの両方を含む
※ レコードの真実性を満たさない

以降の章は省略

- **6. 公開状況の拡張**
  Extended Scenarios
- **7. 他形式のデータの匿名化**
  Anonymizing Other Types of Data
- **8. 他分野でのプライバシ保護技術**
  Privacy-preserving Techniques in Other Domains
- **9. まとめと今後の研究方向**
  Summary and Future Research Directions