系列データの匿名化に関する研究

髙橋　翼

システム情報工学研究科
筑波大学

２０１４年　３月

概要

近年，パーソナルデータが大量に蓄積され，第三者提供や二次活用による活用が期待されている．特に，パーソナルデータのシーケンスである系列データからは，系列中のパターンや因果関係といった高度な分析を実現できる．このような系列データをデータ保有者以外も利活用可能になれば，様々な分野で新たな発見やサービスの発展が期待できる．しかし，パーソナルデータの第三者提供や二次活用に際しては，プライバシへの配慮が必要となる．

プライバシ保護技術の一つとしてデータ匿名化が知られている．データ匿名化は，k-匿名性 [67] や ℓ-多様性 [47] 等の匿名性の指標を充足するようにデータの加工を行う技術である．系列データに対する既存の匿名化技術は，蓄積された複数の属性値の系列に対して静的に匿名化を行う技術であり，連続的に蓄積されていくデータを逐次匿名化することができなかった．また，個人が特定されずともある属性の値が他の属性群から特定されてしまう場合がある．このようなときに，系列データの属性間の関係を多様化するような加工には，属性値や属性間の関係を大きく曖昧化してしまう問題があった．

本研究では，系列データの連続的な利活用をプライバシを考慮しながら実現するために，主に移動軌跡ストリームを対象とした連続的匿名化手法を提案する．また，センシティブ属性間の多様性の保証を，属性値を加工せずに実現する関係多様化を導入し，さらに関係の曖昧性を抑止しながら関係多様化を実現する手法を提案する．

提案手法を用いることで，系列データの提供をプライバシに配慮した形で実現し，特にリアルタイムな移動軌跡の提供や，属性間の関係に多様性 が保証された系列データの提供が，ある一定の精度を維持しつつ匿名性が保証された形で実現される．

目 次

図 目 次

表目次

第1章　緒言

1.1　本研究の背景

近年，個人に関する情報を記録したパーソナルデータが大量に蓄積され，第三者提供や二次活用によるパーソナルデータの活用が期待されている．しかし，パーソナルデータにはデータ主体である個人にとって他人に知られたくない情報（センシティブ属性）が含まれる場合があるため，パーソナルデータを活用する際にはデータ主体のプライバシへの配慮が必要となる．

パーソナルデータの活用の例として，医療機関が保持する患者の医療情報を活用したデータ分析が挙げられる．例えば，日本のセンチネル・プロジェクトに関する提言 [90] では，複数の医療機関が保持するレセプトデータ (診療報酬明細書 [1]) 等の医療情報を分析することで，「ある医薬品の使用者における特定の副作用 (有害事象) の発生頻度を，当該医薬品を使用していない場合の有害事象の発生頻度と比較することが可能」になると言われている．

医療分野においては，米国の HIPAA(Health Insurance Portability and Accountability Act) 法における必要最小限の情報開示の要件 (minimum necessary requirements)[72] では，医療情報を開示する際には開示する情報を必要最小限にすることが求められている．

[1] レセプトデータ (診療報酬明細書) とは，患者が受診した医療費について医療機関が健康保険組合などの保険者に請求する際の明細書のことである．診療報酬明細書は以前は紙であったが，現在は電子化が進んでいる [99].

また，異なる業種が保持するユーザのパーソナルデータを連携することで，新たなサービスが創出されることが期待されている [92]．パーソナルデータはプライバシに関わる情報であると同時に，企業における情報資産とも考えられているため，パーソナルデータを他の機関へ開示することは好ましくない．また，パーソナルデータのデータ主体である顧客やユーザのプライバシへの配慮が求められる．

1.2 匿名化によるプライバシ保護

パーソナルデータには，個人を明示的に識別する社員番号や学籍番号などの直接識別子 (Explicit Identifier) と，個人を特徴付ける生年月日や性別，職業などの準識別子 (Quasi-Identifier) が含まれる．準識別子の属性をいくつか組み合わせることで，パーソナルデータから個人を特定し得る．よって，直接識別子を切り落とすだけでなく，準識別子からの個人特定にも配慮した匿名化が必要となる．

データ匿名化は，所定の匿名性を満たすようにデータセットを加工する技術である．特に，準識別子からの個人特定の困難さを表す k-匿名性 [67] が広く知られている．k-匿名性は同一の準識別子の組を持つレコードが k 以上存在することを表す匿名性の指標であり，k-匿名性を充足することで準識別子から個人を特定が困難になり，一定のプライバシ保護が実現できる．k-匿名性を充足させる処理を k-匿名化と呼ぶ．k-匿名化は広く研究が行われており，様々な手法が提案されている．また，センシティブ属性の特定確率にまで踏み込んだ ℓ-多様性 [47] 等の k-匿名性を拡張した匿名性の指標やそれらを実現するデータ匿名化手法が提案されている．パーソナルデータをデータ匿名化することで，プライバシ侵害のリスクを低減した第三者提供や二次活用が実現できる．

表 1.1: パーソナルデータの例

ID	年齢	性別	傷病名
1	22	男	かぜ
2	28	男	HIV
3	33	男	HIV
4	42	男	かぜ
5	42	女	ガン
6	48	女	ガン

表 1.2: k-匿名化したパーソナルデータの例

年齢	性別	傷病名
[20, 29]	男	かぜ
[20, 29]	男	HIV
[20, 39]	男	HIV
[30, 49]	ANY	かぜ
[30, 49]	ANY	ガン
[30, 49]	ANY	ガン

表 1.1 にパーソナルデータの一例を示す．表 1.1 では，ID が直接識別子，年齢と性別が準識別子，傷病名がセンシティブ属性である．また，表 1.1 では，データ主体を一意に識別する直接識別子を用いずとも，準識別子である (年齢, 性別) の組からある個人のレコードを一意に識別できる．さらに，表 1.2 は，表 1.1 を k-匿名化 ($k = 3$) したパーソナルデータである．表 1.2 では，準識別子である (年齢, 性別) の組が重複するレコードが 3 レコード以上存在し，特定の個人のレコードを識別することが困難になっている．

このようにパーソナルデータに対してデータ匿名化を施すことで，一定の匿名性を持ったデータセットを生成することができ，データ主体のプライバシに配慮したデータセットの第三者提供や公開が実現できる．

1.3 系列データ

系列データは，複数の属性が連なったデータである．系列データを利活用することで，系列を成す属性間からパターンを発見したり，属性間の関係を得ることができる．例えば，位置情報の系列データである移動軌跡からは，人々の移動のパターンを分析でき，商圏分析等に活用することができる．また，診療情報の系列データに対しては，傷病間の因果関係分析や，特定の傷病の患者群の経過観察等を実現できる．さらにこれらの分析を複数の機関から収集したデータに対して実施すること等も期待されている．

系列データは，データ主体の直接識別子に複数のセンシティブ属性が紐づいたデータとする．

$$x = (id, q_1, \ldots, q_m, s_1, \ldots s_d) \tag{1.1}$$

ここで，q_i は準識別子，s_j はセンシティブ属性である．系列データに対してセンシティブ属性を特定しようとする攻撃には，準識別子からセンシティブ属性の特定だけでなく，あるセンシティブ属性から他のセンシティブ属性の特定もある．

表 1.3 は系列データの例の一つである医療データを示している．表 1.3 では，ID が直接識別子，年齢と性別が準識別子，傷病名と処方薬名がセンシティブ属性である．表 1.1 では，センシティブ属性は 1 つであったが，表 1.3 は複数のセンシティブ属性を持ち，センシティブ属性の系列を持っている．表 1.4 も，系列データの一例を示している．表 1.4 は，月ごとの傷病名をセンシティブ属性として持つ時系列データである．このような系列データからはセンシティブ属性間の関係を得ることができ，関係を活かした分析ができる．例えば表 1.1 では傷病名と処方薬名の傾向が分析でき，表 1.4 では，病歴の分析や，経過観察などのコホート分析を実現できる．

表 1.3: 系列データの例 1: 医療データ

ID	年齢	性別	傷病名	処方薬名
1	22	男	a	X
2	28	男	b	Y
3	33	男	b	Y
4	36	男	b	Z
5	42	女	c	W
6	48	女	c	X

表 1.4: 系列データの例 2: 医療データの時系列

ID	年齢	性別	傷病名 [4 月]	傷病名 [5 月]
1	22	男	a	a
2	28	男	b	c
3	33	男	b	a
4	36	男	b	b
5	42	女	c	b
6	48	女	c	c

また，タイムスタンプ毎のレコード (式 1.2) に分散しており，id に基づいて結合することで式 1.3 の形式になるデータも系列データの一つである．

$$x_{id}[t] = (id, t, q_t, s_t) \tag{1.2}$$

$$x_{id} = (id, q_1, \ldots, q_t, s_1, \ldots s_t) \tag{1.3}$$

ここで，q_t, s_t は時刻 (タイムスタンプ)t の準識別子，センシティブ属性である．

式 1.2 や式 1.3 の形式を取る系列データの例として，位置情報のシーケンスである移動軌跡がある (表 1.5, 1.6, 図 1.1)．表 1.5 では，ユーザの滞在先の位置情報を表す l

表 1.5: 系列データの例 3: 移動軌跡

u	l	t
Alice	(10, 5)	1
Alice	(15, 5)	2
Alice	(18, 8)	3
Alice	(18, 10)	4
Bob	(10, 5)	1
Bob	(9, 4)	2
Bob	(8, 3)	3
Bob	(8, 3)	4

表 1.6: 系列データの例 3: 移動軌跡 (系列)

u	$l[t_1]$	$l[t_2]$	$l[t_3]$	$l[t_4]$
Alice	(10, 5)	(15, 5)	(18, 8)	(18, 10)
Bob	(10, 5)	(9, 4)	(8, 3)	(8, 3)

がセンシティブ属性である．表 1.5 では，ユーザの識別子 u によって位置情報 l の系列を得ることができ，式 1.3 の形式に変換したデータが表 1.6 である．

位置情報の形態の一つである，ユーザやオブジェクトの位置を常時測位し，移動軌跡 (位置情報の系列) がリアルタイムに伸長していく移動軌跡ストリーム (表 1.7) も系列データである．移動軌跡ストリームの活用によって群衆の移動パターンのリアルタイムな分析が実現でき，リアルタイムな交通量分析や，移動傾向を利用した情報配信等が可能になる．しかし，リアルタイムなデータの提供によって，ストーキングや監視等のプライバシ侵害も可能になってしまうため，十分にプライバシへ配慮する必要がある．

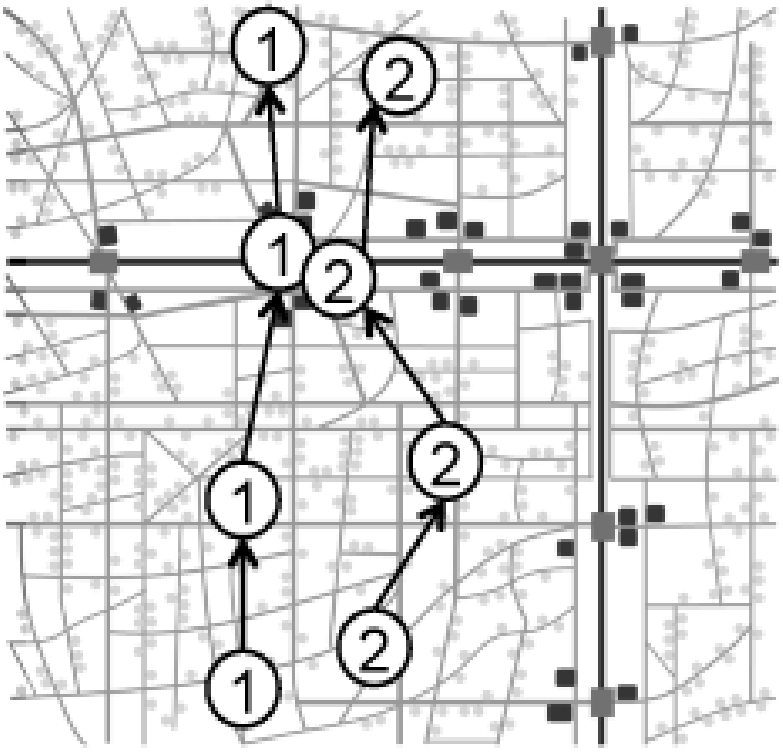


図 1.1: 移動軌跡

表 1.7: 系列データの例 4: 移動軌跡ストリーム

u	$l[t_1]$	$l[t_2]$	$l[t_3]$	$l[t_4]$	$l[t_5]$	...
Alice	(10, 5)	(15, 5)	(18, 8)	(18, 10)	(19, 9)	...
Bob	(10, 5)	(9, 4)	(8, 3)	(8, 3)	(8, 2)	...

1.4 本研究の目的と貢献

本研究では，系列データに対するデータ匿名化の問題を扱う．系列データに対する既存の匿名化技術は，蓄積された属性値の系列に対して静的に匿名化を行う技術であり，連続的に蓄積されていくデータを逐次匿名化することができなかった．また，系列データには複数のセンシティブ属性が含まれるため，あるセンシティブ属性に関する知識から，他のセンシティブ属性の属性値が特定されるというプライバシ侵害も生じ得る．このようなプライバシ侵害を防ぐためには，属性値や，属性間の関係に多様性を保証する必要がある．しかしながら，既存の手法は汎化に基づく手法であるため，属性値の多様性や，属性値間の関係の多様性といったより高度な匿名性を保証する際にセンシティブ属性の属性値が大きく曖昧化されてしまう問題がある．

本研究では，上述の系列データの匿名化に関する問題を解決するために，2 つの匿名化手法を提案する．

- 移動軌跡ストリームの連続的匿名化
- センシティブ属性間の関係多様化

移動軌跡ストリームの連続的匿名化に関する研究では，主に移動軌跡ストリームを対象として，これまで実現されていなかった系列データの連続的匿名化を提案する．提案手法では，移動軌跡ストリーム中の各時刻の位置情報の精度をある水準以上に保ちながらリアルタイムなプライバシ保護出版を実現する．

センシティブ属性間の関係多様化に関する研究では，あるセンシティブ属性から他のセンシティブ属性を特定しようとするプライバシ侵害 (攻撃) を想定する．このとき，センシティブ属性の属性値を汎化せずに，属性間の関係を曖昧化することで属性間の関係多様性を保証する関係多様化を導入し，あるセンシティブ属性から他のセン

シティブ属性の特定を困難にする．また，センシティブ属性間において，ある属性から他の属性がどの程度推測できるかを表す (ℓ_1, ℓ_2)-関係多様性という新たな匿名性の指標を定義する．さらに関係多様化を実現する手法として，関係多様化によって生じる属性間の関係の曖昧化を抑止する手法を提案する．

提案手法を用いることで，様々な種類の系列データの活用を安全に実現することができ，系列データのリアルタイムな活用の実現や，センシティブ属性間の関係の分析の一部を高い精度で実現できるようになる．これによってリアルタイムな移動パターン検出や，精度が維持された月次集計・パターン分析が実現できる．

1.5 本論文の構成

本論文の構成は次の通りである．まず，2 章で関連研究として，データ匿名化やその他プライバシ保護の既存技術について説明する．次に，3 章で移動軌跡ストリームに対する連続的匿名化を提案し，手法の有効性についても評価を行う．4 章にて複数のセンシティブ属性を持つデータセットに対するセンシティブ属性間の関係多様化について述べる．この章では，新たな匿名性指標 (ℓ_1, ℓ_2)-関係多様性を定義し，それを充足する関係多様化を導入する．また，関係の曖昧性を抑制した関係多様化の一手法を提案し，有効性について評価を行う．最後に，5 章にて本論文をまとめる．

第2章　関連研究

本章では，本論文で扱う系列データの匿名化に関連する研究について述べる．まず，パーソナルデータ等のプライバシ侵害の懸念のあるデータを，プライバシを考慮しながら活用するための技術について概観する．その上で，本研究のベースとなる技術であるデータ匿名化とその周辺の技術について述べる．特に本論文で扱う系列データに対する匿名化技術の既存研究を詳細に述べる．さらに，秘匿計算や PPDM，差分プライバシについてもこれまでの研究動向を概説する．

2.1　プライバシを考慮したデータ活用技術

パーソナルデータを活用することで，個々人やある集団に関する新たな知見の発見が期待できる．ここでパーソナルデータとは，個人を特定することができる個人情報にとどまらず，個人に関するあらゆる情報がパーソナルデータに該当する．パーソナルデータを活用するためには，パーソナルデータのデータ主体のプライバシに配慮する必要がある．

プライバシを考慮したデータ活用技術には，統計的開示抑制，データ匿名化 (Privacy Preserving Data Publishing, PPDP)，PPDM (Privacy Preserving Data Mining)，秘匿計算，差分プライバシなど，様々な技術がある．

統計的開示抑制は，国勢調査等の統計データを公開する際に個人が特定されないように統計データを加工する技術である．一定の水準を上回る値をまとめる処理である

トップコーディングや，特定のセルを秘匿状態にして公開しない等の処理を行うことで，機微な情報の漏洩を回避する．

データ匿名化は，統計や分析等を行う前のデータに対して，あるデータが誰に関する情報であるか，ある個人の属性値が何であるかを特定できないように，データを加工する技術である [24][26].

PPDM や秘匿計算は，暗号技術等を用いてデータを秘匿したまま，分析結果や集計結果のみ得る技術である．PPDM や秘匿計算は高い精度で計算結果を得ることができるが，計算を行うためのコストが大きいという課題がある．PPDM や秘匿計算は，計算過程は安全であるが，計算結果は必ずしも何らかのプライバシ保護がされているわけではなく，出力である計算結果から入力である元のデータセットの詳細が推定できる場合がある．このような出力に対するプライバシ保護の技術として出力プライバシがあり，その一つの技術として差分プライバシ [21] が近年注目され，多くの研究成果が創出されている [22][80][41][6][93]．差分プライバシでは，出力に対して摂動を加えることで，元のデータセットの詳細を隠蔽する．

また，データ保有者と利用者が互いにデータやクエリの詳細を明かさずに情報検索や情報推薦を行う Private Information Retrieval (PIR) という技術も存在する [16][20][61]. これまでは暗号を活用してデータやクエリを秘匿する手法が提案されてきた．近年では，クエリやユーザの思考を曖昧化したまま，情報検索や情報推薦を実現する技術も提案されている [63][28].

なお，これらの技術はそれぞれが想定する攻撃に対してのみ，一定の安全性や匿名性が保証される．そのため，想定する環境，アプリケーションのプライバシ保護の要件によっては，いくつかの技術の併用が必要である．

2.2 データ匿名化

プライバシ保護型データ出版 (Privacy-preserving data publishing) に関する技術が様々な分野で研究されている．データベースやデータ工学の分野では，リレーショナルデータベースに対するデータ匿名化が研究されている．データ匿名化とは，データセットから個人の特定を困難にする技術の一つである．データセット中の個人特定の困難さを匿名性と呼び，匿名性の指標として，k-匿名性 [67] が広く知られている．k-匿名性を充足させる処理を k-匿名化と呼び，様々な手法が提案されている．

従来，氏名や会員番号などの直接識別子 (明示的識別子, Explicit Identifier) を削除するといった処理が匿名化として認知されてきたが, k-匿名化に代表されるデータ匿名化では, 1 つ以上の属性の組合せで個人を特定し得る属性である準識別子 (Quasi-identifier, QI) を加工して，特定の個人に関するレコードの特定を困難にする．同一の準識別子を持つレコード群を等価クラス (Equivalence Class) と呼ぶ．データ匿名化では，所定の匿名性を満たすように等価クラスを生成する．さらに，レコードの特定だけでなく，他人に知られたくない属性であるセンシティブ属性 (Sensitive Attribute, SA) の特定を困難にするなど，攻撃のモデルに合わせた匿名化手段が研究されている．また，k-匿名化の実用化に向けた取り組みも行われている [105][104][98][101][103][102].

図 2.1(a) は，パーソナルデータの一例である．属性「識別子」が直接識別子，属性「年齢」と「性別」が準識別子である．例えば，年齢=22 のレコードは user1 のタプルのみであり，一意に識別することができる．属性「疾病名」はセンシティブ属性である．図 2.1(b) は直接識別子である「識別子」を取り除いたデータセットである．この状態であっても，前述の通り準識別子から特定の個人のタプルを特定することが可能である．どの属性が準識別子であるかを機械的に決定することは簡単ではなく，想定する攻撃や保護の要件に合わせて適宜変更する必要がある．

表 2.1: 匿名化の実行例

(a) 元テーブル

識別子	年齢	性別	疾病名
user1	22	男	かぜ
user2	28	男	HIV
user3	33	男	HIV
user4	36	男	かぜ
user5	42	女	ガン
user6	48	女	ガン

(b) 識別子を削除したテーブル

年齢	性別	疾病名
22	男	かぜ
28	男	HIV
33	男	HIV
36	男	かぜ
42	女	ガン
48	女	ガン

(c) 2-匿名化したテーブル

年齢	性別	疾病名
[20, 29]	男	かぜ
[20, 29]	男	HIV
[30, 39]	男	HIV
[30, 39]	男	かぜ
[40, 49]	女	ガン
[40, 49]	女	ガン

(d) 2-多様化したテーブル

年齢	性別	疾病名
[20, 29]	男	かぜ
[20, 29]	男	HIV
[30, 49]	ANY	HIV
[30, 49]	ANY	かぜ
[30, 49]	ANY	ガン
[30, 49]	ANY	ガン

2.2.1 攻撃モデル

データ匿名化では，想定する攻撃モデルに応じたデータの加工を行う．攻撃モデルは，Fung らによって整理されている [24]．最も多くの場面で想定される攻撃モデルに Record Linkage がある．Record Linkage は，特定の個人のレコードを特定する攻撃であり，準識別子の属性の組からレコードの特定を行う．さらに，Attribute Linkage では，特定の個人のセンシティブ属性値を一定の種類未満に，一定の確率以上に絞り込む．Table Linkage は，特定の個人のレコードが特定のテーブルに含まれるかどうかを一定の確率以上に絞り込む攻撃である．レコードの追加や削除，属性値の変化などのイベントの事前確率と事後確率の変化から変化の詳細を推測する Probabilisitic Attack という攻撃もある．

2.2.2 匿名性指標

匿名性の指標は，想定する攻撃を防ぐためにデータセットが満たすべき基準である．k-匿名性 (k-anonymity)[67] は，所定の条件に合致するレコードの重複度合いを表す指標であり，Record Linkage に対応する．対象とするテーブルの形式に応じた k-匿名性として，(X, Y)-Anonymity[74] や MultiR k-anonymity[58] 等が提案されている．

図 2.1(a) を $k = 2$ の k-匿名化したデータセットを図 2.1(c) に示す．準識別子である「年齢」と「性別」の値の組み合わせに対して，同じ値の組み合わせを持つタプルが 2 つ (k 個) 以上存在し，準識別子の値を知っていたとしても特定のデータ主体のタプルを k 未満に絞り込むことができない．

ℓ-多様性 (ℓ-diversity)[47] は，等価クラスに ℓ 種類以上のセンシティブ属性が存在することを表す指標であり，Attribute Linkage に対応する．

図 2.1(a) を $\ell = 2$ の ℓ-多様化したデータセットを図 2.1(d) に示す．準識別子である「年齢」と「性別」の値の組み合わせに対して，同じ値の組を持つタプル集合のセンシティブ属性に 2(ℓ) 種類以上の属性値が存在し，準識別子の値を知っていたとしても特定のデータ主体のセンシティブ属性値を ℓ 種類未満に絞り込むことができない．

t-近接性 (t-closeness)[42] は ℓ-多様性を拡張した指標であり，等価クラスに含まれるセンシティブ属性値の分布と，テーブル全体のセンシティブ属性値の分布との差異を一定以下であることを表す指標である．Confidence Bounding は [75]，センシティブ属性の推測の確信度が一定以下であることを表す．ℓ-多様性の拡張や，類似するプライバシモデルとして，(α,k)-anonymity[76]，LKC-Privacy[54] 等が知られている．

δ-存在性 (δ-presence)[56] は，Table Linkage に対応する匿名性指標であり，任意の個人のレコードが特定のテーブルに存在することの確信度を δ 以上に絞り込めないことを表す．また，k-匿名化を確率的指標へ拡張した Pk-匿名化 [94][96] や，同様に ℓ-

多様性を拡張した $P\ell$-多様性 [95] も提案されている．

Probabilistic Attack に対応する匿名性の指標として，m-invariance[78] や ϵ-differential privacy[21] などが提案されている．

2.2.3 匿名化手法

データ匿名化を行う手法として様々な手法が提案されている．各匿名化手法では，元のデータセットが持つ性質や分析精度がどの程度維持されているかを表す指標である有用性 (Utility) が高い匿名化結果を生成する．有用性の指標として NCP[82] 等が提案されているが，各手法はそれぞれ異なる有用性の指標に基づいた匿名化を行う．

有用性を最大化する最適な k-匿名化手法として，MinGen[67]，Binary Search[65]，Incognito[39] 等の手法が提案されている．しかしながら，これらの最適な k-匿名化は NP-困難な問題として知られている [49]．Datafly[66] は最初のスケーラブルな k-匿名化アルゴリズムである．最適な k-匿名化を行う場合には，一般化階層 (汎化木，タキソノミー) を利用して小さな探索空間で匿名化を行うことが一般的であり，このような方式を採るアルゴリズムとして Incognito[39] が知られている．

最適な k-匿名化は NP-困難であることから，多くの貪欲アプローチの手法が提案されている．Top-Down Specialization[25] や Mondrian[40] は，すべての準識別子を最も汎化した状態から，有用性が大きい状態を探索するアプローチ (トップダウンアプローチ) を採用した匿名化手法である．トップダウンアプローチは，k-匿名性等の所定の匿名性を充足した状態から探索を開始するため，探索先の有用性と匿名性が維持されているか否かのみを評価すればよく，スケーラブルな匿名化を実現可能である．

連続的データ出版

更新されていくデータベースを連続的に匿名化する技術についてもいくつかの技術が研究されている [78][79][27][89].

Xiao と Tao[78] は，動的に変化していくデータベースに対するプライバシ保護の指標として m-不変性 (m-invariance) を提案した．データベースにはレコードや値の追加，変更，削除が発生する．m-不変性はこのデータベースの動的変化によって，変化した部分と特定の個人との対応付けの困難さを表す指標である．m-不変性を充足したレコード群は，当該レコード群の変化を追跡することができず，特定の個人に合致する可能性のあるセンシティブ属性値を常に m 種類未満に絞り込むことができない.

Zhou ら [89] は，データストリームに対する連続的匿名化を提案している．Zhou らの手法では，受信したデータを一定時間蓄積して，類似するデータが k 以上蓄積集まった場合にはそれらを汎化して出版し，集まらなかった場合にはそれらの出版を行わない.

2.2.4 系列データの匿名化

位置情報の匿名化

近年のモバイル端末や各種センサから得られる位置情報の活用事例 [91][38] の増加から，位置情報に対する匿名化技術が広く研究されている．Chow と Mokbel [17], Bonchi ら [12] は，LBS(Location-Based Service) に関するプライバシ保護技術の全体像を報告している.

Beresford と Stajano [11] は位置情報のプライバシ保護の概念として，mix-zone というコンセプトを導入した．mix-zone とは交差点や密集地点などの移動体が交差する地点であり，mix-zone に入った移動体は同じ mix-zone に滞在する他の移動体と識別子の

シャッフルが行われ，mix-zone の入出によって移動体の追跡が困難になる．mix-zone を実世界の問題に適用するための設計方法についても報告されている [62][46].

位置情報に対する k-匿名化は Gruteser と Grunwald [31] らによって初めて導入され，. 位置情報に対する ℓ-多様化 [50] 等，拡張手法も提案されている．その後，位置情報のシーケンスである移動軌跡に対する k-匿名化も研究されている [1][53][86][7][70][57]. Abul ら [1] は，半径 δ 以内に k 以上の移動軌跡が存在することを表す (k, δ)-anonymity を移動軌跡の匿名性指標として導入した．さらに，移動軌跡をクラスタリングし，シリンダー状に汎化した移動軌跡を生成することで (k, δ)-anonymity を充足させる手法 NWA を提案した．Nergiz ら [57] は移動軌跡の要素である (緯度, 経度, 時刻) を直方体状に汎化して，この汎化された (緯度, 経度, 時刻) のシーケンスを生成して，k-匿名化を行う手法を提案している．さらに，汎化後の移動軌跡からランダムに座標を抽出して，粒度の高い移動軌跡をランダムに再構成する手法を提案している．

属性間の関係の曖昧化

テーブルの分割によって属性間の関係を曖昧化し，属性値の推定を困難にする技術が提案されている [77][3][35].

テーブルを分割し，準識別子のみから成るテーブル (QI テーブル) と，センシティブ属性のみから成るテーブル (SA テーブル) とを生成する手法 Anatomy が提案されている．Anatomy では，QI テーブルと SA テーブルとの間をグループ識別子等の曖昧な識別子によって接続することで，QI と SA 間の対応関係を曖昧にし，QI から SA の特定を困難にする．Aggarwal ら [3] は，テーブルの分割を行い，それぞれを論理的に異なる 2 つ以上のサーバに配置することで，分割された属性間の関係の再構築を抑止する手法を提案している．Jiang ら [35] は，関係従属性のある属性間の関係をテーブル分割によって曖昧化し，ℓ-多様性を充足させる手法を提案している．

トランザクションデータの匿名化

トランザクションデータは，各アイテムの有無をバイナリ形式で表すセンシティブ属性と見なすと，系列データの一種であると言える．トランザクションデータに対する匿名化手法もいくつか提案されている [71][83][33][45][15][13][84]．また，トランザクションデータとリレーショナルデータのハイブリッドなデータに対する匿名化手法も提案されている [69][64].

いずれの手法もトランザクション中のアイテム集合の重複性に基づく匿名化手法である．さらに，あるアイテムを知識とした他のアイテムの推定確率を一定以下に抑えることによるプライバシ保護手法についても提案されている [83][13].

トランザクションデータに対する匿名性の指標として k^m-匿名性 [71] が提案されている．k^m-匿名性は，m 個のアイテムが同一のレコードが k 以上存在することを表す．$m = \infty$ のときはすべての組合せを考慮するため，k-匿名性と同一である．

Terrovits ら [71] は，アイテムの概念階層を定義した一般化階層を用いて，アイテムを一般化することで k-匿名化する手法を提案している．He ら [33] は，一般化階層を用いて，特定のレコードの特定のアイテムだけを一般化 (局所的再符号化) することで k-匿名化する手法を提案している．Xu ら [83] は，匿名性の違反の要因となるアイテムをテーブルから削除することで，一般化階層を用いずに匿名化を実現する手法を提案している．

分散データの匿名化

複数の機関が分散して保持するテーブルを結合して匿名化する処理を分散匿名化 (Distributed Anonymization) と呼ぶ [51][73][34][36] [107][108][106]．分散匿名化が対象とするデータセットでは，共通の識別子によって結合することで分散されていた属

性の系列が得られる．そのため，系列データの一種と考えることができる．

分散匿名化は，パーソナルデータの分割形態の違いにより垂直分割と水平分割に分類される．垂直分割とは，パーソナルデータが属性毎に異なる機関に保持された分割形態である．水平分割とは，パーソナルデータがユーザ毎に異なる機関に保存された分割形態である．

垂直分割での分散匿名化としては [51][73][34] などが存在する．パーティ間での匿名化をセキュア計算 (secure computation)[44][85] を組み合わせて実現する手法が提案されている [51][73]．水平分割での分散匿名化としては，[36] が知られている．

2.3 秘匿計算と PPDM

2.3.1 セキュア計算と Multi Party Computation

セキュア計算とは，複数の機関が持つ値を互いに秘密にしながらそれらの値を入力とした演算を実現する暗号プロトコルである [44]．セキュア計算の暗号プロトコルは，Yao による研究 [85] が始まりとされている．Yao[85] は，信頼のおける第三者 (Trusted Third Party, TTP) が存在しないという仮定において，2 機関がそれぞれ持つ秘密の値を引数とする任意の関数が計算可能であることを示した．その後 [30][29] において，複数機関が持つ秘密の値に対応するように拡張され，Multi Party Computation(MPC) と呼ばれている [10][9].

2.3.2 Privacy Preserving Data Mining

PPDM (Privacy Preserving Data Mining) とは, 複数の機関が持つ値を, 互いに秘密にしながらデータマイニングを行った結果を得る技術である [97][100][2][88][43][4][23][18].

PPDM は，データ匿名化とは異なり，対象とするデータに対してデータマイニングを行う点が大きな違いである．

PPDM では, MPC やセキュア計算などの暗号プロトコルを利用する手法と, ノイズを付加する手法とが存在する．例えば暗号プロトコルを利用する手法 [88][43][4][23][18] では，セキュア計算を用いた近傍検索を行う手法 [88][18][4] や，分類木を作成する手法 [43] などが提案されている．

ノイズを付加する手法としては [5] が良く知られている．この手法は，ある確率分布のノイズを付加したデータから分類木を作成する手法である．まず，ある機関が持つ秘密の値 $\{x_1, \cdots, x_n\}$ に対して確率分布 Y の乱数 $\{y_1, \cdots, y_n\}$ を付加し，乱数が付加された値 $\{w_1 = x_1 + y_1, \cdots, w_n = x_n + y_n\}$ を公開する．そして，この乱数が付加された値を受け取った機関は，確率分布 Y を知っている前提において，公開された $\{w_1 = x_1 + y_1, \cdots, w_n = x_n + y_n\}$ から，元の値である $\{x_1, \cdots, x_n\}$ の確率分布を推定する．[5] では，ベイズの定理を用いて元の値の確率分布を推定する手法を提案している．つまり，たとえ乱数が付加されたとしても，乱数の確率分布を知っていれば元の値の分布を推定でき，分類木を作成可能である．

2.4 差分プライバシ

差分プライバシ (Differential Privacy) は，データベースに対する応答から，データベース中の各レコードの詳細や機微な変化を隠蔽するプライバシモデルである．差分プライバシは，PPDP と PPDM の双方で利用できる技術であり，近年注目されているプライバシモデルである．ϵ-differential privacy[21] はデータベースの 1 レコードの変化を秘匿する差分プライバシを実現するための指標の一つである．また，差分プライバシのフレームワークとして PINQ(Privacy Integrated Queries) が提案されている [48].

差分プライバシでは，データベース応答に対して所定の性質を満たすノイズを加える．このノイズを付加するメカニズムとして Laplace Mechanism が広く知られている [21]．差分プライバシによるプライバシモデルでは，1 レコードの変化によってどの程度データベース応答が変化するかという考え (Sensitivity) に基づいて，ノイズを加える．また，複数回の問い合わせや，応答に与える影響力が高い問い合わせに対しては大きなノイズを加える必要がある．出力されるデータの有用性を高める (付与されるノイズを抑制する) ために，目的に併せたノイズを付与するメカニズムの研究が広く行われている [41][32][52][81][87][59].

第3章　移動軌跡ストリームの連続的匿名化

3.1　概要

近年，携帯端末や自動車の位置情報を取得し，情報配信等を行う位置情報サービス (Location-Based Service, LBS) として様々なサービスが展開されている．しかしながら，収集される位置情報は自宅や勤務先，通院先などの他人には知られたくない場所への滞在情報を表し，高いプライバシ性を有する．さらに近年では，位置情報をリアルタイムかつ連続的に測位してユーザの移動軌跡や行動履歴を収集するサービスも存在する．そのような環境では，ユーザの移動軌跡が一種のデータストリームとして収集され，移動パターンのリアルタイム分析等に活用することができる．移動軌跡のストリーム (移動軌跡ストリーム) がリアルタイムに公開されれば，大規模な移動軌跡ストリームを活用してユーザの導線解析によるマーケティングへの活用や，移動パターン解析による伝染病の感染経路のシミュレーション等を，今現在移動している移動体の位置情報を活用して実現することが可能になる．

しかしながら，移動軌跡は個々人に特有の情報であるため，明示的な識別子や氏名等の個人を識別する情報を削除したとしても，移動軌跡と対応する個人を紐づけることは容易である．例えば，自宅と勤務先の位置情報の組み合わせからでも，高い確率で対応する個人の移動軌跡を一意に絞り込むことができる．移動軌跡を特定され

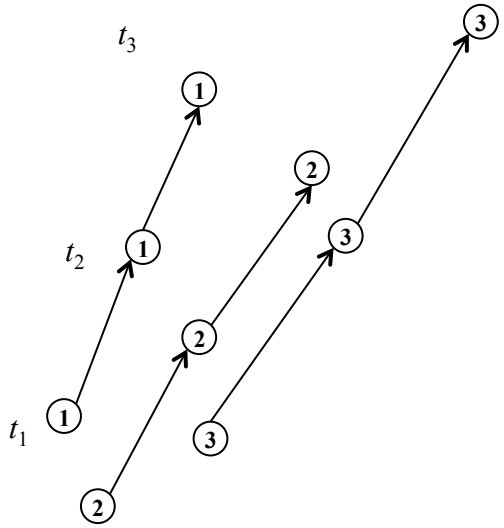


図 3.1: 移動軌跡

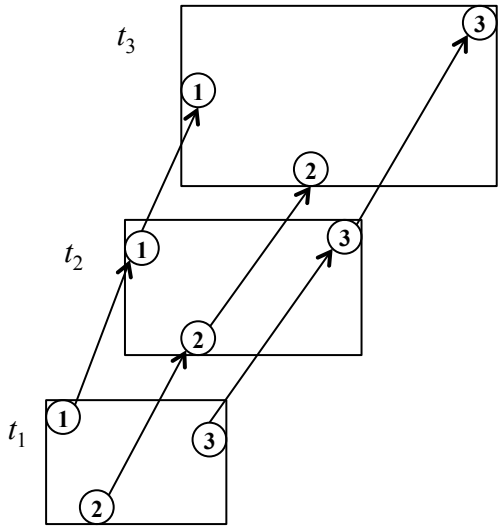


図 3.2: 3-匿名化した移動軌跡

ると，すべての位置情報が露になり，プライバシ性の高い場所への滞在，不在が漏洩してしまう．さらに，リアルタイムに移動軌跡ストリームを提供するような状況下では，常に追跡や監視といった脅威に晒される可能性がある．そこで，高いプライバシ性を持つ移動軌跡ストリームを第三者に提供する際には，移動軌跡のデータ主体のプライバシに配慮する必要がある．

2 章でも述べたように，データセットから個人の特定を困難にする技術の一つとしてデータ匿名化が知られている．移動軌跡に対するデータ匿名化として，蓄積された移動軌跡を静的に匿名化する技術がいくつかの手法が研究されている [1][57][70][53][86][7]. 位置情報の系列データである移動軌跡が k-匿名性を満たすためには，k 個の移動軌跡が常に同じ場所に滞在する必要がある．図 3.1 は移動軌跡の例であり，図 3.2 は k-匿名化 ($k = 3$) した移動軌跡の例である．Abul ら [1] は，k-匿名性 [67] を拡張した (k, δ)-anonymity という匿名性指標を定義し，蓄積された静的な移動軌跡をチューブ状に汎化する匿名化手法を提案している (図 3.3)．Nergiz ら [57] らは (緯度, 経度, 時刻) を汎化したシーケンスを生成して k-匿名化する手法を提案している (図 3.4).

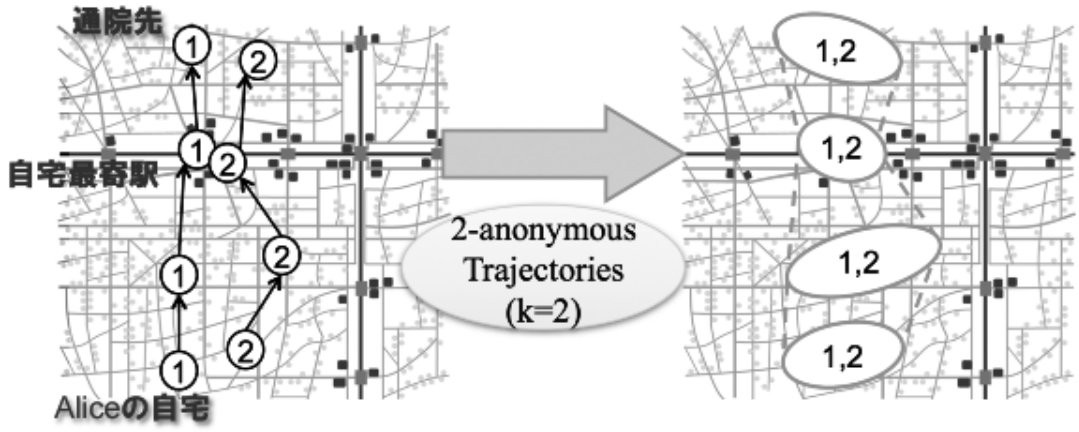


図 3.3: NWA による匿名化

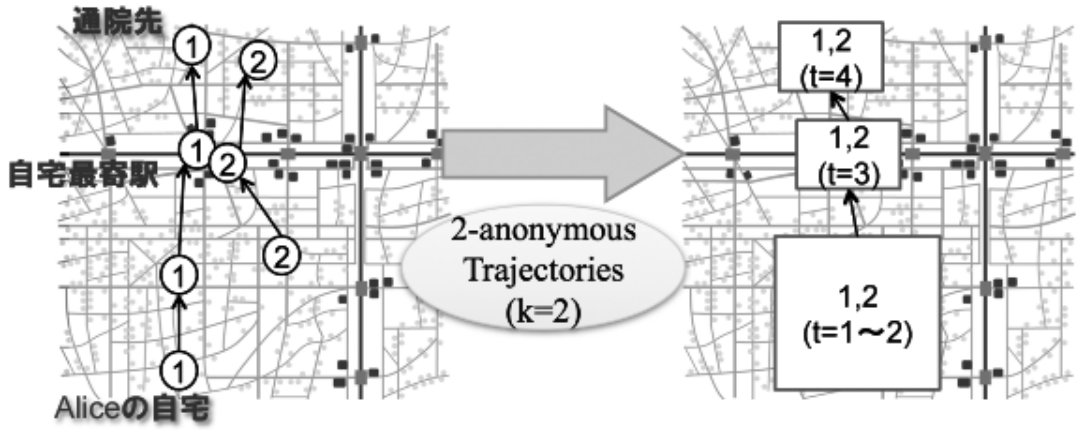


図 3.4: Nergiz の手法による匿名化

ただし，これらの手法はリアルタイム環境における連続的な匿名化を考慮したものではない．既存の手法は，既に蓄積された移動軌跡に対してその全容を用いた移動軌跡のグルーピングを行う技術であるため，時々刻々と変化する状況に応じた匿名化を実現できない．そのため，リアルタイム環境で適用した際には抽象度の増大 (有用性の低下) といった問題が生じる．

本研究では，上述のように機微性の高い移動軌跡ストリームに対するリアルタイムな匿名化に関する問題を扱う．時々刻々と測位される移動軌跡からユーザを特定できないように連続的に匿名化を行うことで，プライバシを保護した形で人々の移動軌跡ストリームをリアルタイムに利用可能にすることを目的とする．提案匿名化手法 CMOA (Continuous Moving Objects Anonymization) は，移動軌跡ストリームを成す位置情報群のうち，新たに測位された位置情報を既に匿名化されている情報を考慮しながらリアルタイムかつ連続的に匿名化を行う手法である (図 3.2)．また，時々刻々と変化する人々の移動に合わせて匿名グループの構成を動的に再構成することで，過度な抽象化を抑制する．評価実験では，提案手法が一定の解像度を維持しながら連続的に匿名化できることを確認した．また，10 万人程度の移動軌跡ストリームを毎分リアルタイムに匿名化できることを確認した．

本章の以降の構成は以下の通りである．3.2 節では，本研究が取り組む問題の定義と，提案手法の論述に必要な基本事項について述べる．3.3 節では，移動軌跡ストリームに対する連続的匿名化手法を提案する．3.4 節では，提案手法の有用性を評価するために行った評価実験について報告する．最後に 3.5 節で，本章をまとめる．

3.2 問題定義と準備

本研究では，オリジナルの位置情報を l，l を加工した範囲を持つ位置情報 (エリア) を l^* とする．各タイムスタンプには，すべてのデータ主体の位置情報が欠損なく生成されることを前提とする．

3.2.1 移動軌跡ストリーム

各ユーザの位置情報 l は一定のインターバル毎に測位され，タイムスタンプ t を付加して信頼できるプラットフォームに蓄積されるものとする．また，T をすべてのタイムスタンプの集合とする．l の例として，緯度・経度によって表される座標等がある．

定義 1 *(移動軌跡ストリーム):* データ主体 u の移動軌跡ストリームは位置情報の時系列で表される*:* $\tau_u = \{(u, l_0, t_0), (u, l_1, t_1), \ldots, (u, l_{last}, t_{last})\}(t_0 < t_1 < \cdots < t_{last})$．ここで，$t_{last}$ は最新のタイムスタンプ *(*現在時刻*)* とする．また，時間 $[t_i, t_j]$ の移動軌跡を以下で表す*:* $\tau_u[t_i, t_j] = \{(u, l_i, t_i), (u, l_{i+1}, t_{i+1}), \ldots, (u, l_j, t_j)\}(t_i < \cdots < t_j)$．ここで，時刻 t の位置情報は $\tau_u[t_i](= l_i)$ と表す．移動軌跡 τ_u において，時間 $[t_i, t_{i+1}]$ にユーザ u は l_i から l_{i+1} へ移動する．

移動軌跡ストリームの例を表 3.1 に示す．表 3.1 では，Alice の移動軌跡ストリームは $\tau_{Alice} = \{(Alice, (10, 5), 1), (Alice, (15, 5), 2), (Alice, (18, 8), 3)\}$ である．また Bob の時刻 t_2 の位置情報は $\tau_{Bob}[t_2] = (9, 4)$ である．

3.2.2 移動軌跡ストリームのプライバシ

移動軌跡ストリームの最新の位置情報を受信するたびに受信した位置情報を第三者へ提供することを想定する．このとき第三者の信頼性に関しては仮定をおかない．

表 3.1: 移動軌跡

u	l	t
Alice	(10, 6)	0
Alice	(10, 5)	1
Alice	(15, 5)	2
Alice	(18, 8)	3
Bob	(10,6)	0
Bob	(10,5)	1
Bob	(9, 4)	2
Bob	(8, 3)	3

表 3.2: 汎化移動軌跡 (2-匿名化)

	tid	l^*	t
Alice	1	([10,10], [6,6])	0
Alice	1	([10,10], [5,5])	1
Alice	1	([9,15], [4,5])	2
Alice	1	([8,18], [3,8])	3
Alice	1	([10,10], [6,6])	0
Bob	2	([10,10], [5,5])	1
Bob	2	([9,15], [4,5])	2
Bob	2	([8,18], [3,8])	3

よって，移動軌跡ストリームは悪意のある第三者 (以降，攻撃者) に暴露する可能性がある．ここで，攻撃者はデータ主体 u の移動軌跡ストリーム τ_u 中の位置情報もしくはエリアをいくつか知っているものとする．このとき，攻撃者が，保有する τ_u に関する知識を用いて移動軌跡ストリームのデータセットから τ_u を特定しようとすることを攻撃として想定する．攻撃が成功した場合，対象のデータ主体の移動軌跡が特定され，攻撃者は事前知識以上の情報を得ることができる．

定義 2 *(攻撃者の知識):* 攻撃者はデータ主体 u の軌跡 τ_u の一部 $A_u \subseteq \tau_u$ を既知であるとする．このとき，A_u をデータ主体 u に関する攻撃者の知識とする．

A_u の任意の要素の組み合わせは，攻撃対象 u の移動軌跡ストリームを特定し得る．もし u の移動軌跡ストリームが移動軌跡ストリーム集合から特定されてしまえば，攻撃者は事前知識 A_u に有していなかった新たな知識 (滞在情報) を発見できる．

定義 3 *(プライバシ侵害):* 攻撃者の事前知識 A_u に一致する移動軌跡ストリームの数が所定の閾値 k 未満に絞り込まれたとき，データ主体 u にプライバシ侵害が生じるとする．

攻撃者が Alice に関して $A_{Alice} = \{((10,5),1),((15,5),2)\}$ を持ち，$k = 2$ であるとする．このとき，表 3.1 では Alice にプライバシ侵害が生じる．なぜならば，A_{Alice} を用いると，Alice の移動軌跡ストリームを k 個未満に特定できてしまうためである．このようなプライバシ侵害を防ぐためには，オリジナルの移動軌跡ストリームを匿名化する必要がある．特に，本研究では k-匿名性を充足するようなプライバシ保護を連続的に施すことで，移動軌跡ストリームを匿名化する問題を扱う．

3.2.3 移動軌跡ストリームの連続的出版における匿名性

移動軌跡ストリームの特定を困難にするには，移動軌跡ストリーム中の各々の位置情報が，他の移動軌跡ストリーム中にも出現する必要がある．このように，異なるデータ主体が特定の時間に同じ位置に滞在することを Co-local と呼ぶ．

定義 4 *(Co-locality):* 移動軌跡ストリーム τ_1 と τ_2 が時刻 t に $\tau_1[t] = \tau_2[t]$ であるとき，τ_1 と τ_2 は時刻 t に *Co-local* である．また，$\forall t \in [t_i, t_j]$ に関して $\tau_1[t] = \tau_2[t]$ であるとき，τ_1 と τ_2 は時間 $[t_i, t_j]$ に *Co-local* である．なお，特に時間を指定せずに *Co-local* と表現する場合には，$\forall t \in T$ において $\tau_1[t] = \tau_2[t]$ であることを指す．

一般に，位置情報のシーケンスである移動軌跡ストリームでは，複数の時刻において Co-locality を保証することは難しい．そのため，複数のデータ主体の移動軌跡ストリーム中の位置情報 l を範囲を持った情報 (エリア)l^* に変換 (汎化) することで，複数の移動軌跡ストリームで Co-locality を保証する．

定義 5 *(*汎化移動軌跡ストリーム*):* 汎化移動軌跡ストリーム τ^*_{tid} はエリアと時刻のシーケンスである*:* $\tau^*_{tid} = \{(tid, l^*_1, t_1), (tid, l^*_2, t_2), \ldots, (tid, l^*_m, t_m)\}$ $(t_1 < t_2 < \cdots < t_m)$. tid は汎化移動軌跡ストリームの識別子であり，ランダムに割り当てた値を用いる．汎化移動軌跡ストリームの時刻 t_i のスナップショットであるエリアを $\tau^*_{tid}[t_i](= (l^*_i))$ で表す．汎化移動軌跡ストリーム τ^*_1 と τ^*_2 が $\forall t \in [t_i, t_j]$ に関して $\tau^*_1[t] = \tau^*_2[t]$ であるとき，τ^*_1 と τ^*_2 は時間 $[t_i, t_j]$ に *Co-local* である．なお，特に時間を指定せずに *Co-local* と表現する場合には，$\forall t \in T$ において $\tau^*_1[t] = \tau^*_2[t]$ であることを指す．

表 3.2 に表 3.1 を加工した汎化移動軌跡ストリームを示す．表 3.2 の Alice と Bob の移動軌跡ストリームが時間 $[1, 3]$ に Co-local になるように汎化している．

定義 6 *(*等価クラス*):* 時刻 t において，t 以前のすべての時刻で互いに *Co-local* な汎化移動軌跡ストリーム *(*tid*)* の集合を等価クラス *(Equivalence Class)* と呼ぶ．時刻 t におけるある等価クラスを $EC_{cid}[t]$ で表す．ここで cid は等価クラスの識別子である．

定義 7 *(EC* 履歴*): EC* 履歴 h_u をデータ主体 u が属する等価クラス $EC[t]$ のシーケンスとする．u が t に属する EC のメンバを $h_u[t](= EC_{cid}[t])$ で表す．

表 3.3 は EC 履歴の一例を示している．表 3.3 には 7 人のデータ主体の移動軌跡ストリームの汎化移動軌跡ストリームが属する．表中の各々セルに記載された EC_i は各 tid の汎化移動軌跡ストリームが時刻 t_j に属する EC である．

定義 8 *(*移動軌跡ストリームの k-匿名性*):* 移動軌跡ストリーム τ_u*(*または汎化移動軌跡ストリーム τ^*_{tid}*)* は，時間 $[t_i, t_j]$ に *co-local* な他の移動軌跡ストリームが $k-1$ 個存在するとき，時間 $[t_i, t_j]$ に k-匿名性を満たす．また，*EC* 履歴において $|\cap_{t\in[t_i,t_j]} h_u[t]| \geq k$ を満たすとき，データ主体 u に関する汎化移動軌跡ストリームは，k-匿名性を満たす．

移動軌跡ストリームに対して連続的に匿名化する場合には，既に k-匿名性を満たす EC を鑑みて，新たに到着した位置情報の汎化を行う必要がある．

表 3.3: EC 履歴

tid	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7
1	EC_1	EC_1	EC_1	EC_1	EC_1	EC_5	—	—
2	EC_1	EC_1	EC_1	EC_1	EC_1	EC_5	—	—
3	EC_2	EC_2	EC_2	EC_3	EC_3	EC_5	—	—
4	EC_2	EC_2	EC_2	EC_3	EC_3	EC_5	—	—
5	EC_2	EC_2	EC_2	EC_3	EC_3	EC_5	—	—
6	EC_2	EC_2	EC_2	EC_4	EC_4	EC_4	EC_4	EC_4
7	EC_2	EC_2	EC_2	EC_4	EC_4	EC_4	EC_4	EC_4
8	—	—	—	—	—	—	EC_6	EC_6
9	—	—	—	—	—	—	EC_7	EC_7
10	—	—	—	—	—	—	EC_7	EC_7
11	—	—	—	—	—	—	EC_7	EC_7
12	—	—	—	—	—	—	EC_6	EC_6

表 3.2 は表 3.1 を $k = 2$ の k-匿名化した汎化移動軌跡ストリームである．表 3.2 の Alice と Bob の移動軌跡ストリームを汎化することで 2-匿名性を達成している．

3.3 提案手法: CMOA

移動軌跡ストリームをインターバル毎に連続的に k-匿名化する手法 CMOA (Continuous Moving Object Anonymization) を提案する．各タイムスタンプでは，新たな位置情報が到着し，既に k-匿名化された過去の移動軌跡を鑑みて，新たな位置情報を匿名化する．この匿名化を次の位置情報が到着するまでに完了する必要がある．

移動軌跡ストリームに対する連続的匿名化では，将来の位置情報を把握することが困難であるため，将来に渡って最適な k-匿名化を行うことは困難である．また，次の位置情報が到着するまでに匿名化を完了する必要があるため，各時刻の中で最適な組み合わせで k-匿名化を行うことも困難である．

3.3.1 基本的なアイディア

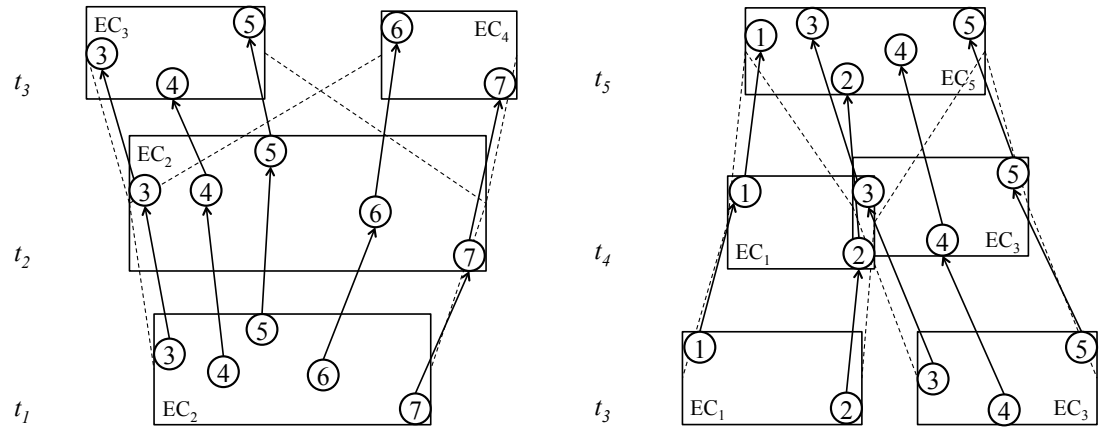
新しい時刻 t_{last} では，$EC[t_{last}]$ を EC 履歴に基づいて生成する．このとき，位置情報 l を汎化してエリア l^* を生成する．特に本稿では，エリア l^* を $EC[t_{last}]$ の移動軌跡の位置情報すべてを包含する最小包囲矩形として生成する．

まず，CMOA は初期座標である位置情報の集合に対して**初期分割**を行い，k-匿名性を充足する汎化移動軌跡ストリームと EC 履歴を生成する．以降は，各時刻 t_i では移動軌跡ストリームの新たな位置情報に対して，$EC[t_{i-1}]$ を継続して $EC[t_i]$ を生成する (**継承匿名化**).

解像度とトレーサビリティ

前述の継承匿名化では, 現在の位置情報と過去の移動軌跡のみを考慮して EC を生成する．EC のメンバを継続していくことで，EC 履歴の k-匿名性は保たれるが，メンバの移動方向が変化していくことで EC のエリアが過度に大きくなってしまう (図 3.2). エリアが大きくなっていくと，位置情報の精度が劣化する．ここで，位置情報やエリアの面積を S とし，位置情報の精度を表す尺度として解像度 (Resolution)$r = S^{-1/2}$ を導入する．

解像度を一定レベルに保つために，CMOA は**動的再構成 (Dynamic Reconstruction, DR)** を導入する．本研究の DR では，等価クラス EC の**分割**をベースとする．EC を複数の EC に分割することでエリアの小さい EC を生成する．ただし，k-匿名性を維持するためには，メンバ数 k 未満の EC に分割することができない．メンバ数が小さい EC は分割不可であるため，このような EC を複数集約する**併合**を行うことで，再び分割可能とする．



(a) 分割　　(b) 併合

図 3.5: 動的再構成

図 3.5(a), 図 3.5(b), 図 3.6 に DR の例を示す．図中の円が移動軌跡ストリームの位置情報 l を表し，矩形が汎化移動軌跡ストリームのエリア l^* を表す．円の中に記載された数字は汎化移動軌跡ストリームの tid である．

分割と併合によって動的再構成を行った場合には，EC のメンバ数が k 以上であっても k-匿名性に違反する場合がある．図 3.6 では，$EC_1[t_3]$ と $EC_3[t_3]$ が併合され $EC_5[t_4]$ が生成された．そして t_5 では， $EC_5[t_4]$ が分割されて $EC_6[t_5]$ と $EC_7[t_5]$ が生成された．このとき，$EC_6[t_5]$ と $EC_7[t_5]$ は共にメンバ数が k 以上である．しかし，Alice の汎化移動軌跡ストリーム τ_1^* は $h_{Alice}[3] \cap h_{Alice}[5] < k$ であるため，k-匿名性を損失している．同様に τ_2^* と τ_3^* も k-匿名性を損失している．

k-匿名性の損失を解消するために，CMOA は tid の付け替え (bf TID 再割当) を行う．tid を全く異なる値の新しいものに付け替えることで，その前後で移動軌跡をトレースすることができなくなる．そのため，tid の再割当前と後でそれぞれ k-匿名性は維持され，TID 再割当の前後をまたがった対応付けは困難になる．よって，k-匿名性の違反が解消される．表 3.3 では, Alice, Bob, Chris, David, Ellen の tid がランダム

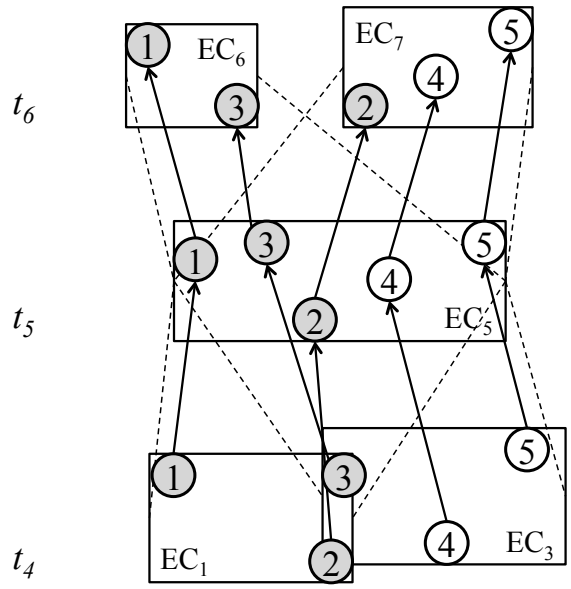


図 3.6: k-匿名性の損失

に再割当されている．再割り当て後は，例えば Alice の場合，$h_{Alice}[3] \cap h_{Alice}[5] = \emptyset$ となる．

このようにして tid を用いた明確な対応付けは困難になる．その一方で k-匿名性の代わりに移動軌跡のトレーサビリティが損失する．たとえトレーサビリティが損失したとしても，プライバシ侵害を生じさせないためにも k-匿名性を充足させる必要がある．よって，本研究では，k-匿名性の充足を優先する．

しかし，トレーサビリティができるだけ保たれることは，移動軌跡の有用性の観点では重要である．そこで，DR では解像度に加えてトレーサビリティについても考慮して再構成を行う．分割が頻発すると，トレーサビリティの損失を招いてしまうため，エリアの面積が閾値 σ を超えた場合に分割を行う．σ を超えない場合は直近の EC のメンバをそのまま引き継ぐ．さらに，併合ではトレーサビリティが高くなるように複数の EC の集約を考える．

クラスエネルギー

併合において，トレーサビリティを維持できる度合いを測る指標としてクラスエネルギーを導入する．分割可能回数が多く，面積が小さいクラスは，TID 再割当が生じるまでの猶予が長く，トレーサビリティが高いと考えることができる．よって，クラスエネルギーはこの TID 再割当が生じるまでの猶予を表す指標とする．

定義 9 *(クラスエネルギー): $EC[t]$ のクラスエネルギー $E(EC[t])$ を以下の式で表す:*

$$E(EC[t]) = \frac{\sigma}{S(EC[t])}(1 + \log_p \frac{|EC[t]|}{k}) \tag{3.1}$$

ここで，p は分割数であり，$S(EC[t])$ は $EC[t]$ のエリアである．$\log_p \frac{|EC[t]|}{k}$ は分割可能回数の期待値である．

3.3.2 CMOA のアルゴリズム

これまで述べた基本的なアイディアを基に，CMOA は以下の手順によって移動軌跡ストリームを連続的に匿名化する．

1. 新たな位置情報を受信し，$t_{last} = t_0$ なら 2. へ，$t_{last} > t_0$ なら 3. へ
2. 初期分割によって EC を生成し，6. へ
3. 継承匿名化によって EC を生成
4. 分割による動的再構成
5. 併合による動的再構成
6. 匿名化したデータを出版し，1. へ

CMOA はまず初期分割により, 移動軌跡ストリームの初期座標から k-匿名な EC を生成する (ステップ 2). 各時刻 t_i では, t_{i-1} のメンバ構成を引き継いで $EC[t_i](= EC[t_{i-1}])$ を編成する (ステップ 3). 続いて, 各 EC のエリアの面積が閾値 σ を超えたら動的再構成 (DR) を行う. DR はエリアの面積を σ 以下に保つために, まず分割を試行し (ステップ 4), その後併合を試行する (ステップ 5). このとき, 分割によって k-匿名性に違反した場合には TID 再割当を行う. 以降の各節で, それぞれについて詳細に述べる.

3.3.3 初期分割

移動軌跡ストリーム群の初期座標である位置情報を受信したら, CMOA は初期化として初期分割を行う. 初期分割では, 各位置情報が k-匿名性を満たすように EC を生成する. このとき, 将来の移動方向が全く未知であるため, 分割によって移動方向が類似する EC へと分割するための余地を残すために, EC のエリアの面積は σ 以下になったら, それ以上の分割を行わない.

EC を生成するために, クラスタリングによる空間分割を行う. まず, すべての位置情報から一つのクラスタを生成する. 次に, p-分割 (以降, 問題を簡単にするために $p = 2$ を想定する) を再帰的に繰り返すことで, クラスタの面積を σ 以下にする. クラスタの面積が σ 以下になる前に, クラスタのメンバ数が k を下回ってしまった場合は, 当該クラスタに属する位置情報の移動軌跡ストリームを**秘匿状態**とする. ここで秘匿状態とは, CMOA で処理した結果を出版しない状態である. 秘匿状態の移動軌跡ストリームは, 継承匿名化の対象とならず, 併合処理で他の EC に併合されるまでは匿名化結果として出版されない. 秘匿状態の移動軌跡ストリームの集合を **SUP** とする.

クラスタリング後, 秘匿状態でないクラスタから EC を生成する. EC 中のすべての

移動軌跡ストリームはk以上のメンバからなり，すべてのメンバの位置情報を包含する最小包囲矩形Rへと座標(x, y)が汎化される．生成したECを匿名化データとして出版する．このとき，秘匿状態の移動軌跡ストリームはk-匿名性を満たさないため，匿名化結果には含めない．秘匿状態の移動軌跡ストリームは動的再構成のDRで他のECに併合されるまで出版されない．

クラスタリング手法の一つであるk-means法では，クラスタのセントロイドと各点の二乗距離が小さくなるようにクラスタリングを行う．本研究では，k-means法を拡張したk-means++法 [8] を再帰的に利用して初期分割を行う．クラスタリングは，位置情報中の座標情報である(x, y)を対象とする．

3.3.4 動的再構成

動的再構成DRは，ECのメンバ構成を変更し，より解像度が高い，またはよりクラスエネルギーの高いクラスを生成する処理である．本研究では，分割と併合を主たる操作として用いる．両操作は，ECの面積$S(EC[t])$が閾値σを超えた際に行う．DRによって元のECは新たなECで置き換わる．

分割

分割では，1つのECをp個の独立したECに分割する．初期分割と同様に，本研究では問題を簡単化するために$p = 2$で分割することを考える．

分割はエリアの面積$S(EC[t])$がσを超えたECを対象に行う．ここで，分割対象のECをEC_oとする．また，EC_oを分割して生成された2つのECをEC'_a, EC'_bとする．EC'_aとEC'_bは互いに独立であり，共にEC_oの部分集合である．EC'_aとEC'_bが共にkメンバ数以上で成り立っていれば，EC_oと置き換える．

Algorithm 1 動的再構成

Require: equivalence classes $\mathbf{EC}[t]$, suppressed trajectories **SUP**

1: $M \leftarrow partition(\mathbf{EC})$
2: $merge(M, \mathbf{SUP})$

表 3.3 では，図 3.5(a) のように EC_2 が EC_3 と EC_4 に分割されている．EC_3 と EC_4 は k-匿名性を充足している．分割することで，エリアの面積も小さくなっていることがわかる．

分割の結果，メンバ数が k 未満の EC が生じる場合がある．メンバ数 k 未満の EC は k-匿名性を満たさないため，出版できない．そのため秘匿状態とする必要がある．一方で，メンバ数が k 未満の EC が外れ値であれば，この EC を取り除くことで他の移動軌跡ストリームは解像度を維持できる．

分割の結果，メンバ数が k 未満の EC が生じた場合には以下の 2 つの選択肢が考えられる．

1. 分割を承認してメンバ数が k 未満の EC の移動軌跡ストリームを秘匿状態にする
2. 分割を承認せず，ロールバックして併合に回す

本研究では，分割した結果，メンバ数が k 以上のクラスのクラスエネルギーが分割前と比較して向上していれば，分割を承認するものとする．ここで，メンバ数が k 以上のクラスを EC'_1, k 未満のクラスを EC'_2 とする．このときの条件は以下のように表現できる:

$$E(EC'_1[t]) > E(EC_o[t]) \tag{3.2}$$

分割のアルゴリズムを Algorithm2 に示す．

Algorithm 2 分割

Require: equivalence classes **EC**$[t]$, suppressed trajectories **SUP**

$M \leftarrow \emptyset$
for all $EC_o[t] \in$ **EC**$[t]$ **do**
  if $S(EC_o[t]) > \sigma$ **then**
    $\{EC'_a[t], EC'_b[t]\} \leftarrow$partitioning($EC_o[t]$)
    if $E(EC'_a[t]) > E(EC_o[t])$ OR $E(EC'_b[t]) > E(EC_o[t])$ **then**
      Replace $EC_o[t]$ with $EC'_a[t]$ and $EC'_b[t]$.
      if $EC'_a[t]$ or $EC'_b[t]$ lose k-anonymity **then**
        Reassign tid for members of $EC'_a[t]$ and $EC'_b[t]$.
      end if
      for all $EC'_i[t] \in \{EC'_a[t], EC'_b[t]\}$ **do**
        if $EC'_i[t] < k$ **then**
          SUP$\leftarrow$**SUP**$\cup EC'_i[t]$
        end if
      end for
    else
      $M \leftarrow M \cup EC_o[t]$
    end if
  end if
end for

併合

併合では，分割が不可能な EC を複数併合して 1 つの EC を作る．併合することで，再び分割できるようにすることが併合の目的である．

むやみに EC を併合すると，分割したとしても解像度の高い EC が得られず，再び併合が必要になる．そこで，EC のペアに対してクラスエネルギーが向上できるか否かを判定して，向上が可能なペアを併合する．

ここで EC_a と EC_b を併合対象の EC とする．また，EC'_m を EC_a と EC_b を併合した EC とする．EC_a と EC_b を併合した際には，互いにクラスエネルギーが増加することが望まれる．例えば，離れた位置に存在する EC 同士を併合した場合には，エリ

Algorithm 3 併合

Require: equivalence classes **EC**$[t]$, suppressed trajectories **SUP**, M

1: **for all** $EC_c[t] \in M$ **do**
2: **if** $EC_c[t] < 2k$ **then**
3: $N \leftarrow$ neighbors of $EC_c[t]$
4: Find $bestpartner$ from $N \cup$**SUP**.
5: $EC'_m[t] \leftarrow$merging($EC_c[t], bestpartner$)
6: Replace $EC_c[t]$ and $bestpartner$ with $EC'_m[t]$.
7: **end if**
8: **end for**

アの面積が大きくなってしまいクラスエネルギーが併合前よりも小さくなってしまう．よって，$E(EC'_m[t]) > E(EC_a[t])$ と $E(EC'_m[t]) > E(EC_b[t])$ が満たされる EC_a と EC_b を併合する必要がある．また，併合の処理対象である EC すべてのクラスエネルギーを最大化する併合の組み合わせの生成は，計算コストが高いため，移動軌跡ストリームをリアルタイムに匿名化するという観点からヒューリスティックな方法で併合のペアを考える．

ここでは，ある EC を中心に隣接する他の EC と併合することを考える．併合の中心とする EC を EC_c で表す．また，EC_c から一定の範囲内に存在し，EC_c との併合の候補に成り得る EC を隣接クラス (neighbors) と呼ぶ．EC_c の隣接クラスの集合を $N(EC_c)$ とする．隣接クラスは，$E(EC'_m[t]) > E(EC_c[t])$ と $E(EC'_m[t]) > E(EC_l[t])$ を満たす EC_l である．この条件を満たし得る隣接クラスは，以下の補題を充足する．

補題 1 分割数が $p = 2$ のとき，EC_c の隣接クラスになり得るクラスは以下を満たす*:* $dw < w(\frac{3}{2(1+h^{-1}dh)} - 1)$ ここで w *(*h*)* は EC_c のエリアの幅 *(*高さ*)* であり, dw *(*dh*)* は EC_c と併合後のクラス EC'_m と EC_c の幅 *(*高さ*)* の変化量である．

証明 1 *(*補題 *1): *$EC_c[t]$ と $EC_n[t]$ の併合を考える．$EC'_m[t]$ を $EC_c[t]$ と $EC_n[t]$ を併合したクラスとする．$EC_c[t]$ が $EC_n[t]$ と併合して $EC'_m[t]$ を成すためには，$EC'_m[t]$ は

少なくとも $E(EC'_m[t]) > E(EC_c[t])$ を満たさなければならないため，$EC_c[t]$ と $EC_n[t]$ の間には以下が成り立たなければならない:

$$\frac{\sigma(1+\log_p \frac{|EC'_m[t]|}{k})}{S(EC'_m[t])} > \frac{\sigma(1+\log_p \frac{|EC_c[t]|}{k})}{S(EC_c[t])}, \tag{3.3}$$

$$S(EC'_m[t]) < \frac{1+\log_p(|EC_c[t]|+|EC_n[t]|)/k}{1+\log_p |EC_c[t]|/k} S(EC_c[t]). \tag{3.4}$$

ここで, $S(EC_c[t]) = wh$，$S(EC'_m[t]) = (w+dw)(h+dh)$ とすると，

$$(w+dw)(h+dh) < \frac{1+\log_p(|EC_c[t]|+|EC_n[t]|)/k}{1+\log_p |EC_c[t]|/k} wh, \tag{3.5}$$

$$dw < \frac{1}{h+dh}(\frac{1+\log_p(|EC_c[t]|+|EC_n[t]|)/k}{1+\log_p |EC_c[t]|/k} - (wh+wdh)). \tag{3.6}$$

となる．また，併合においては $|EC_c[t]| < 2k$，$|EC_n[t]| < 2k$ であり，本研究では $p=2$ を想定しているため，以下のようになる:

$$dw < w(\frac{3}{2(1+h^{-1}dh)} - 1). \tag{3.7}$$

□

隣接クラスの探索を効率化するために，グリッド状の索引を用いて EC を管理する．グリッド索引のセルのサイズは σ(1 辺が $\sigma^{1/2}$ の正方形) とする．

EC_c は $N(EC_c)$ の隣接クラスのうち，併合後のクラスエネルギーが最大になるクラス ($bestpartner$) と併合する．併合によって生成したクラスは併合処理の対象のクラス集合に追加し，他のクラスとの併合の対象とし，よりメンバ数が多く密なクラスの生成を図る．

すべてのクラスで併合を試みたあと，併合が行われなかったクラスはクラスを解体し，当該クラスに属していた移動軌跡ストリームを秘匿状態にする．

分割のアルゴリズムを Algorithm3 に示す．

表 3.3 では，時刻 t_5 に EC_1 と EC_3 が併合されて，EC_5 が生成されている (図 3.5(b))．

境界の曖昧化

この節では CMOA を拡張する手法の 1 つを紹介する．CMOA では最小包囲矩形で位置情報を汎化したエリアを生成している．EC のメンバ数が比較的少数なとき，1 つの移動軌跡ストリームの移動が最小包囲矩形の形状や面積に大きな影響を与えることがあり，そのような情報から個人が特定されてしまう可能性がある．これは単に k-匿名化の問題とは異なるプライバシ侵害の問題である．

このプライバシ侵害を緩和する一手段として，最小包囲矩形として生成したエリアにノイズを加えて，境界を曖昧化する方法がある．また，エリアを常に一定の大きさ (例えば σ) にする方法等が考えられる．

3.4 評価

CMOA によって生成された汎化移動軌跡ストリームの有用性 (Utility) を実験によって測定し，評価する．また，クラスエネルギーの妥当性について，他の指標と比較することで検証を行う．

有用性の評価として，匿名化によって生成した汎化移動軌跡ストリームの解像度とトレーサビリティを評価する．さらに，実行時間を計測し，計算効率やリアルタイム処理に適用可能かについても検証する．解像度およびトレーサビリティの評価は独自の指標を定義して評価する．

解像度の評価では，既存の静的な匿名化手法との比較を行う．比較対象の手法として，Abul らが提案する *NWA*[1] と Nergiz らの提案手法 [57](ここでは *Nergiz* と呼ぶ) を用いた．両手法ともに，始点から終点までの移動軌跡の全容が与えられた状況において，k-匿名性を満たす匿名化移動軌跡を一度に生成する手法である．*NWA* は，半径 δ 以内のシリンダー状の匿名化移動軌跡を生成する手法である．δ は，CMOA と条件を合わせるために $\delta = |\sigma^{1/2}|$ とした．*Nergiz* らの手法は，実際には時刻についても汎化を行う 3 次元の汎化手法であるが，CMOA と条件を合わせるために，緯度，経度のみを匿名化の対象とした．また，*NWA* は公開されているソースコード [1] を用い，*Nergiz* は論文を基に実装した．

3.4.1 実験環境

データセット

本評価では，人工データセットである PFLOW[14] を用いた．PFLOW は関東地方の約 72 万人分の移動軌跡を含んだデータセットであり，被験者に対する一日の滞在先，移動方法に関するアンケート調査結果から，滞在先間の移動を補完した移動軌跡を Nakamura らの手法 [55] によって生成されている．よって，PFLOW データセットの移動軌跡は実際の人々の移動パターンが反映されたものであると言える．

本評価では，10 万人分の移動軌跡をランダムに抽出し，その中から 2 時間分のデータを用いた．インターバルは 1 分とし，一つの移動軌跡には 120 個の位置情報を含む．

PFLOW は静的なデータセットであるため，毎分位置情報を送信するシミュレータとして移動軌跡ストリームシミュレータを開発し，移動軌跡ストリームを生成した．移動軌跡ストリームシミュレータは，毎分，すべての移動軌跡の位置情報を送信する．

計算機環境

評価実験は仮想マシン上で実施した．仮想マシンには，4 コアの CPU と 32GB のメモリ，100GB の外部記憶 (HDD) を割り当てた．OS は CentOS 5.6 を用いた．なお，ホストサーバーは 2.4GHz の 12 コア CPU を持ち，196GB のメモリ，6TB の HDD を持つ．実行時間の計測は，他の仮想マシンを動作させていない状況で行った．

CMOA は Java 1.6.0_17 で開発した．また, 開発した CMOA では移動軌跡ストリームシミュレータから送信される位置情報を，キーバリューストア型 DBMS である Apache Cassandra 0.6.6 [1](以降，Cassandra) に格納し，Cassandra から特定のタイムスタンプの位置情報群を取得して匿名化を行う．

3.4.2 評価指標

本評価では，汎化移動軌跡ストリームの解像度，tid の継続時間，tid の変更回数，CMOA の処理性能について評価を行う．tid の継続時間と変更回数は汎化軌跡ストリームのトレーサビリティを評価するために用いる．

本節では，評価実験のために新たに導入した評価指標の定義や性質について述べる．特に, 対象としたデータセット全体の位置情報の解像度を測る指標 RM(Resolution Metric) と，トレーサビリティを測る指標 MD(Maximum Duration) を導入する．

[1] http://cassandra.apache.org/

解像度

あるタイムスタンプにおけるデータセット全体の位置情報の解像度を評価する指標として解像度指標 (Resolution Metric, RM) を導入する．

$$RM[t] = \sum_{\tau^* \in D^*} S(\tau^*[t])^{-1/2}. \tag{3.8}$$

RM は，EC 毎にエリアの対角線あたりの移動体 (移動軌跡ストリーム) の数を各タイムスタンプで算出し，タイムスタンプ毎にその総和を計算した値である．汎化された度合いが小さいほど，RM の値は高くなり，オリジナルの位置情報に近い精度を持っていることを表す．RM=1 のとき，すべての位置情報がオリジナルの位置情報と等価であることを表す．時刻 t において秘匿状態の移動軌跡ストリームは RM 値に計上しないため，秘匿状態の移動軌跡ストリームが多いほど RM は小さくなる．

最大継続時間

トレーサビリティの評価尺度として最大継続時間 (Maximum Duration, MD) を導入する．データ主体ごとに各 tid を継続した時間を計測し，そのうち継続した時間が最大の値をデータ主体 u の MD 値とする．

$Dur(tid)$ を tid を継続した時間とする．MD を以下の式で定義する:

$$MD(u) = max_{tid \in TID(u)} Dur(tid) \tag{3.9}$$

ここで $TID(u)$ はデータ主体 u の移動軌跡ストリームが匿名化された際に付与された tid の集合である．

3.4.3 評価結果

本評価実験では，解像度，トレーサビリティ，実行時間について評価する．また，クラスエネルギーの妥当性検証のための評価結果についても述べる．

解像度

まず，位置情報の解像度を RM を用いて評価した結果を示す．図 3.7 と図 3.8 は，データセット全体の解像度を測る評価値である RM 値の $[t_{i-9}, t_i](i = 0, 10, 20, \ldots, 120)$ の区間での平均値の変化を示している．

図 3.7 は，提案手法 CMOA と，CMOA で動的再構成をしない場合 (*NoReform*) を比較した結果である．*NoReform* は動的再構成ができないため，初期分割で形成したクラスを維持し続ける．$k = 5$ で評価を行った．

図 3.7 では，*NoReform* の RM 値が単調減少していることがわかる．一方，CMOA は，RM 値の増減はあるがある一定以上の解像度を維持していることがわかる．よって，動的再構成が解像度の維持にある一定の効果を持っていると言える．

次に，CMOA と既存の静的な移動軌跡に対する k-匿名化手法である *NWA* と *Nergiz* との比較を行う．*NWA* は今回実施したデータセットの規模で匿名化を実施することができなかったため，*NWA* のみ 60 分間のデータセットに対して匿名化を行っている．いずれの手法も $k = 5$ で評価を行った．

図 3.8 は，静的手法である *Nergiz* と *NWA* は，CMOA よりも高い RM 値を示している．CMOA は未知な将来の移動に対して解像度を維持するために，面積が σ 以下の場合はクラスを小さくしない．そのため，RM 値が静的手法には及ばなかったと考えられる．しかしながら，$t_i = 20$ では，CMOA($\sigma = 250 \times 250[m^2]$) が静的手法の RM 値に漸近しており，CMOA の匿名化結果は必ずしも低い解像度ではないと言える．

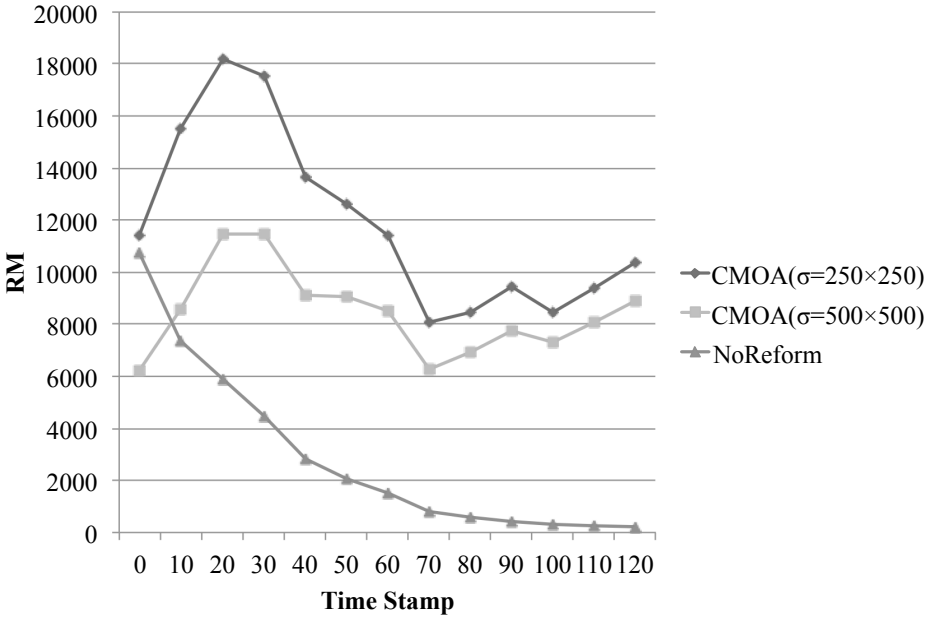


図 3.7: RM(ナイーブ手法との比較)

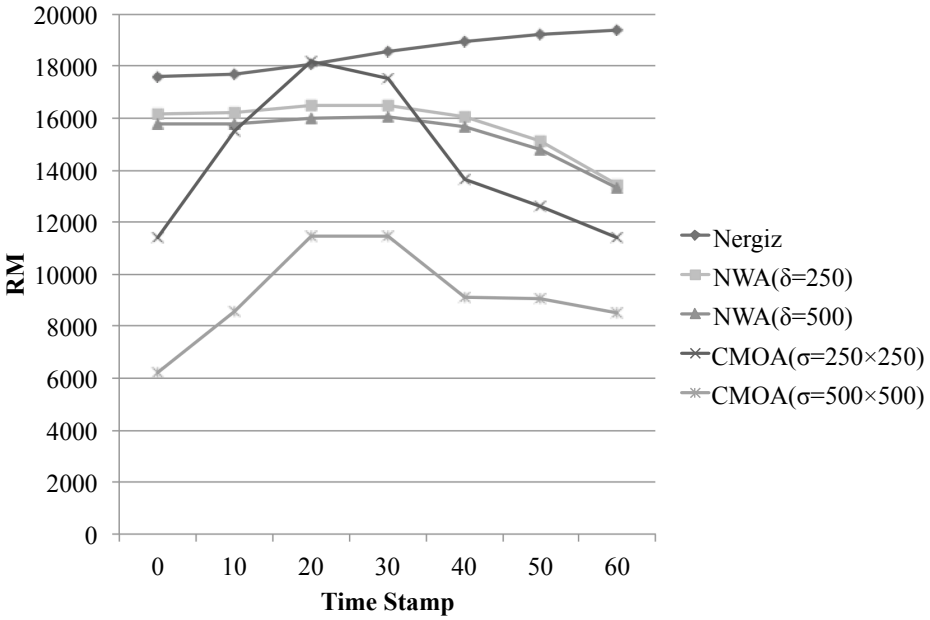


図 3.8: RM(静的手法との比較)

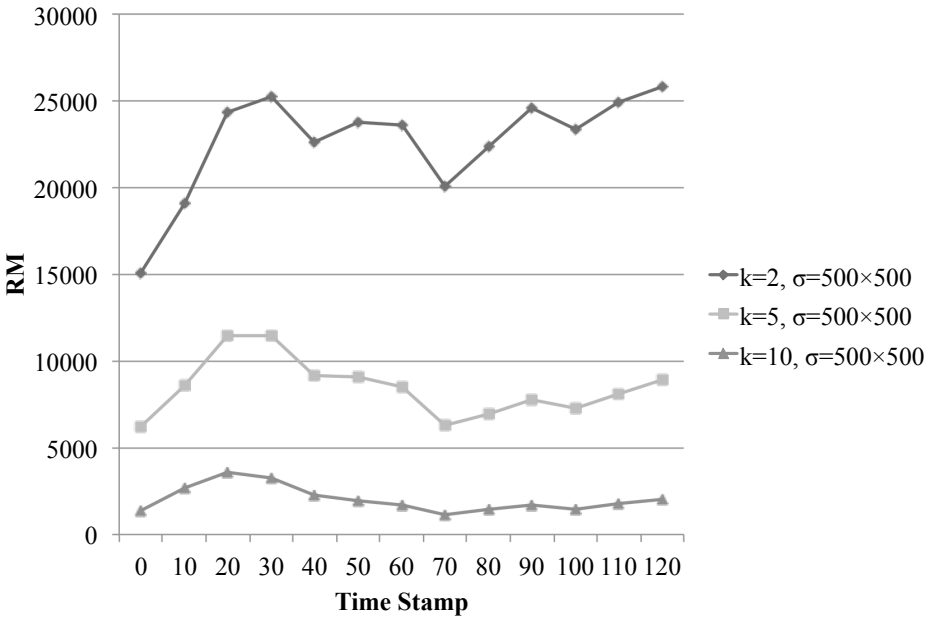


図 3.9: RM (k を変更)

図 3.9 は k を変更した際の匿名化結果の変化を示している．k と σ はそれぞれ k=2, 5, 10, $\sigma = 500 \times 500[m^2]$ とした．また，図 3.10 は図 3.9 と同じ設定において，クラスエネルギーの平均値の変化を示している．図 3.10 では，クラスエネルギーの増減が見てとれる．クラスエネルギーが減少している点では，解像度が増加している．$k = 2$ のときは，十分なクラスエネルギーが確保されているため，高い解像度を維持できている．k が大きくなるにつれて，k-匿名性の充足に多くのメンバが必要なため，クラスエネルギーが不足しやすいこともわかる．

以上より, 提案手法 CMOA が解像度をある一定レベルに維持できることがわかった.

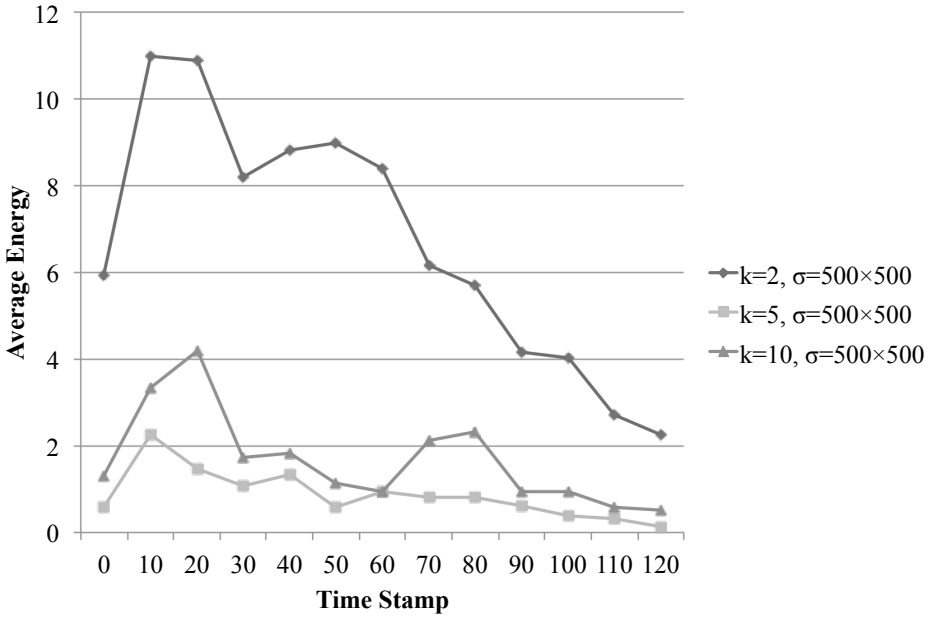


図 3.10: 平均クラスエネルギー

トレーサビリティ

図 3.11 は $k = 2, 5, 10$，$\sigma = 500 \times 500[m^2]$ で匿名化した際の各 MD 値を示している．図 3.12 は $k = 5$，$\sigma = 250 \times 250, 500 \times 500, 1000 \times 1000[m^2]$ で匿名化した際の各 MD 値を示している．横軸が MD の値の範囲であり，縦軸がデータ主体 (移動軌跡ストリーム) の割合を示している．MD は k の値が大きいと，小さくなることがわかる．これは，k の値が大きいほど，k-匿名性の充足に多くの移動軌跡ストリームが必要となり，また分割可能な回数も少なくなってしまうためと考えられる．よって，k の値が大きいほどトレーサビリティは損失しやすい．一方，閾値 σ が大きいほど，解像度に余裕が生まれるためトレーサビリティは大きくなり，長時間トレース可能な汎化移動軌跡ストリームを提供できる．

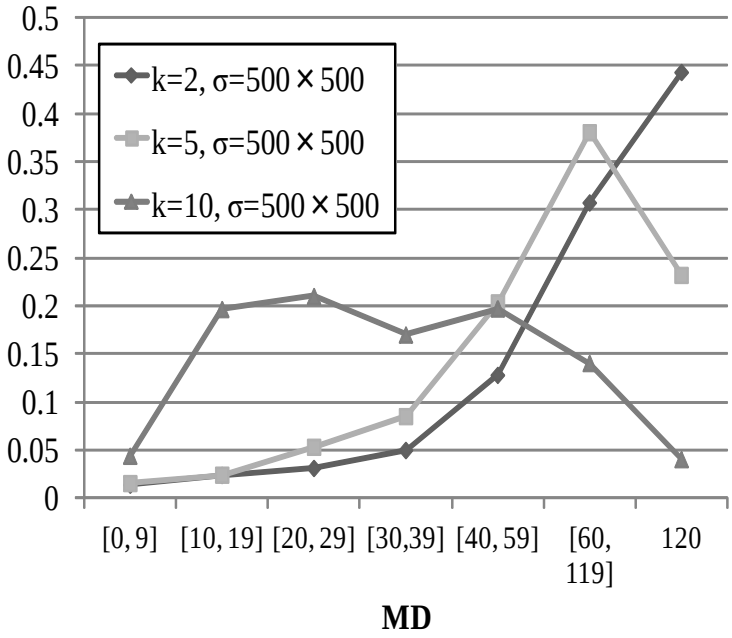


図 3.11: MD(k を変更)

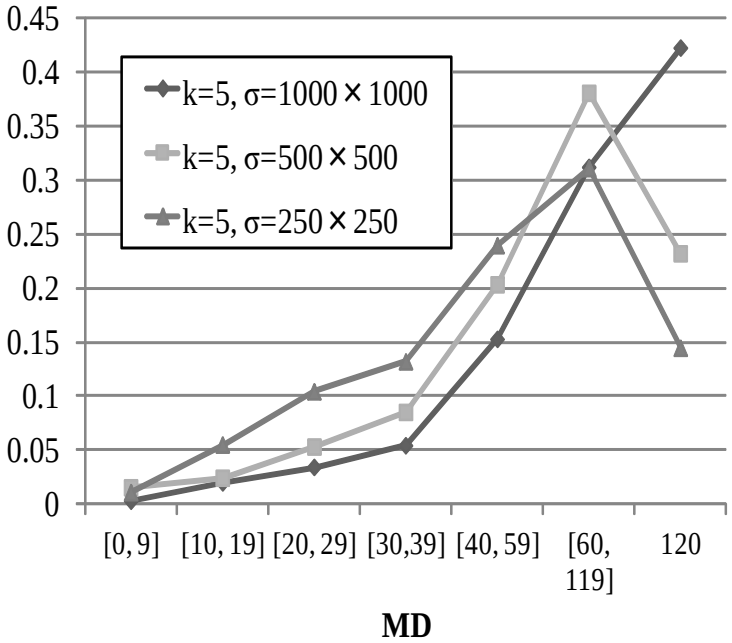


図 3.12: MD(σ を変更)

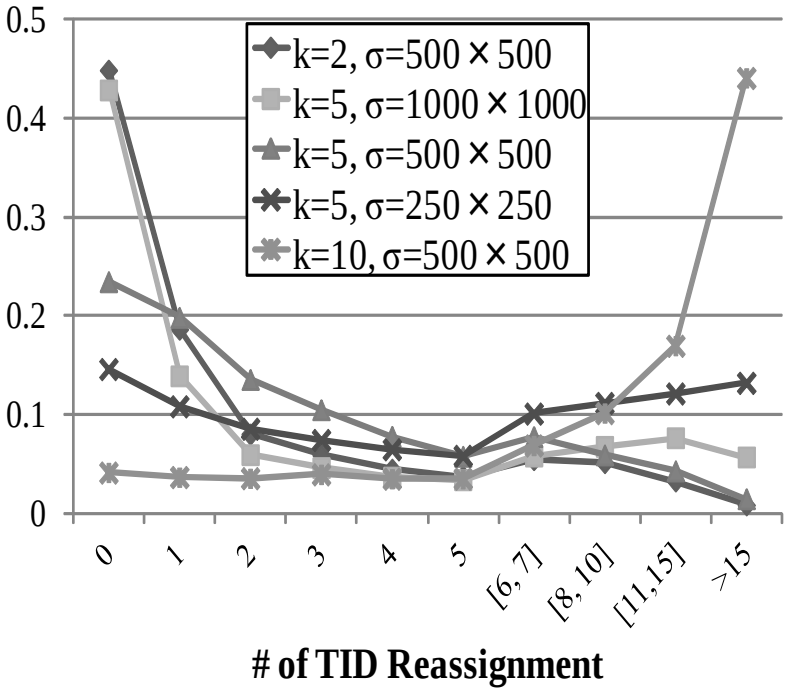


図 3.13: tid 再割当の発生回数

次に，汎化移動軌跡の識別子 tid の再割当の発生頻度について評価を行った．図 3.13 は tid の再割当の発生頻度を $k = 2, 5, 10$ の場合に比較した結果を示している．$k = 2$ の場合は約半数の移動軌跡ストリームは高々 1 回の tid の再割当回数である．一方で，$k = 5, 10$ の場合は複数回の再割当が生じている．特に，$k = 5, \sigma = 250 \times 250$ は，10 回以上の再割当が生じたものが半数以上である．このことから，k と σ は慎重に決定することが必要なパラメータであることがわかる．

クラスエネルギーの妥当性検証

本節では，クラスエネルギーを基にした併合の妥当性を検証する．CMOA では，クラスエネルギー (式 (3.1)) によって動的再構成を行う．特に併合では，クラスエネルギーが高いクラスを成すことによって，TID の再割当が生じづらく，トレーサビリティが高いクラスを目指している．

クラスエネルギーを基にした併合の妥当性を検証するため，クラスエネルギー以外の指標によって併合を行った場合と，TID の変更回数と，秘匿状態の移動軌跡数を比較する．併合の指標として，クラスエネルギーと同様に，併合後のクラスのメンバ数と面積によって評価する 2 つの指標を用いた．1 つがクラスの密度 (Density) であり，以下の式で表す．

$$E_{Density}(EC[t]) = \frac{|EC[t]|}{S(EC[t])} \tag{3.10}$$

2 つ目の指標がクラスのメンバ数 (Cardinality) であり，以下の式で表す．

$$E_{Cardinality}(EC[t]) = |EC[t]| \tag{3.11}$$

クラスエネルギー (Class Energy)，密度 (Density)，メンバ数 (Cardinality) を用いた CMOA を，$k = 5$，$\sigma = 500 \times 500[m^2]$ の設定で，10 万件の移動軌跡ストリームに対して匿名化を行った．

図 3.14 は，TID の再割当によって 1 回以上 TID を変更した移動軌跡に対して，TID の変更回数が $m(\geq 1)$ 回以下の移動軌跡の割合を示している．図 3.14 では，面積の小ささを考慮していないメンバ数 (Cardinality) による手法よりも，クラスエネルギーや密度を考慮した手法の方が TID の変更回数が少ない移動軌跡ストリームの割合が多い．また，密度に基づく手法が最も TID の変更回数が少ないことがわかる．

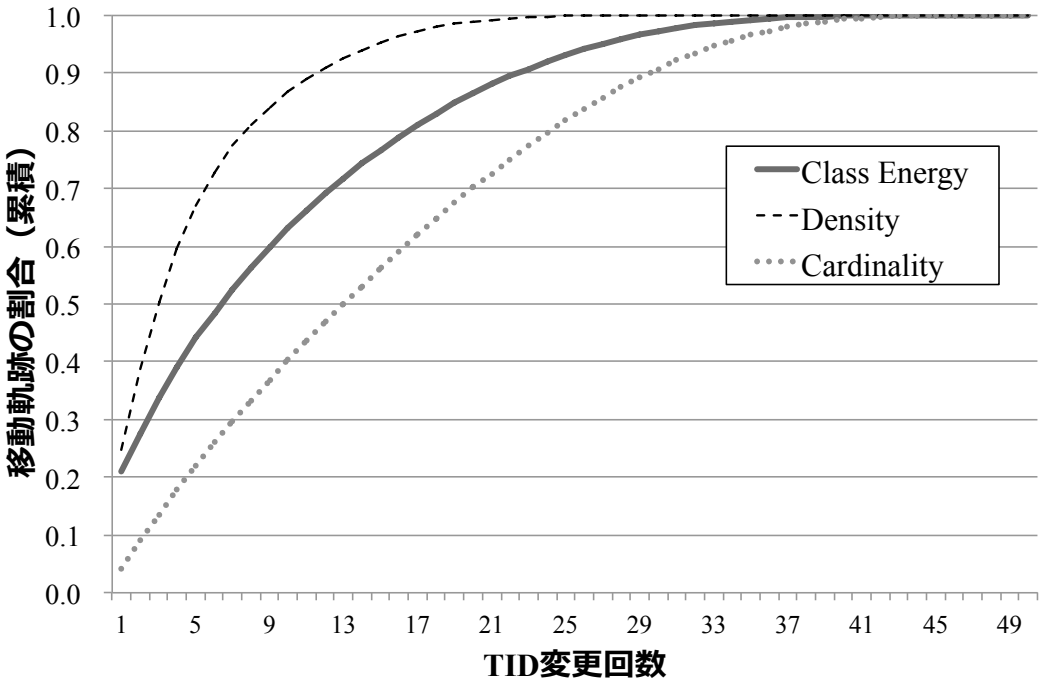


図 3.14: TID の変更回数

図 3.15 は，外れ値であり秘匿状態となった移動軌跡ストリームの割合の，時間変化を示している．図 3.15 では，密度による手法の秘匿状態の移動軌跡ストリームの割合が，単調増加していることがわかり，最終的には他の手法と比較して 2 倍以上の移動軌跡ストリームが秘匿状態になっていることがわかる．秘匿状態の移動軌跡ストリームは，k-匿名性を満たしていないため，出版することができない．よって，秘匿状態の移動軌跡ストリーム数が単調増加していると，より長い時間匿名化を行った場合に，出版可能な移動軌跡ストリームが僅かになってしまう可能性がある．併合時の密度によるクラスの評価は，メンバ数の影響度がクラスエネルギーよりも大きい．そのため，小規模なクラスは併合の対象に成りづらく，秘匿状態のクラスが増加しているのではないかと考えられる．

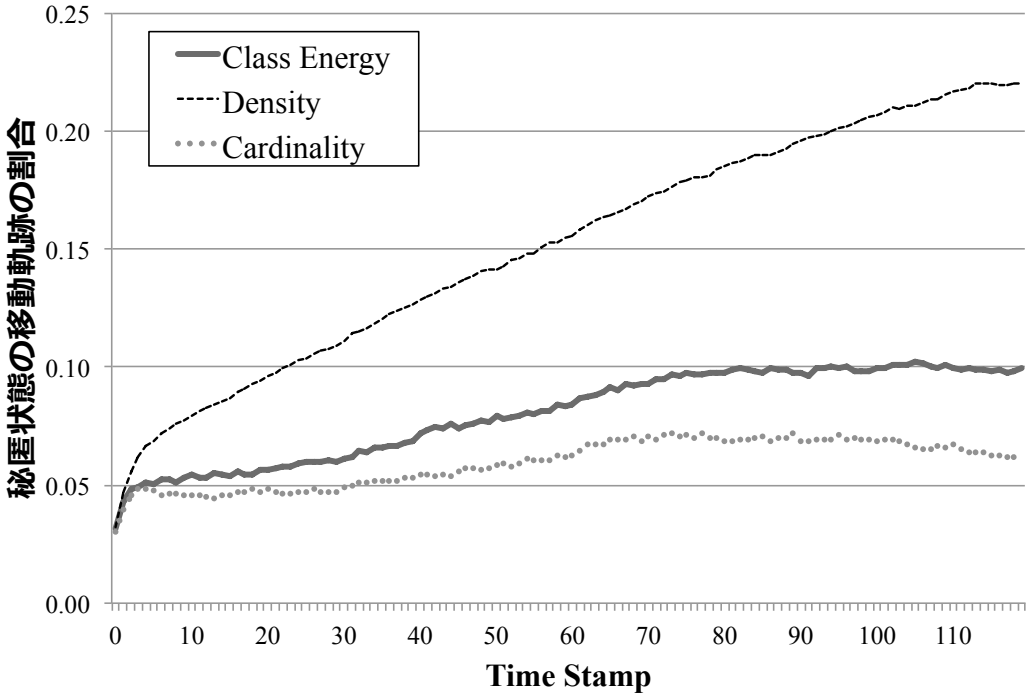


図 3.15: 秘匿状態の移動軌跡数

以上より，クラスエネルギーを基にすることで秘匿状態の移動軌跡ストリーム数, TID の変更回数を共に少なくできることが分かり，併合の指標としてクラスエネルギーが一定の優位性を持っていることが分かった.

処理性能

最後に，提案手法の計算効率を評価するために実行時間の計測を行った．図 3.16, 図 3.17, 図 3.18 はそれぞれの設定における平均実行時間を示している.

まず，CMOA の対象とするデータセットの規模毎に実行時間を計測し，提案手法のスケーラビリティを評価する．図 3.16 は，1 万，2 万，4 万，6 万，8 万，10 万件で匿名化した際の実行時間を示している．実行時間を初期分割 (Initial P.)，分割 (Partition-

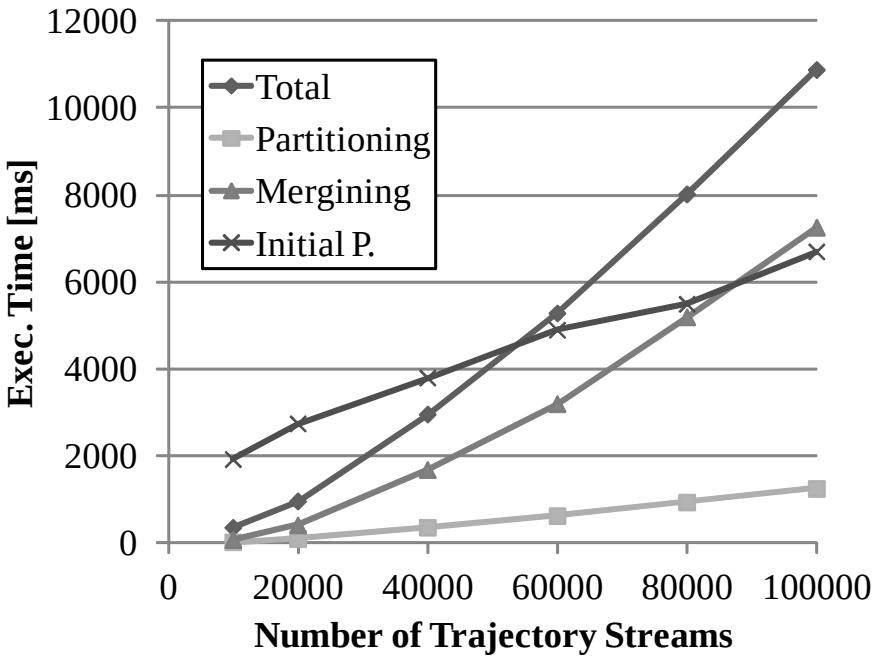


図 3.16: 実行時間 ($|D|$ を変更)

ing)，併合 (Merging)，トータル (Total) に分けて示す．Initial P. は t_0 の実行時間である．Partitioning, Merging, Total は時刻 t_1 以降の平均値である．Total には Partitioning, Merging に加えて，1 つ前の時刻の EC を参照して EC を生成する時間や，データの読み書きに要する時間などが含まれる．トータルの実行時間は，データの規模の増加に合わせてほぼ線形に近い増加を見せている．他の実行時間の増加傾向も線形に近い．よって，本評価では 10 万件までの評価であったが，より大規模なデータセットであっても計算リソースの増強によって対応することが可能であると考えられる．また，10 万件を約 10 秒で匿名化できていることから，今回用いた 1 分間のインターバルで生成される移動軌跡ストリームをリアルタイムに匿名化できることが確認できた．

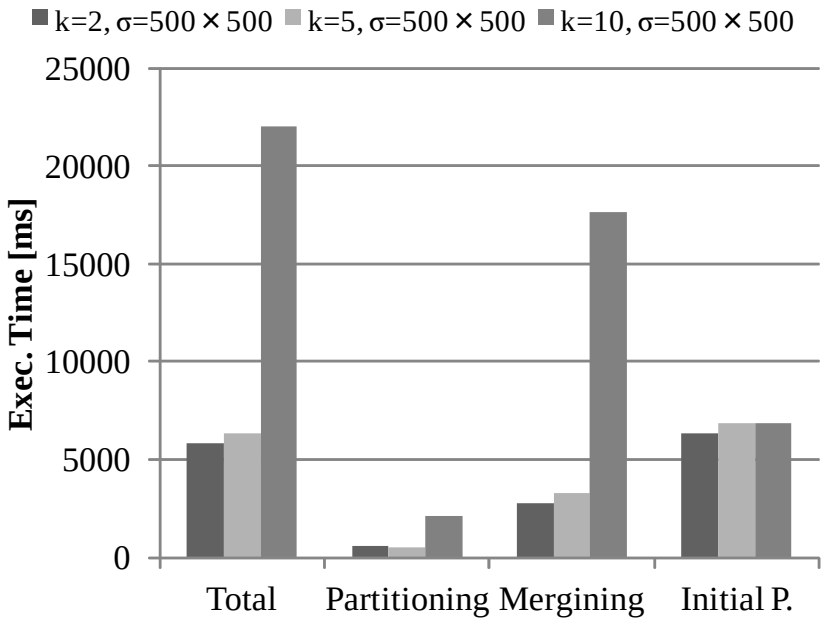


図 3.17: 実行時間 (k を変更)

図 3.17 は k=2, 5, 10，$\sigma = 500 \times 500[m^2]$ で匿名化を行い，k を変更した際の実行時間の変化を示している. 図 3.18 は k=5，$\sigma = 250 \times 250, 500 \times 500, 1000 \times 1000[m^2]$ で匿名化を行い，σ を変更した際の実行時間の変化を示している. $k = 10$ の際には，併合に多くの時間を要していることがわかる．$k = 10$ の際には，$\sigma = 500 \times 500[m^2]$ というエリアの面積の閾値がタイトであり，多くの EC で併合が必要となったためと考えられる．$k = 2, 5$ では併合に要した時間に大きな差はないため，σ の設定が $k = 10$ の場合には適切でなかったと考えられる．図 3.18 では，σ の値が大きくなり，閾値の設定が緩くなっていくにつれて併合に要する時間も減少している．よって，所定の時間内に匿名化を行うためには，適切なパラメータの設定が必要であると考えられる.

以上より，10 万件程度の移動軌跡ストリームを毎分リアルタイムに匿名化可能であることを確認できた．しかしながら，適切な設定の指針については検討が必要である.

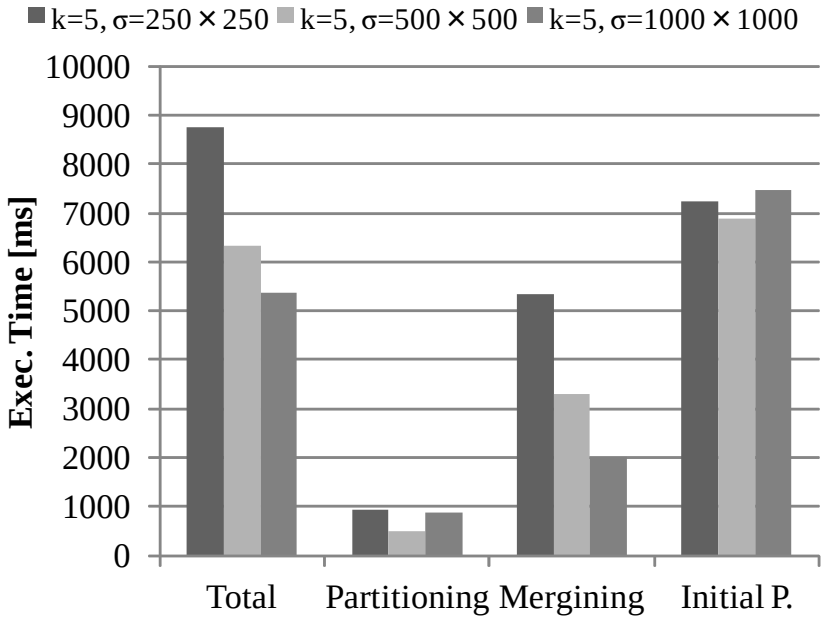


図 3.18: 実行時間 (σ を変更)

3.5 まとめ

本研究では，移動軌跡ストリームを連続的かつリアルタイムにプライバシ保護出版する問題に取り組み，連続的 k-匿名化手法 CMOA を提案した．

既存の移動軌跡に対する匿名化手法では，蓄積された移動軌跡に対する静的な匿名化を対象としていたため，時々刻々と蓄積していく移動軌跡を連続的に匿名化することができなかった．提案手法 CMOA は，位置情報の解像度を一定以上に保ち，また可能な限り位置情報の系列を維持しながら連続的な匿名化を実現した．

評価実験では，CMOA が一定の解像度を維持したまま，リアルタイムに k-匿名化できることを示した．静的な手法との解像度の比較では，静的な手法よりも解像度が劣ることがわかった．これは，移動軌跡の全容を把握した上で匿名化を行う静的な手法とは異なり，過去の軌跡と現在の位置のみを利用していること，また，将来の移動

が未知であることから動的再構成の余力を持たせるために，曖昧化された面積が一定以下の場合はより面積の小さい (解像度が高い) 構成へと動的再構成を行わない方針を取っているためである．加えて，CMOA が導入した動的再構成では，解像度を維持するために移動軌跡ストリームのトレーサビリティが損失してしまう場合がある．解像度 (エリアの面積) の閾値がタイトな場合，このトレーサビリティの損失が頻繁に生じることもわかった．本研究では，解像度を保証すべき指標として考慮したが，トレース可能な時間を重視するユースケースでは，トレース可能な時間を保証する手法も必要であると考えられる．評価実験では，10 万人の移動軌跡ストリームを毎分匿名化するという条件において，10 秒程度で匿名化できることを確認した．また，処理時間がほぼ線形に近い推移を見せていることから，より大規模な移動軌跡ストリームを対象とする場合でも，計算リソースの増強で対応できると考えられる．

第4章　センシティブ属性間の関係多様化

4.1　はじめに

診療履歴やサイト訪問履歴といったパーソナルデータが，サービスを受ける度に蓄積されている．近年，ビッグデータ活用のニーズが高まり，これらの蓄積されたパーソナルデータを第三者のサービスや事業に活用する二次活用の期待が高まっている．

センシティブ属性の系列を持ったパーソナルデータからは，属性間の相関や変化を観察することができる．例えば，表 4.1 は傷病と薬剤を記録したテーブルであり，傷病と薬剤との相関が得られる．表 4.2 は 4 月，5 月，6 月の傷病の履歴を記録した時系列データであり，傷病の変化やその経過観察を行うことができる．

しかしながら，1 章で述べたように傷病や薬剤のようにデータ主体に関する機微な情報（センシティブ属性）は，第三者に知られたくない情報であるため，二次活用の際にはデータ主体のプライバシへの配慮が必要となる．

複数のセンシティブ属性を含むパーソナルデータでは，特定のデータ主体に関するあるセンシティブ属性に関する知識から他のセンシティブ属性値が特定されるプライバシ侵害が生じ得る．このプライバシ侵害を防ぐためには，あるセンシティブ属性から他のセンシティブ属性が一意に対応付かないように，センシティブ属性間の対応関係が多様 (多対多) になることを保証する必要がある．

表 4.1: 系列データの例

ID	傷病名	薬剤名
1	a	x
2	a	y
3	b	x
4	b	y
5	a	x
6	a	z
7	c	z
8	c	w

表 4.2: 系列データの例 2

ID	傷病名 (4 月)	傷病名 (5 月)	傷病名 (6 月)
1	a	x	x
2	a	y	x
3	b	x	e
4	b	y	f
5	a	x	e
6	a	z	w
7	c	x	w
8	c	w	e

属性間の対応が多対多になるように属性値を汎化した系列データの例を表 4.3 と表 4.4 に示す．表 4.3 と表 4.4 における「a/b」という表記は，属性値が a と b のいずれかであることを示している．表 4.3 は表 4.1 のセンシティブ属性値を汎化して，属性間の対応を多様化している．同様に表 4.4 は表 4.2 のセンシティブ属性値を汎化して，属性間の対応を多様化している．

表 4.3 と表 4.4 の生成に用いた汎化は，既存の多くのデータ匿名化技術で用いられてきた操作である．しかしながら，属性間の対応関係の多様化を実現するためには，表 4.3 と表 4.4 のように属性値の大きな汎化が必要になり，有用性が大きくて低下し

表 4.3: 匿名化した系列データの例

傷病名	薬剤名
a / b	x / y
a / b	x / y
a / b	x / y
a / b	x / y
a / c	x / z
a / c	x / z
a / c	x / w
a / c	x / w

表 4.4: 匿名化した系列データの例 2

傷病名 (4 月)	傷病名 (5 月)	傷病名 (6 月)
a / b	x / y	x / e
a / b	x / y	x / e
a / b	x / y	x / f
a / b	x / y	x / f
a / c	w / x / z	e / w
a / c	w / x / z	e / w
a / c	w / x / z	e / w
a / c	w / x / z	e / w

てしまう問題がある．

そこで本研究では，新たに系列データを分割して，センシティブ属性間の二項関係へと変換する関係多様化という操作を導入する．関係多様化では，センシティブ属性間の対応関係を曖昧にすることで，属性間の対応が多対多であること (関係多様性) を保証する．このとき，センシティブ属性を汎化せずに関係多様性を保証することができるという利点を持つ．

表 4.5(a)，4.5(b) は，表 4.1 を関係多様化した一例を示している．複数のレコードが共通の識別子 CID を持ち，CID と CID' によって属性値間の関係が多対多に多様化

表 4.5: 関係多様化データ

(a) 傷病名

CID	CID'	傷病名
11	21	a
11	21	b
12	22	a
12	22	b
13	23	a
13	23	c
14	24	a
14	24	c

(b) 薬剤

CID	CID'	傷病名
21	—	x
21	—	y
22	—	x
22	—	y
23	—	x
23	—	w
24	—	x
24	—	z

表 4.6: 関係多様化データ

(a) 4 月

CID	CID'	傷病名
11	21	a
11	21	b
12	22	a
12	22	b
13	23	a
13	23	c
14	24	a
14	24	c

(b) 5 月

CID	CID'	傷病名
21	31	x
21	32	y
22	32	x
22	31	y
23	33	x
23	34	w
24	34	x
24	33	z

(c) 6 月

CID	CID'	傷病名
31	—	x
31	—	f
32	—	x
32	—	e
33	—	e
33	—	w
34	—	e
34	—	w

されていることがわかる．同様に，表 4.6(a)，4.6(b)，4.6(c) は表 4.1 を関係多様化した一例である．いずれの例も，センシティブ属性値が元の状態に保たれている．

しかしながら，関係多様化はセンシティブ属性間の対応関係に曖昧性を生じさせるため，分析精度を劣化させる可能性がある．本研究では，2 つのセンシティブ属性を持つパーソナルデータを対象として，関係多様性を関係の曖昧化の度合いを抑制しながら実現するデータ匿名化に取り組む．まず，センシティブ属性間の関係多様性である

(ℓ_1, ℓ_2)-関係多様性を提案する．さらに，関係の曖昧化を抑制しつつ効率的に (ℓ_1, ℓ_2)-関係多様化を実現する手法を提案する．

評価実験では，提案手法が関係の曖昧性を抑止しつつ効率よく (ℓ_1, ℓ_2)-関係多様化を実現でき，ナイーブな手法と比較して大幅に関係の曖昧性を抑止でき，10 倍から 100 倍の高い効率性を有していることを確認した．また，データ分析時にデータの操作に工夫を行うことで，誤差の小さいデータ分析が実現できることを確認した．

本章の貢献は，以下の 3 点である．

- 系列データに対する関係多様化の導入
- 関係多様化データの関係多様性指標として (ℓ_1, ℓ_2)-関係多様性の定義
- 関係多様化に伴いセンシティブ属性間の関係が曖昧化される問題に対して，関係の曖昧性を抑止可能な関係多様化手法を高い効率性を有しながら実現する手法の実現

これによって，系列データを活用する際に問題となる属性値の過度な汎化，もしくは属性間の関係の過度な曖昧化の両方を抑止したデータ匿名化が実現される．

本稿の以降の構成は以下の通りである．4.2 節では，本稿の論述に必要な基本的事項の導入を行う．4.3 節では，センシティブ属性間の関係多様化と関係多様性の指標である (ℓ_1, ℓ_2)-関係多様性を提案する．4.4 節では，(ℓ_1, ℓ_2)-関係多様化の実現手段の一例としてクラスタリングによる手法を述べる．4.5 節では，(ℓ_1, ℓ_2)-関係多様化を効率よく実現するための考察を行う．4.6 節では，4.5 節の議論に基づいたヒューリスティクスを用いた効率的な (ℓ_1, ℓ_2)-関係多様化手法を提案する．4.7 節では，提案手法の有効性について評価実験を通して論じる．4.8 節では，関係多様化データの活用例とその方法，およびデータ分析時の精度について述べる．4.9 節では，関連研究を紹介し，最後に 4.10 節にて，本章の結論を述べる．

4.2 準備

本節では，本稿で対象とするデータやプライバシモデルなど，本章の理解，論述に必要な基本的な事項について述べる．

4.2.1 対象とするデータ

対象とするデータは，$S_1, \ldots, S_d\,(d > 1)$ のセンシティブ属性を持つ $T = \{t_1, \ldots, t_n\}$ のテーブルである．タプル t のセンシティブ属性 S_i の値を $t[S_i]$ で表す．

4.2.2 攻撃モデルと保護モデル

複数のセンシティブ属性を持つ系列データに対しては，ある属性に関する知識から他の属性の属性値が特定されてしまう場合がある．

定義 10 *(攻撃者の知識):* 攻撃者は, $X \in S$ を攻撃対象の属性とし, $Y_1, \ldots, Y_{d-1} \in S \backslash X$ の属性値を既知とする．このとき，X を攻撃対象属性, $(y_1, \ldots, y_{d-1}) \in Y_1 \times \cdots \times Y_{d-1}$ を攻撃者の知識とする．ただし，攻撃者がセンシティブ属性の集合 S のうち，どのセンシティブ属性を攻撃対象 X としているかを事前に知ることはできない．

攻撃者は, $(y_1, \ldots, y_{d-1}) \in Y_1 \times \cdots \times Y_{d-1}$ に該当するタプルの集合 $T(y_1, \ldots, y_{d-1}) = \{t | t[Y_i] = y_i \wedge \cdots \wedge t[Y_{d-1}] = y_{d-1}\}$ に含まれる属性 X の値を特定しようとする．このとき, $(y_1, \ldots, y_{d-1})$ に該当するタプル集合の X の値の種類数は, $|\{t[X] | t \in T(y_1, \ldots, y_{d-1})\}|$ である．そこで，任意の属性の組から特定可能な X の属性値を複数 (ℓ_X) 種類になること (関係多様性) を保証することが求められる．

定義 11 *(関係多様性):* ある $X \in S, Y = S \backslash X$ について, $|\{t[X] | t \in T(y_1, \ldots, y_{d-1})\}| \geq \ell_X$ であるとき，X は ℓ_X-関係多様性を満たすとする．

定義 11 の関係多様性を保証する方法の一つに，センシティブ属性値の汎化がある．すべての属性 $X \in S$ が関係多様性を満たすためには，表 4.3 や表 4.4 のように，各センシティブ属性値を大きく汎化する必要がある．

4.3 関係多様化によるプライバシ保護

本節では，センシティブ属性間の関係多様化と関係多様性の指標である (ℓ_1, ℓ_2)-関係多様性を提案し，それぞれの定義について述べる．また，関係多様化によるセンシティブ属性間の関係の曖昧化に関して，関係の曖昧性の評価指標を導入する．

4.3.1 関係多様化

定義 11 の関係多様性を達成するための操作として，テーブル T をセンシティブ属性毎に分割する関係多様化を導入する．関係多様化は，センシティブ属性の系列 $S = \{S_1, \ldots, S_d\}$ を持ったテーブル T を，センシティブ属性として $S_i \in S$ のみを持ったテーブル T_i^* へと分割し，複数のレコードに共通のクラス識別子 CID を T_i^* と T_{i+1}^* に付与した新たなデータ形式へと変換する操作である．

定義 12 *(*関係多様化*):*

関係多様化 f_{rd} は，$f_{rd}(T) = \{T_1^*, \ldots, T_d^*\}$ のように，テーブル T をセンシティブ属性毎のテーブル T_i^* に分割する．T_i^* は CID, CID', S_i を持つテーブル $T_i^* = \{t_{i,1}^*, \ldots, t_{i,n}^*\}$ である．T_i^* の CID' には T_{i+1}^* の CID の値が入り $(t_{i,j}^*[CID'] \in T_{i+1}^*.CID)$, T_i^* の CID' と T_{i+1}^* の CID によって，S_i と S_{i+1} の二項関係が表される．また，$t_{i,j}^*[S]$ を $t_{i,j}^*$ のセンシティブ属性値とする．

複数の $t^*_{i,j}$ が同じ CID を持つことによって，T^*_i の CID' と T^*_{i+1} の CID の結びつきが弱くなり，センシティブ属性間の関係が多様化される．

定義 13 *(関係多様化クラス):* 関係多様化されたテーブル T^*_i において同じ $cid \in CID$ を持つ $t^*_{i,j}$ の集合を関係多様化クラス $c_{cid} = \{t^*_{i,j} | t^*_{i,j}[CID] = cid\}$ とする．また, T^*_{i-1} のセンシティブ属性値の集合を $c_{cid}[S_{pre}] = \{t^*_{i-1,j}[S] | t^*_{i-1,j}[CID'] = cid\}$ とし, c_{cid} の**前提部**と呼ぶ．同様に，T^*_i のセンシティブ属性値の集合を $c_{cid}[S_{con}] = \{t^*_{i,j}[S] | t^*_{i,j}[CID] = cid\}$ とし，c_{cid} の**結論部**と呼ぶ．

定義 14 *(オリジナルの関係集合):* c_{cid} に属する $t^*_{i,j}$ の元のタプル t_j におけるセンシティブ属性 S_{i-1} と S_i の二項関係の集合を，オリジナルの関係集合とし，$R(c_{cid}) = \{(t_j[S_{i-1}], t_j[S_i]) | t_j \wedge t^*_{i,j}[CID] = cid\}$ とする．

4.3.2 提案指標：(ℓ_1,ℓ_2)-関係多様性

4.2.2 節で示した攻撃モデルによるプライバシ侵害を防ぐために，テーブルが満たすべき関係多様性の指標（4.3.2 節）と，関係多様化による関係の曖昧性指標（4.3.3 節）を導入する．

定義 15 *((ℓ_1,ℓ_2)-関係多様性):* $\forall cid \in T^*_i[CID]$ に対して，$|c_{cid}[S_{pre}]| \geq \ell_1$ および $|c_{cid}[S_{con}]| \geq \ell_2$ を満たすとき，T^*_i は *(ℓ_1,ℓ_2)*-関係多様性を満たす．

$\{T^*_1, \ldots, T^*_d\} \in f_{rd}(T)$ において，T^*_{i-1} と T^*_i が (ℓ_1, ℓ_2)-関係多様性を満たすとき，$Y = S \setminus S_i$ から $X = S_i$ について，S_i は ℓ_2-関係多様性を満たす．同様に，$Y = S \setminus S_{i-1}$ から $X = S_{i-1}$ について，S_{i-1} は ℓ_1-関係多様性を満たす．

4.3.3 関係の曖昧性指標

(ℓ_1,ℓ_2)-関係多様性を保証する (ℓ_1,ℓ_2)-関係多様化を行うと，クラス c の $c[S_{pre}]$ は ℓ_1 種類以上，$c[S_{con}]$ は ℓ_2 種類以上になり，$c[S_{pre}]$ と $c[S_{con}]$ の関係は，$(c[S_{pre}], c[S_{con}])$ のセンシティブ属性値の集合の二項関係へと汎化される。また，この汎化された関係は以下のように表すこともできる．

$$R^*(c) = \{(s_{pre}, s_{con}) \in c[S_{pre}] \times c[S_{con}]\} \tag{4.1}$$

例えば，(3, 2)-関係多様性を保証した場合，前提部に $\{a, b, c\}$，結論部に $\{x, y\}$ なる二項関係を持つ (3, 2)-関係多様化されたクラスが存在することを考える．このクラスを成すタプルのオリジナルの関係がそれぞれ (a, x)，(b, y)，(c, x) とする．このとき，前提部 $\{a, b, c\}$，結論部 $\{x, y\}$ から推測可能な関係 (a, x)，(a, y)，(b, x)，(b, y)，(c, x)，(c, y) には，オリジナルの関係には存在しない関係 (a, y)，(b, x)，(c, y) が含まれていることが分かる．関係多様化によって混入するオリジナルの関係には存在しない関係をノイズ関係（Noisy Relation）と呼び，以降，単にノイズと呼ぶ．一方，オリジナルの関係集合が (a, x)，(a, y)，(b, x)，(b, y)，(c, x)，(c, y) であるタプルからクラスが作られている場合には，推測可能な関係とオリジナルの関係が同じになり，ノイズが混入しない．ノイズが混入すると，様々なデータ分析の精度に影響を与える可能性があるため，後者の例のようにノイズが混入しないことが望ましい．

ここで，クラス中のオリジナルの関係とノイズとの比を表す関係ノイズ比 RNR を以下のように定義する．

定義 16 （関係ノイズ比）

$$RNR(c) = \frac{|R^*(c)|}{|R(c)|} \tag{4.2}$$

$RNR(c)$ は 1 以上の値を取り，最小値（1）のとき，クラスにノイズが混入していないことを表す．ノイズの混入のないクラスを**ノイズレスクラス**と呼ぶ．

本稿の関係ノイズ比は，クラス単位の局所的な視点に基づくものであり，テーブル全体の関係ノイズ比を全域的な視点で評価したものではない．また，本稿で定義した関係の曖昧性指標はノイズの頻度や偏りを捉えられておらず，それらによる関係の曖昧化を小さく見積もってしまう場合がある．本稿では，評価指標を簡略化するためにノイズの有無による指標を用いた．ノイズの頻度や偏りを考慮した評価指標の確立は今後の課題とする．

4.4 ナイーブな (ℓ_1,ℓ_2)-関係多様化手法

関係の曖昧化を抑止した関係多様化テーブルの生成について考える．最適な関係多様化は，関係多様化後の各タプルの関係ノイズ比の総和が最小となる関係多様化である．

最適な k-匿名化は，NP 困難な問題であることが知られている [49]．また，関係多様化と同様に，属性間の関係が曖昧化されるようにテーブルを分割する手法である Anatomy[77] においても，準識別子の曖昧性を最小化する最適解の導出は NP 困難であるとされている．

そこで本稿では，ヒューリスティックな手法によってできるだけ関係ノイズ比が小さいクラスを生成することを考える．ここでは，クラスタリングによって関係多様化を実現する手法を提案する．4.4.1 節ではクラスタリングによる関係多様化手法を述べ，4.4.2 節では関係ノイズ比を考慮した類似度指標を述べる．

4.4.1 クラスタリングによる関係多様化

まず，関係の曖昧性については考慮せず，タプル群のクラスタリングによって関係多様化を実現するナイーブな手法について述べる．

本節で述べるクラスタリングによる手法では，テーブル T のセンシティブ属性 S_i とその前提部のセンシティブ属性 S_{i-1} を対象とし，の関係多様化後のテーブル T_i^* を生成する．よって，d 次元のセンシティブ属性を持つ場合は，d 回のクラスタリングを行う．

以降，議論を簡単にするために，(S_{i-1}, S_i) の二項関係 $r_{i,j} = (t_j[S_{i-1}], t_j[S_i])$ を対象とする．$r_{i,j}$ から定義 13 のクラスを生成すると (1,1)-関係多様性を持ったクラスになる．このクラスを複数併合し，共通の CID を付与することで (ℓ_1, ℓ_2)-関係多様性を充足させることを考える．

クラスタリングの手順

クラスタリング手法として，凝集型の階層的クラスタリング [19] を用いる．関係多様化に用いる凝集型の階層的クラスタリングは，以下のステップのアルゴリズムを採用する．

- ステップ 1: 各関係から 1 つの関係だけを含むクラス c を生成
- ステップ 2: クラス間の類似度 $sim(c_1, c_2)$ を計算し，類似度が高いペアを併合
- ステップ 3: 併合したクラスがノイズレスクラスになったらクラスタリングの対象から除く
- ステップ 4: ステップ 2 で併合が行われればステップ 2〜3 を繰り返し，併合が行われなければ終了

(ℓ_1,ℓ_2)-関係多様化のためのクラス間類似度

クラス間の類似度 $sim(c_1, c_2)$ は，2 つのクラスを併合した場合における (ℓ_1,ℓ_2)-関係多様性の充足性（充足の度合い）によって評価する．

ここで, クラス c の前提部の多様性は $div_{pre}(c) = |c[S_{pre}]|$, 結論部の多様性は $div_{con}(c) = |c[S_{con}]|$ である．2 つのクラス c_1 と c_2 を併合したクラス $c' = c_1 \cup c_2$ の前提部，結論部の多様性はそれぞれ $div_{pre}(c_1 \cup c_2)$，$div_{con}(c_1 \cup c_2)$ である．

c_1 と c_2 を併合した際の多様性の変化量を式 4.3 と式 4.4 に示す．

$$\triangle div_{pre}(c_1, c_2) = div_{pre}(c_1 \cup c_2) - max(div_{pre}(c_1), div_{pre}(c_2)) \tag{4.3}$$

$$\triangle div_{con}(c_1, c_2) = div_{con}(c_1 \cup c_2) - max(div_{con}(c_1), div_{con}(c_2)) \tag{4.4}$$

$\triangle div_{pre}(c_1, c_2)$ や $\triangle div_{con}(c_1, c_2)$ が正のとき，クラス c_1 と c_2 を併合することで，(ℓ_1,ℓ_2)-関係多様なクラスに近づくことができる．(ℓ_1,ℓ_2)-関係多様性の充足性を式（4.5）で表す．

$$rdiv(c_1, c_2) = \frac{div_{pre}(c_1 \cup c_2) + div_{con}(c_1 \cup c_2)}{\ell_1 + \ell_2} \tag{4.5}$$

$rd = rdiv(c_1, c_2)$ は 2/(ℓ_1 + ℓ_2) $\leq rd \leq$ 1 の値を取り，1 のとき (ℓ_1,ℓ_2)-関係多様性を満たすことを表す．よって，$rdiv$ 値が高いクラスペア同士を併合すると，(ℓ_1,ℓ_2)-関係多様性の充足に大きく近づく．

これまでの議論より，(ℓ_1,ℓ_2)-関係多様性の充足性の高いクラスペアを発見するためのクラス間の類似度指標として，以下の式（4.6）の DG(Diversity Gain) を導入し，凝集型の階層的クラスタリングで関係多様化を行う．

$$DG(c_1, c_2) = \begin{cases} rdiv(c_1, c_2) & (\triangle div_{pre}(c_1, c_2) > 0) \\ rdiv(c_1, c_2) & (\triangle div_{con}(c_1, c_2) > 0) \\ 0 & (otherwise) \end{cases} \tag{4.6}$$

4.4.2 関係ノイズ比を考慮した関係多様化

本節では，関係ノイズ比を考慮してクラスタリングを用いて関係多様化する手法 $DGRL$ について述べる．クラスタリングは，4.4.1 節と同じ手順に従う．

関係ノイズ比を考慮したクラス間の評価指標を考える．まず，関係ノイズ比 RNR を考慮して，関係の損失を表す評価値 RL（Relation Loss）を導入する．

$$RL(c_1, c_2) = exp(RNR(c_1 \cup c_2) - 1) \tag{4.7}$$

DG と RL を用いて，関係多様性の充足率と，関係ノイズ比の両方を加味した評価値 $DGRL$ を式（4.8）のように定義する．

$$DGRL(c_1, c_2) = \frac{DG(c_1, c_2)}{RL(c_1, c_2)} \tag{4.8}$$

上述の $DGRL$ をクラス間類似度 $sim(c_1, c_2)$ として凝集型の階層的クラスタリングを用いることで，関係ノイズ比が小さい (ℓ_1,ℓ_2)-関係多様なクラスを生成できる．

例 1 *(a, x)*，*(b, y)*，*(a, y)*，*(b, x)* をセンシティブ属性として持つタプルを対象にして，*(2,2)*-関係多様化させる場合を考える．*(a, x)*，*(b, y)* の併合は，DG 値が *2* の併合であり，それぞれ *(2,2)*-関係多様性を満たす．しかし，関係ノイズ比 RNR は *2* であり，各クラスにはノイズ関係が混入してしまう．さらにこの併合は RL 値は $exp(2-1) > 2$

であり，$DGRL$ 値は *2/e* である．一方で，*(a, x)* と *(a, y)* の併合は，ノイズ関係を混入せずに *(1,2)*-関係多様性を得られる．このとき，DG 値は共に *2*，RL 値は共に *1* であり，$DGRL$ 値は *1* である．よって，$DGRL$ 値の高い後者の組み合わせを採用し，同様に *(b, x)* と *(b, y)* を併合する．続いて，併合された *2* つのクラスをさらに併合すると，ノイズ関係のない *(2,2)*-関係多様化を実現できる．

4.4.3 ナイーブ手法の課題

凝集型の階層的クラスタリングは各ステップで最良の評価値を持つクラスタペアを探索するクラスタリング手法である．タプル数 $|T|$ に対して最悪のケースで O($|T|^3$) の計算量を要する [60]．そのため，スケーラビリティに課題があり，大量のタプルを持つテーブル T を対象とした際に大きな計算時間を要する．また，各クラス間の併合においては RNR を極小化できるが，必ずしも最終的に生成されたクラスの RNR の小ささを保証するものではない．

4.5 効率的なノイズレスクラス生成のための考察

本節では，関係ノイズ比 RNR のの小さなクラスを効率よく生成するために，ノイズレスクラスを効率的に生成する方法について議論する．特に，ノイズレスクラスを成すための条件や，ノイズレスクラスを成すタプル群が持つ性質について考え，効率化に資する補題等を導く．

以降の議論では，S_{i-1}，S_i をそれぞれ V，U とし，V と U の二項関係を持つデータ $R = \{r_1, \ldots, r_n\}$，$r = (v, u)$，$v \in V$，$u \in U$ を対象に議論を進める．

4.5.1 関係ベクトル

前提部 v_1 を持つ関係 $(v_1, u_1), \ldots (v_1, u_{\ell_2})$ が，ノイズなしに (ℓ_1, ℓ_2)-関係多様性を満たすためには，$\{v_2, \ldots, v_{\ell_1}\} \times \{u_1, \ldots, u_{\ell_2}\}$ のすべての関係を用いてクラスを成す必要がある．

ここで，前提部 v，結論部 u を持つ関係の多重集合を $R(v, u)$ とする．また，前提部 v を持ち，結論部は任意の値を持つ関係の多重集合を $R(v, *)$ と表し，前提部関係集合と呼ぶ．v_1 と多くのノイズレスクラスを形成できる $v_2 (\neq v_1)$ は，$R(v_1, *)$，$R(v_2, *)$ 間において，結論部の値に共通の値を多く含み，結論部の値の出現パターンが類似するものと言える．

補題 2 *(h, ℓ_2)*-関係多様性を満たすノイズレスクラスを生成可能な前提部関係集合 $R(v_1, *), \ldots, R(v_h, *)$ は，共通の結論部値を ℓ_2 以上持つ．

結論部の出現パターンの類似性を測るために，前提部関係集合毎に結論部のセンシティブ属性値の出現回数をベクトル形式で表現し，ベクトルの類似度によって出現パターンの類似性を測ることを考える．

まず，前提部関係集合毎に結論部のセンシティブ属性値の頻度をベクトル形式で表現する．

$$\boldsymbol{v_i} = (v_{i,1}, \ldots, v_{i,|S_{con}|})^T \tag{4.9}$$

ここで，$f_{i,j}$ を R_i 中の (v_i, u_j) の数とし，$v_{i,j} = f_{i,j} / \sum_{\ell=1}^{|S_{con}|} f_{i,\ell}$ とする．以降，$\boldsymbol{v_i}$ を**関係ベクトル**と呼ぶ.

ベクトルの類似度は，ベクトル間の内積（コサイン類似度）等によって求めることができる．ベクトル間の内積は，共通要素の出現パターンの類似性を表す．このとき，補題 2 より，少なくとも ℓ_2 種類の結論部が共起しない関係ベクトルの類似度は 0 とする．

これまでの議論から，関係ベクトル間の類似性（ノイズレスクラスの生成可能性）を以下の式で定義する．

$$sim(\boldsymbol{v_p}, \boldsymbol{v_q}) = \begin{cases} \boldsymbol{v_p} \cdot \boldsymbol{v_q} & (cmn(\boldsymbol{v_p}, \boldsymbol{v_q}) \geq \ell_2) \\ 0 & (otherwise) \end{cases} \tag{4.10}$$

ここで，$cmn(\boldsymbol{v_p}, \boldsymbol{v_q})$ は，$\boldsymbol{v_p}$ と $\boldsymbol{v_q}$ の共通の結論部値の種類数である (式 4.11).

$$cmn(\boldsymbol{v_p}, \boldsymbol{v_q}) = \sum_{y_j \in U} sgn(v_{p,j}, v_{q,j}) \tag{4.11}$$

4.5.2 類似度グラフ

式 4.10 の関係ベクトル間の類似度からは，類似度行列を生成できる．この類似度行列に基づいて，各関係ベクトルを頂点とし，類似度が 0 より大きい頂点にエッジを張ることで，関係ベクトル間の類似度を表す無向グラフ $G := (\hat{V}, E)$ が生成できる (図 4.1)．ここで，$\hat{v}_i \in \hat{V}$ は，センシティブ属性値 v_i を表す頂点であり，$e_{i,j} \in E$ は $\hat{v}_i$ と $\hat{v}_j$ の間のエッジであり，類似度が 0 より大きい頂点間に張られる．この無向グラフ G を**類似度グラフ**と呼ぶ．エッジで接続された頂点同士は，$(2, \ell_2)$-関係多様なノイズレスクラス生成の候補である．

ノイズレスクラスを多数生成するためには，類似度の高い ℓ_1 個のベクトル群を抽出することが望まれる．しかしながら，ℓ_1 個のベクトル間の類似度計算は計算コストが高い．例えば，ℓ_1 個の関係ベクトル間の類似度計算は，各頂点を始点とした深さ優先探索によって実現することが可能である．この類似度計算のための深さ優先探索には，深さ優先探索が $\mathrm{O}(|\hat{V}| + |E|)$ の計算量を有することから，$\mathrm{O}(|\hat{V}|(|\hat{V}| + |E|))$ の計算量を要すると考えられる．前述の全探索と比較すると計算コストは小さいが，類似度グラフが更新される度に深さ優先探索を実施することを考えると，依然として計算コストが高い．

提案手法では，本節で導入した関係ベクトル，類似度グラフを前提とし，2 頂点間の類似度をベースとしたヒューリスティックな手法でノイズレスクラス生成を行う．そのために，以降では，ヒューリスティックな手法を実現するためのいくつかの性質について考える．

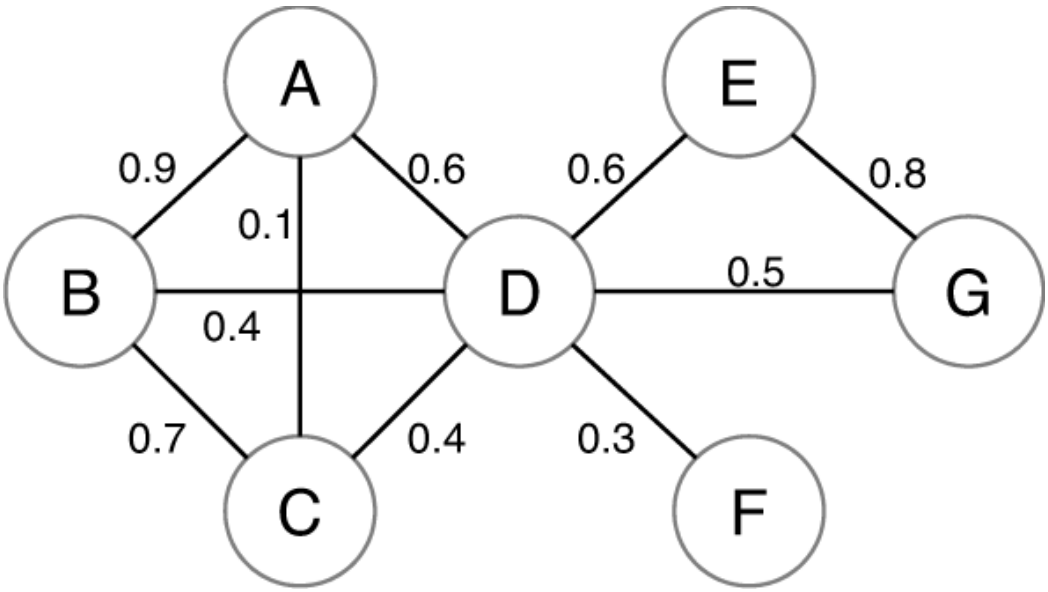


図 4.1: 類似度グラフ (初期状態)

類似度グラフの枝刈り

まず，ノイズレスクラスを成す関係ベクトル集合が持つ性質について考える．これによって類似度グラフを枝刈りし，ノイズレスクラスを成し得ない関係ベクトルをノイズレスクラス生成の候補から除外する．

ノイズレスクラスを成すためには結論部の共起が必要である．そのため，類似度グラフ上において，各頂点は $\ell_1 - 1$ 個の他の頂点と接続されている必要がある．

補題 3 *(ℓ_1, ℓ_2)-*関係多様性を満たすノイズレスクラスを生成可能な頂点 $\hat{v} \in \hat{V}$ は，エッジ数が ℓ_1*-1* 以上の頂点だけである．

補題 3 より，グラフ G からノイズレスクラスに成り得ない頂点を簡単に枝刈りすることができる．

さらに，ノイズレスクラスを成すためには，ℓ_1 個の関係ベクトルが互いに共通の結論部を ℓ_2 個持つ必要がある．そのため，類似度グラフ上において，少なくとも互いにエッジが張られていることが必要であり，互いに接続された ℓ_1 個の頂点だけがノ

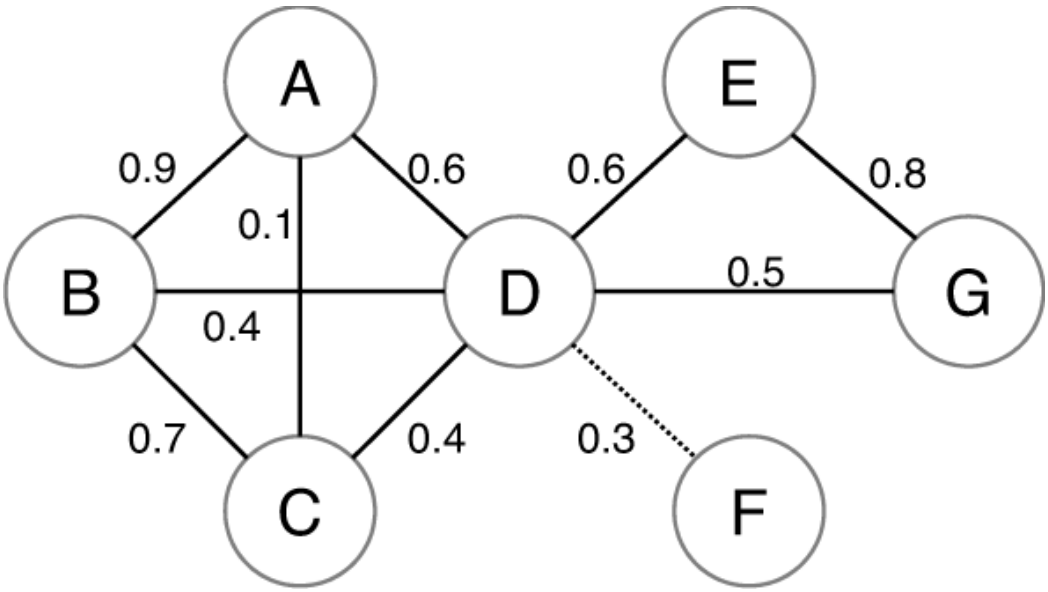


図 4.2: 類似度グラフ (枝刈り)

イズレスクラスを成し得る．言い換えると，頂点集合 $\forall\hat{v}_1,\ldots,\hat{v}_h \in \hat{V}$ からノイズレスクラスを生成するには，$\forall\hat{v}_i,\hat{v}_j(i \neq j) \in \{\hat{v}_1,\ldots,\hat{v}_k\}$ にエッジが張られたクリークを成すことが必要である．

補題 4 *(ℓ_1, ℓ_2)*-関係多様性を満たすノイズレスクラスが生成可能な頂点の集合 $\hat{V}' \subseteq \hat{V}$ は，$|\hat{V}'| \geq \ell_1$ であり，$\forall\hat{v} \in \hat{V}'$ で ℓ_1 次のクリークを成す $\hat{V}'$ だけである．

補題 4 を用いると，多くのエッジを枝刈りできる可能性がある．しかしながら，特定のグラフから指定された大きさのクリークを探索する問題は NP 完全であることが知られている [37]．さらにノイズレスクラス生成によって関係ベクトル間の類似度が変化し，グラフが更新される度にクリークの探索を行うことはリーズナブルでない．

本稿では，クリークの探索によるノイズレスクラス生成は行わず，補題 3 によってノイズレスクラスを成し得る関係ベクトルを絞り込んだ上で，2 頂点間の類似性が高い頂点集合からノイズレスクラスを生成する．

4.6 効率的なノイズレスクラス生成手法 NLC

本章では，4.4.3 節で述べた凝集型クラスタリングによる (ℓ_1,ℓ_2)-関係多様化の課題である効率性と，関係多様化の精度を，関係多様化の際に利用できるいくつかの性質を用いて改善する手法を提案する．

提案手法では，センシティブ属性 S_i とその前提部である S_{i-1} を対象に，関係ノイズ比を低減した (ℓ_1, ℓ_2)-関係多様化を行う．その際に，4.5 節の議論に基づいてノイズレスクラスを優先的に (貪欲に) 生成し，関係ノイズ比の低減を図る．

提案手法は以下の手順のアルゴリズムによってノイズレスクラスを生成する．

1. 関係ベクトルと類似度グラフの生成
2. 対象データの選択（4.6.2 節）
3. 前提部の多様化 ((ℓ_1, 1)-関係多様化)（4.6.3 節）
4. 結論部の多様化 ((ℓ_1, ℓ_2)-関係多様化)（4.6.4 節）
5. 類似度グラフの更新（4.6.5 節）
6. 2.〜4. をノイズレスクラスが生成できなくなるまで繰り返す
7. ノイズレスクラスを成していないタプル群を関係多様化クラスタリング（4.4 節）

ステップ 1 では，4.5.1 節と 4.5.2 節で議論した関係ベクトルと類似度グラフの生成および枝刈りを行う．ステップ 2 では，関係ベクトル間の類似性からノイズレスクラス生成の対象とする ℓ_1 個の関係ベクトルを生成する．次にステップ 3 では，ステップ 2 で抽出した関係ベクトル群のタプルを用いて，同じ結論部を持ち，異なる前提部を持つ (ℓ_1, 1)-関係多様化したクラスを生成する．ステップ 4 では，ステップ 3 で生成し

た (ℓ_1, 1)-関係多様化したクラスを併合して (ℓ_1, ℓ_2)-関係多様性を満たすクラスを生成する．ステップ 5 では，ノイズレスクラスを形成したタプルに関する情報を関係ベクトル，類似度グラフに反映し，これらの更新を行う．最後に，ノイズレスクラスを形成できなかったタプルについては，4.4 節に示した関係多様化クラスタリングによって関係多様化を行う（類似度には $DGRL$ を用いる）．

4.6.1 関係ベクトルと類似度グラフの生成

4.5 節の 4.5.1 節と 4.5.2 節に基づいて，前提部関係集合から関係ベクトルを生成し，関係ベクトル間の類似度を求めて類似度グラフを生成する (図 4.1)．その上で，補題 2 と補題 3 によってグラフ G を枝刈りし，ノイズレスクラスを生成し得る頂点を絞り込む (図 4.2)．

4.6.2 対象データの選択

次に，頂点 $\hat{v}$ にとって最良な併合対象である隣接頂点の選択を行う．ここで，$\hat{v}$ の隣接頂点の集合を $n_G(\hat{v})$ とする．このとき，$\hat{v}$ と類似度の高い隣接頂点 $\hat{u_1} \in n_G(\hat{v})$ はノイズレスクラスを生成できる可能性も高い．そこで，$\hat{v}$ との類似度が高い上位 ℓ_1-1 個の頂点の集合を $\hat{v}$ のノイズレスクラス生成候補とする．

続いて，どの頂点を中心としたノイズレスクラス生成を行うべきかを判断するために，各頂点がどれだけノイズレスクラス生成に適しているかを評価する．この評価では，$\hat{v}$ を隣接頂点の類似度の積によって評価する．ここで，頂点 $\hat{v}$ の隣接頂点集合 $n_G(\hat{v})$ のうち，$\hat{v}$ との類似度が高い上位 ℓ_1-1 頂点の集合を $n^*_G(\hat{v})$ とする．頂点 $\hat{v}$ の評

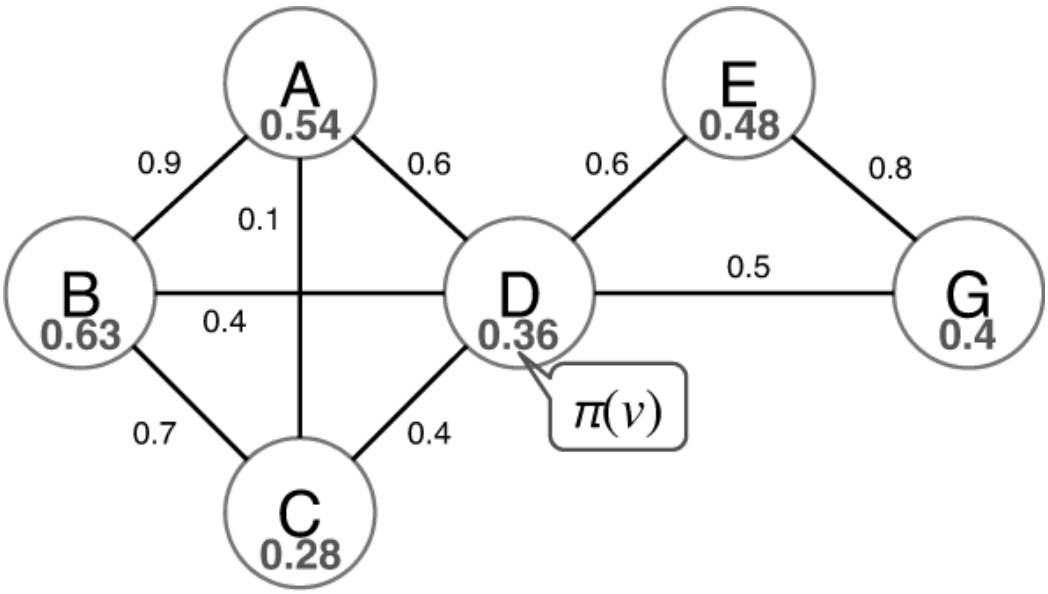


図 4.3: 類似度グラフ (頂点の評価)

価値 $\pi(\hat{v})$ を式（4.12）のように定義する．

$$\pi(\hat{v}) = \prod_{\hat{w} \in n_G^*(\hat{v})} sim(\boldsymbol{v}, \boldsymbol{w}) \tag{4.12}$$

$\pi(\hat{v})$ が最大の頂点 $\hat{v}_{max}$ とノイズレスクラス生成の基準とし，$\hat{v}_{max}$ と $\hat{w} \in n_G^*(\hat{v}_{max})$ の頂点集合 $\hat{V}'$ からノイズレスクラスを生成する．

4.6.3 前提部の多様化

ノイズレスクラス生成を行うための候補となる関係ベクトル群を抽出したら，それらを用いてノイズレスクラスを生成する．まず，結論部に基づいて，ノイズのない (ℓ_1, 1)-関係多様性を満たすクラスを生成する．

頂点 $\hat{v'}_1, \ldots, \hat{v'}_{\ell_1} \in \hat{V}'$ に関するタプル集合を用いて，ノイズレスクラスの生成を行

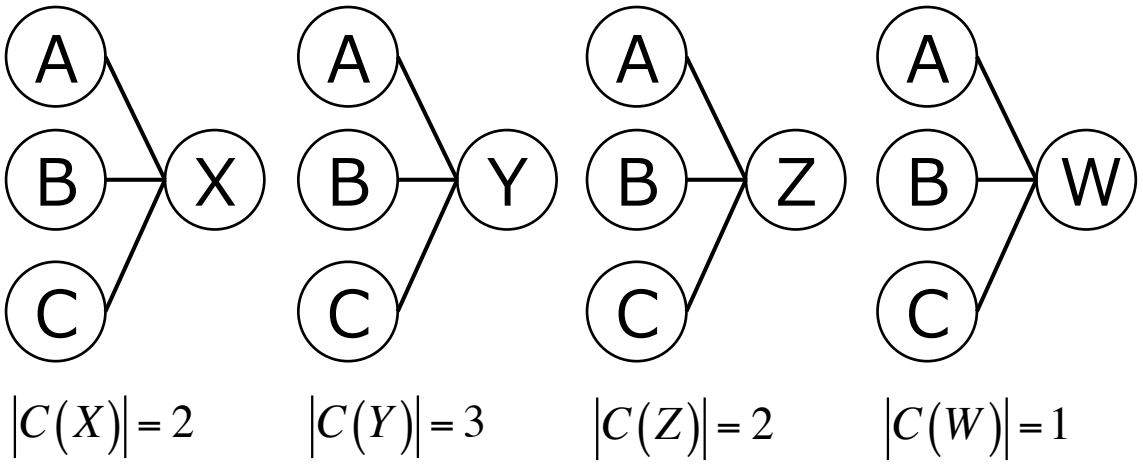


図 4.4: 結論部に基づく前提部の ℓ_1-多様化 ($(\ell_1, 1)$-関係多様化)

う．このとき，同一の結論部値と，異なる前提部値を持つタプル群によって (ℓ_1, 1)-関係多様性を満たすクラスの生成を行う (図 4.4).

結論部値 $u_j \in U$ に対して，関係 (v'_i, u_j) を $R(v'_1, *), \ldots, R(v'_{\ell_1}, *)$ から抽出し，$\{(v'_1, u_j), \ldots, (v'_{\ell_1}, u_j)\}$ から成るクラスを生成する．これによって，(ℓ_1, 1)-関係多様なノイズレスクラスが生成される．ここで生成した結論部値に u_j を持つクラスの集合をクラス集合 $C(u_j)$ とする．この操作を $R(v'_1, *), \ldots, R(v'_{\ell_1}, *)$ から 1 つでも関係 (v'_i, u_j) を抽出できなくなるまで繰り返す．さらに，すべての $u_j \in U$ に対して同様に実施する．

4.6.4 結論部の多様化

次に，結論部に基づくクラス併合で (ℓ_1, 1)-関係多様化したクラス群を，(ℓ_1, ℓ_2)-関係多様化する．

ここで，クラス集合 $C(u_j)$ のクラス数を $|C(u_j)|$ とする．$|C(u_j)|$ が大きい上位 ℓ_2 個のクラス集合の集合 $C(u_1), \ldots, C(u_{\ell_2})$ を選択する．$C(u_1), \ldots, C(u_{\ell_2})$ からクラスを 1

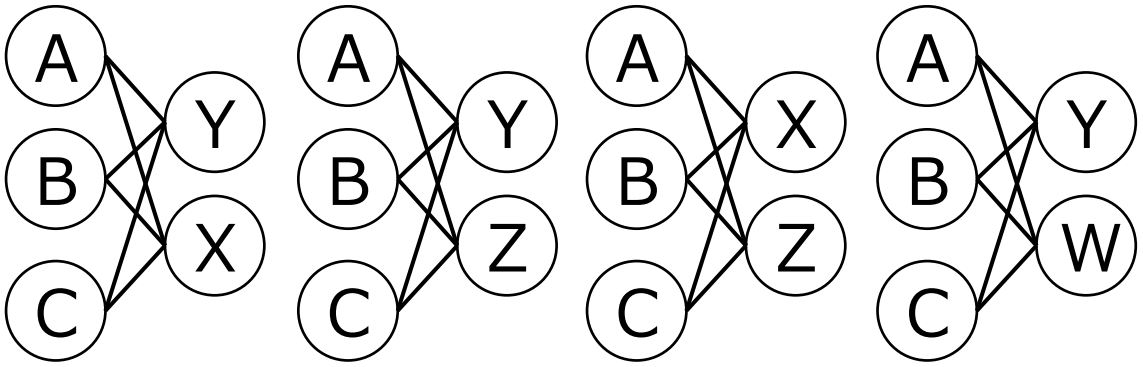


図 4.5: 前提部に基づく結論部の ℓ_2-多様化 ((ℓ_1, ℓ_2)-関係多様化)

つずつ選択し，1 つのクラスへ併合する．これによって (ℓ_1, ℓ_2)-関係多様性を充足するノイズレスクラスが生成される (図 4.5).

例 2 図 *4.4* と図 *4.5* は，*(3, 2)*-関係多様化の例を示している．まず結論部値 X, Y, Z, W 毎にクラス集合 $C(X)$, $C(Y)$, $C(Z)$, $C(W)$ を生成する（図 *4.4*）．次に，頻度の高い上位 ℓ_2=*2* 件のクラス集合 $C(Y)$, $C(X)$ から *1* つずつクラスを抽出して，抽出したクラス群を *1* つのクラスに併合する（図 *4.5*）．この操作を，上位 ℓ_2=*2* 件のクラス集合から *1* つずつクラスが抽出できなくなるまで繰り返す．

4.6.5 類似度グラフの更新

4.6.3 節と 4.6.4 節のグルーピングでノイズレスクラスを成した関係を，各前提部関係集合から取り除く．これに伴い関係ベクトルと関係ベクトル間の類似度を更新し，類似度グラフ G を更新する．G の頂点数 $|\hat{V}|$ が ℓ_1 個以上存在する間，ノイズレスクラス生成を繰り返す.

4.7 評価

提案手法である効率的ノイズレスクラス生成手法 NLC（4.6 節）の有効性を評価するために，評価実験を行った．

4.7.1 評価内容

比較対象として本稿で示したナイーブな関係多様化手法である関係ノイズ比を考慮したクラスタリングによる手法 DGRL（4.4.2 節），関係ノイズ比を考慮しないクラスタリングによる手法 DG（4.4.1 節）を用いた．

提案手法と上記 2 手法の関係ノイズ比，ノイズレスクラスの割合，関係多様化の計算時間を比較する．

関係ノイズ比の評価では，4.3.3 節で導入したクラス毎関係ノイズ比 RNR を利用して，すべてのクラスの RNR の平均値を評価する．ノイズレスクラスの割合は，すべてのレコードの内，ノイズレスクラスに属するレコードの割合を評価する．

NLC，DGRL，DG は，本稿に記載した各手法を実装し，評価に利用した．

4.7.2 評価環境

評価用のデータセットとして，2 種類の人工データを生成した．生成した人工データは，2 つのセンシティブ属性値にそれぞれ 10 種類の値を持つデータセット SA10 と，50 種類の値を持つデータセット SA50 であり，各タプルのセンシティブ属性にはランダムに値を割り当てた．本評価では 1,000，2,000，3,000，5,000，10,000 件のタプルを持つ SA10，SA50 を生成した．なお，本評価では，関係多様化によるノイズの発生度合いを純粋に評価するために，準識別子を含まない人工データを対象とした．

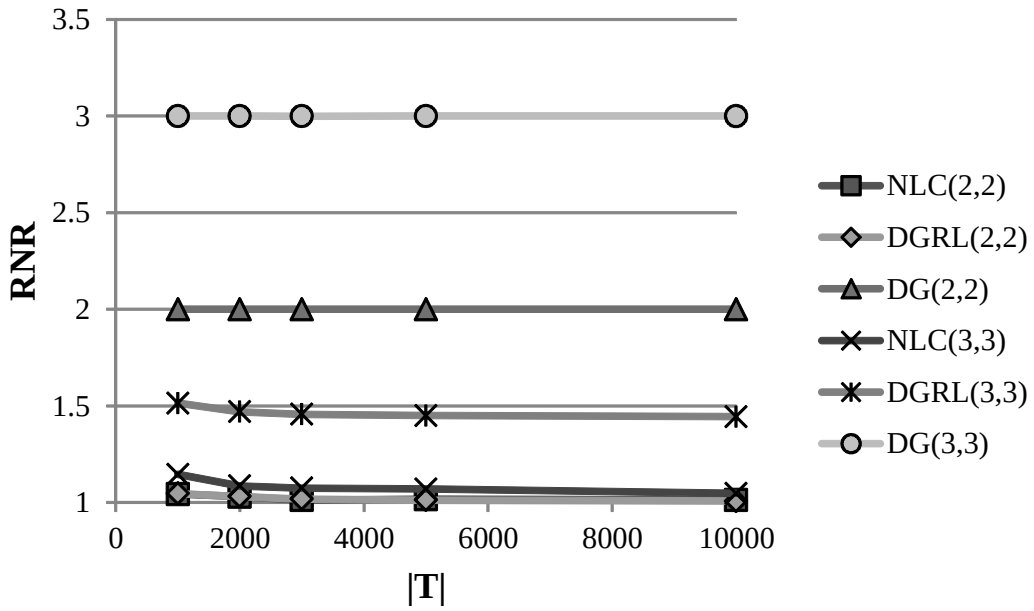


図 4.6: RNR の平均値 (手法の比較)

本評価では, 仮想マシン上で評価を行った．評価で用いた仮想マシンは, 4 コア CPU と，32GB のメモリ，120GB のディスクを持つ．仮想マシンのホストは，12 コア（24 スレッド）の 2.4GHz の CPU と，192GB のメモリ，6TB のディスクを持つ．仮想マシン上で Java 言語（Java 1.6.0_32）によって実装した．匿名化対象のデータセットは PostgreSQL（PostgreSQL 8.4.13）に格納し，匿名化結果も PostgreSQL に格納する．

4.7.3 評価結果

関係ノイズ比

図 4.6 は，(2, 2), (3,3)-関係多様性を充足させた 3 手法の RNR をデータセットサイズごとにプロットして示している．関係多様化の対象としたデータセットは SA10 である．

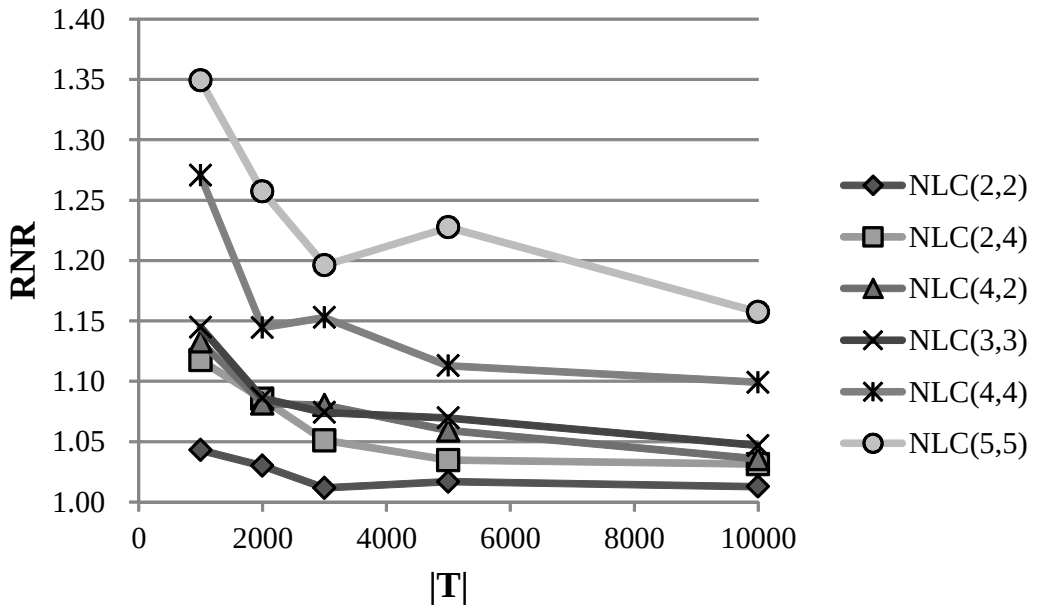


図 4.7: RNR の平均値 (SA10)

提案手法の NLC は，(2,2)-関係多様化，(3,3)-関係多様化において，RNR が 1 に近い値であり，関係多様化によるノイズの混入が非常に少ない．DGRL は，(2,2)-関係多様化においてはノイズの混入率は非常に少ないが，(3,3)-関係多様化の際にはノイズの混入によって本来の関係数の 1.5 倍程度の関係数に見えてしまう．RNR を考慮しないナイーブな手法である DG はノイズの割合が非常に多く，本来観測され得る関係を発見することが困難になってしまっている．以上より，クラス単位での局所的な関係ノイズ比においては，提案手法が低く抑えることができたと言える．

次に提案手法 NLC について，さらに詳細な評価を行った結果について示す．図 4.7 は SA10 に対して，(2,2), (2,4), (4,2), (3,3), (4,4), (5,5)-関係多様性をそれぞれ充足させた場合の RNR を示している．充足すべき多様性が小さいほど，RNR は小さい．(2,4), (4,2), (3,3)-関係多様化結果を比較すると，(3,3)-関係多様化した場合が最も RNR が高い．これは，ノイズレスクラス生成に必要となる最小レコード数が，(2,4), (4,2)-

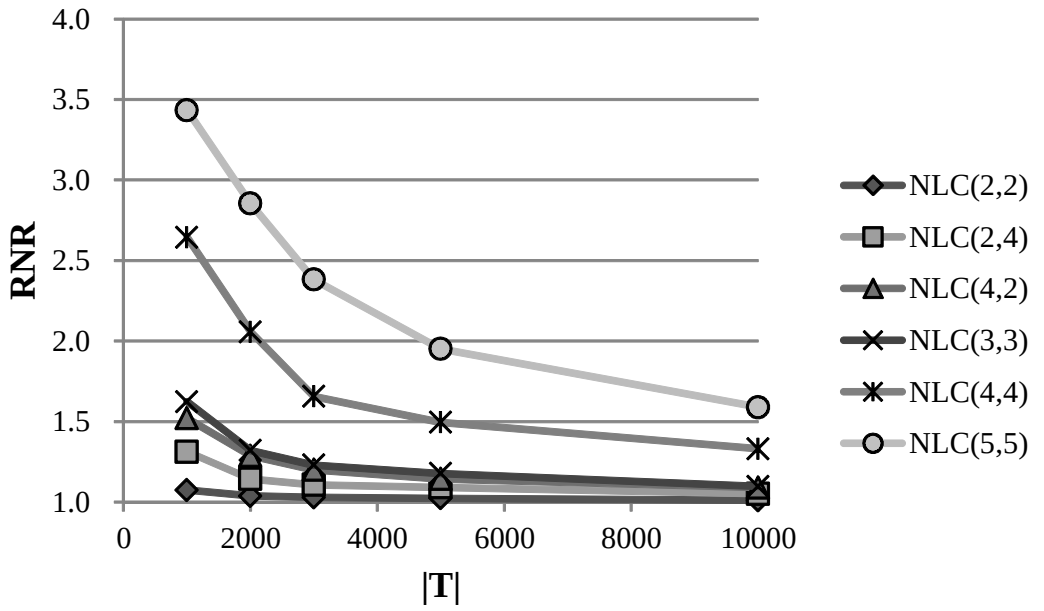


図 4.8: RNR の平均値 (SA50)

関係多様化の場合は 8 であるのに対して，(3,3)-関係多様化の場合は 9 であり，ノイズレスクラスを生成するためにより多くの種類の関係を持つレコードを必要とし，ノイズレスクラスの実現の困難さが高いためと考えられる．

図 4.8 は SA50 に対して同様に関係多様化をした際の RNR を示している．図 4.7 と図 4.8 を比較すると，属性値の種類数が多い SA50 を対象とした図 4.8 の結果が明らかに RNR が高い．同一のレコード数である際には，属性値の種類数が多いほど同一の属性値を持つレコードが少なくなり，ノイズレスクラスを生成することが困難になるためと考えられる．

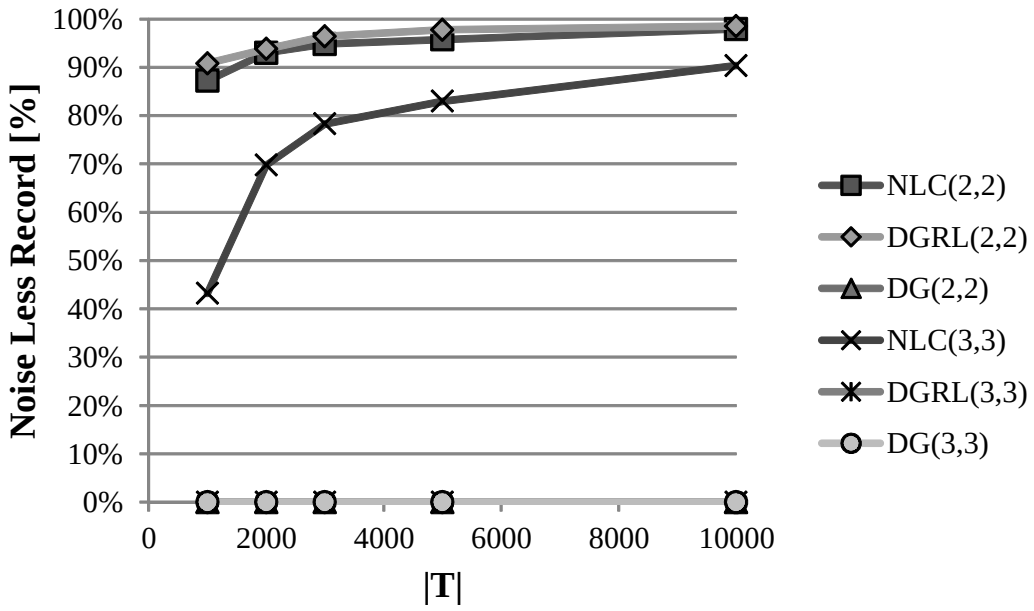


図 4.9: ノイズレスレコードの割合 (手法の比較)

ノイズレスレコードの割合

次に関係多様化テーブル中のノイズレスクラスに属するレコード（ノイズレスレコード）の割合を評価した．図 4.9, 図 4.10, 図 4.11 には，関係多様化テーブルのレコードのうち，ノイズレスレコードの割合を示している．

図 4.9 は各手法のノイズレスレコードの割合を比較した結果である．提案手法 NLC は (2,2)-関係多様化の際には，90% 以上のレコードがノイズレスレコードである．また，提案手法のようにノイズレスクラス生成に関するヒューリスティクスを用いない DGRL も (2,2)-関係多様化において，85% 以上のレコードがノイズレスレコードである．しかしながら，RNR を考慮しない DG はノイズレスレコードが存在しない．よって，RNR を考慮して関係多様化することで，多くのレコードのセンシティブ属性の二項関係を過度に曖昧化しないことがわかる．(3,3)-関係多様化の際には，他の

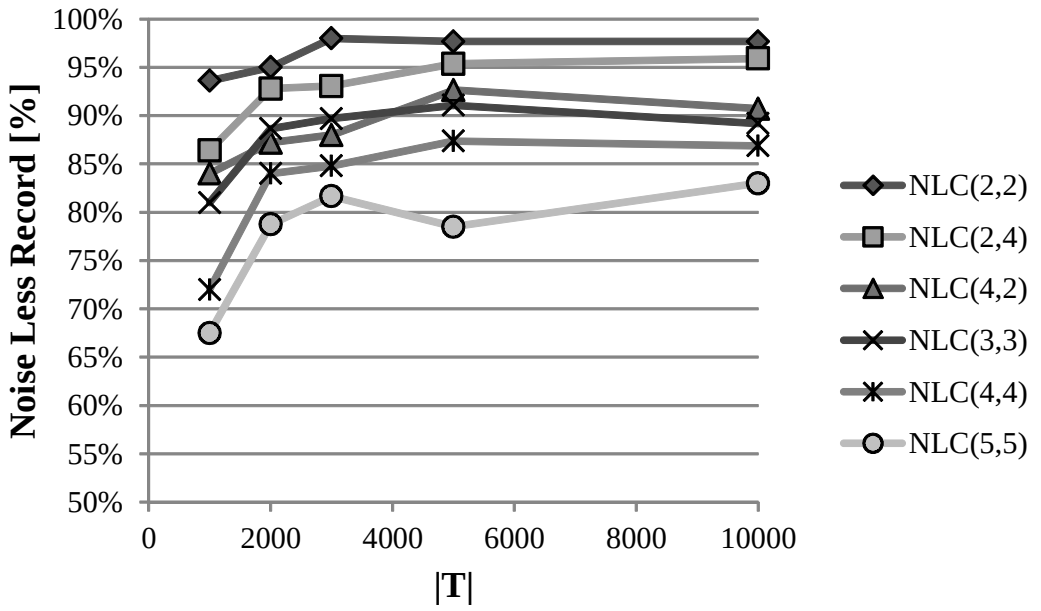


図 4.10: ノイズレスレコードの割合 (SA10)

手法がほぼノイズレスレコードを生成できていないことに対して，提案手法 NLC は多くのノイズレスレコードを生成している．これは，より多くのレコードがノイズレスクラスを成すようなレコードのグループ化処理を導入しているためと考えられる．

さらに，提案手法 NLC について，いくつかの関係多様性を充足させた場合のノイズレスレコードの割合を図 4.10 と図 4.11 に示す．充足すべき関係多様性が小さいほど，多くのノイズレスレコードが生成できていることが分かる．しかしながら，ノイズレスクラスを優先的に生成する NLC であっても，充足すべき関係多様性が高い場合には，ノイズレスクラスがほとんど生成できないことがわかる．特に，図 4.11 に示した SA50 を対象とした場合には，(5,5)-関係多様化においてノイズレスレコードが 1 つもない．よって，データの性質を棄損せずに関係多様化をしたい場合には，データセットの性質や規模によって ℓ_1 や ℓ_2 を調整する必要がある．

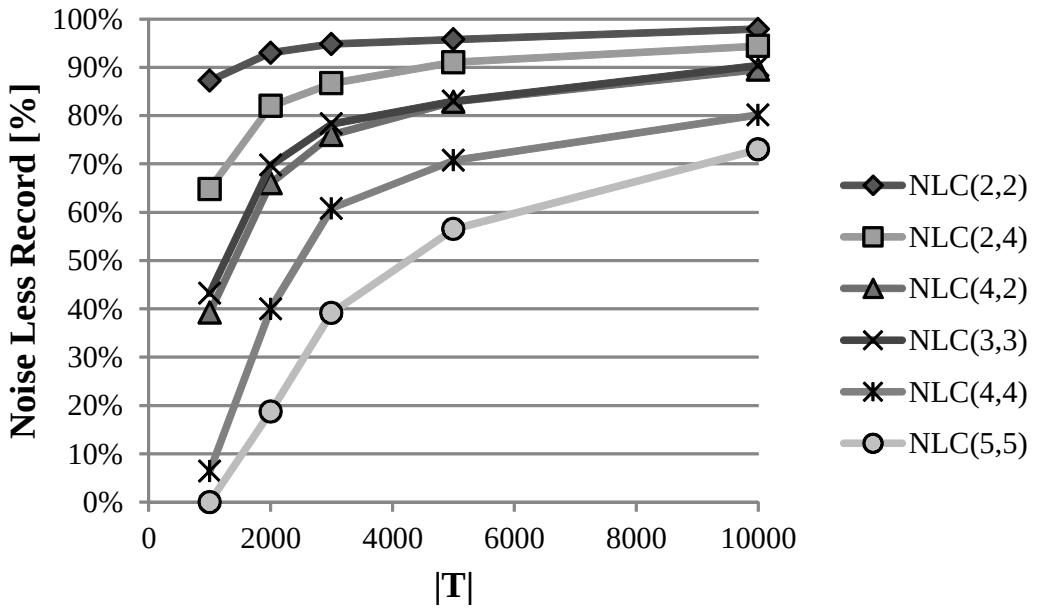


図 4.11: ノイズレスレコードの割合 (SA50)

計算時間

関係多様化手法のスケーラビリティを評価するために，各手法の計算時間を計測した．図 4.12 は各手法の計算時間を示している．提案手法 NLC は他の手法と比較して 10 倍以上高速であることが分かる．また，データサイズの増加に対して計算量が比較的増大しやすい傾向にあることが分かる．DGRL と DG はほぼ同程度の計算時間である．これは，DGRL と DG は距離計算の方法以外は同一の凝集型クラスタリングによって実現しているためである．NLC は凝集型クラスタリングに先立って，ヒューリスティクスを用いてノイズレスクラスを生成する．このノイズレスクラスが非常に効果的であると考えられ，かつ，関係多様化のために必要となる類似度計算の対象数を大幅に減らすことによって，DGRL と DG と比較して，大幅に高速化が実現できていると推察される．

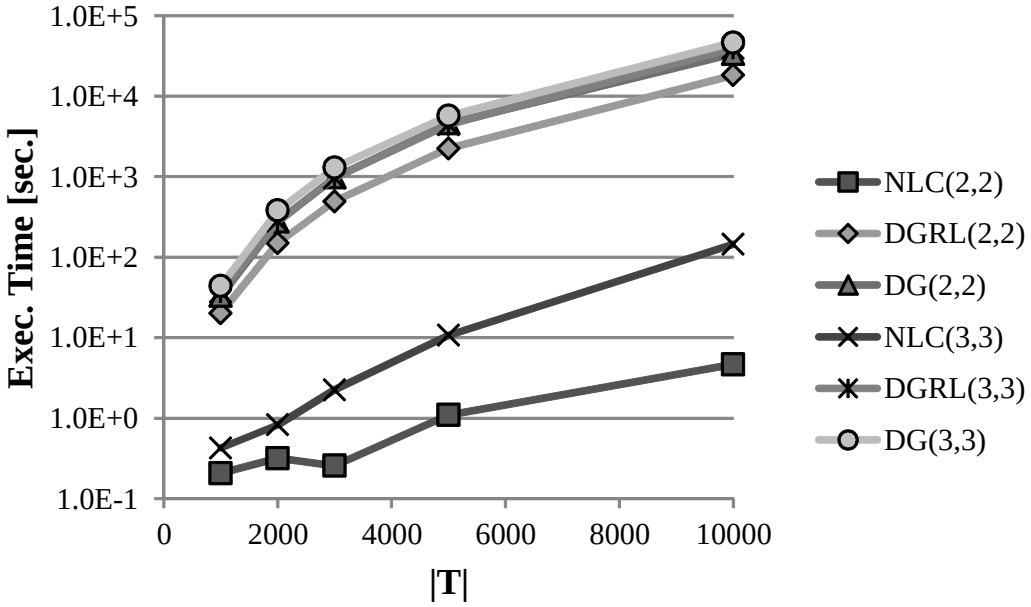


図 4.12: 計算時間 (手法の比較)

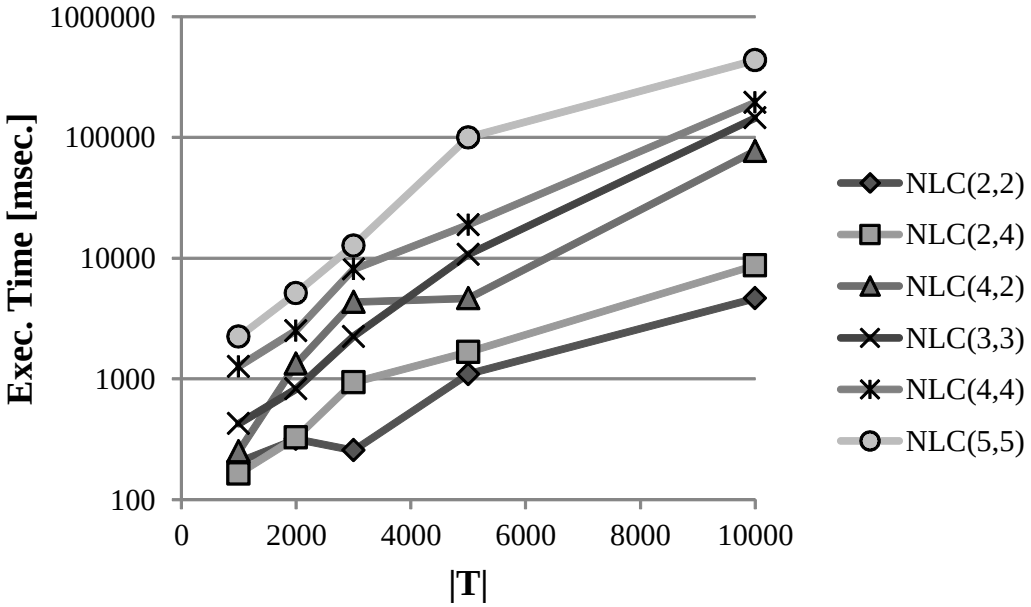


図 4.13: 計算時間 (SA10)

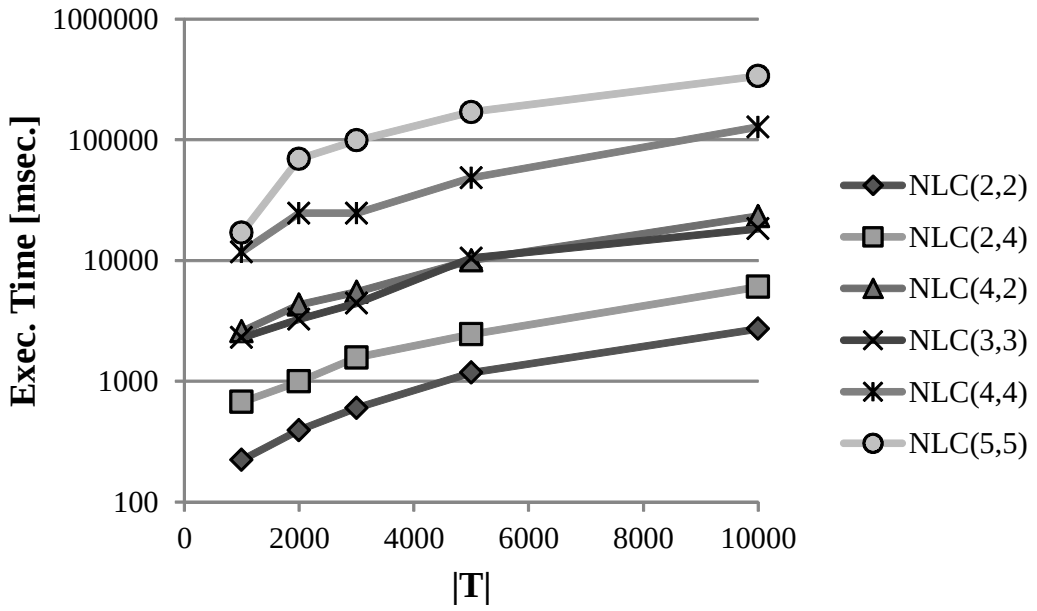


図 4.14: 計算時間 (SA50)

図 4.13 と，図 4.14 は SA10，SA50 それぞれに対して NLC で関係多様化した際の計算時間を詳細に示している．データ数が少ない場合，センシティブ属性の値の種類が少ない方が計算時間が短い．これは，ノイズレスクラスを多く生成し，凝集型クラスタリングのコストを低減することで計算時間が短縮されると考えられるため，センシティブ属性値の種類数が多い SA50 において，時間を要したのではないかと推察される．同様に，充足すべき関係多様性が高いほどノイズレスクラス生成が困難であるため計算時間が長いのではないかと考えられる．

以上より，提案手法 NLC は高い効率性を有しながら，関係多様化による関係の曖昧化を抑止できることが示された．

4.8 活用例と精度

複数のセンシティブ属性を持つパーソナルデータの活用方法として，センシティブ属性間の相関分析が挙げられる．

相関分析の際には，属性値間の共起頻度を算出することで，頻度の高い関係を分析対象のパーソナルデータの特徴的な関係として抽出できる．

しかしながら，関係多様化によってプライバシ保護を施した場合，センシティブ属性毎にテーブルが分割されたりと，テーブルの構造が変化してしまう場合がある．そのような場合には，元データと同様にデータを扱うことができず，工夫を要する．

本節では，関係多様化データを相関分析する際に利用者に必要とされる操作の一例を紹介すると共に，その際の分析精度について評価する．

4.8.1 データ操作の工夫

表 4.7 と表 4.8 は (2,2)-関係多様化したデータセットである．これらのデータを用いて SA1 と SA2 の相関関係を分析するために，SA1 と SA2 の共起頻度を求める．

表 4.7 と表 4.8 は CID と CID' によって曖昧に接続されている．例えば，TID=11 は TID=21, 22, 23, 24 と接続しており，TID=11 に該当するデータ主体の SA1 と SA2 の値の組が (a, x), (a, y), (a, x), (a, y) の 4 つのいずれかであるのかを特定することができない．よって，各関係が出現することの確からしさはそれぞれ 1/4 である．

表 4.7 と表 4.8 からは確かな関係を得ることができないため，表 4.7 と表 4.8 の間に存在し得る関係を列挙して，列挙された関係の出現の確からしさを考慮した共起頻度の算出を行う必要がある．表 4.9 は，表 4.7 と表 4.8 の間に存在し得る関係を列挙し，関係の確からしさ c を併記している．c は同じ TID(T1) を持つ関係数を m としたとき，$c = 1/m$ で導出した．

表 4.7: SA1 のテーブル (T_1)

TID	CID	CID'	SA1
11	G11	G21	a
12	G11	G21	a
13	G11	G22	b
14	G11	G22	b
15	G12	G22	a
16	G12	G22	a
17	G12	G21	c
18	G12	G21	c

表 4.8: SA2 のテーブル (T_2)

TID	CID	CID'	SA2
21	G21	—	x
22	G21	—	x
23	G21	—	y
24	G21	—	y
25	G22	—	x
26	G22	—	x
27	G22	—	z
28	G22	—	w

表 4.9 の c を，同じ値の組毎に集計し，総和を計算した結果 sum(c) を表 4.10 に示す．sum(c) は各関係の出現頻度の期待値である．表 4.10 には，関係多様化していない元データにおける関係の出現頻度 f_0 を併記している．

これまで述べたようなデータ操作を行うことで，関係多様化したデータセットからも，属性値間の共起頻度や相関関係を分析することができる．

表 4.9: SA1 と SA2 の共起 (全列挙)

TID(T1)	TID(T2)	SA1	SA2	c
11	21	a	x	0.25
11	22	a	x	0.25
11	23	a	y	0.25
11	24	a	y	0.25
12	21	a	x	0.25
12	22	a	x	0.25
12	23	a	y	0.25
12	24	a	y	0.25
13	21	b	x	0.25
13	22	b	x	0.25
13	23	b	y	0.25
13	24	b	y	0.25
14	21	b	x	0.25
14	22	b	x	0.25
14	23	b	y	0.25
14	24	b	y	0.25
15	25	a	x	0.25
15	26	a	x	0.25
15	27	a	z	0.25
15	28	a	w	0.25
16	25	a	x	0.25
16	26	a	x	0.25
16	27	a	z	0.25
16	28	a	w	0.25
17	25	c	x	0.25
17	26	c	x	0.25
17	27	c	z	0.25
17	28	c	w	0.25
18	25	c	x	0.25
18	26	c	x	0.25
18	27	c	z	0.25
18	28	c	w	0.25

表 4.10: SA1 と SA2 の共起関係

SA1	SA2	sum(c)	f_0
a	x	2.00	2
a	y	1.00	1
a	z	0.50	1
a	w	0.50	0
b	x	1.00	1
b	y	1.00	1
c	x	1.00	1
c	z	0.50	0
c	w	0.50	1

4.8.2 分析精度

図 4.15 に提案手法 NLC によって (2,2)-関係多様化したデータに対して，関係の共起頻度を算出した結果とその誤差を示す．評価には 1 万件の SA10 のデータセットを利用した．図 4.15 は，元データにおける各関係の出現頻度順に各関係の出現頻度を示している．センシティブ属性値はそれぞれ 10 種類ずつであるため，トータルで 100 種類の関係がある．関係多様化をしていない元データに対して分析をした場合と，(2,2)-関係多様化した場合とで誤差は生じている．しかし，平均で 0.7%，最大で 11.2%と小さな誤差であることがわかる．

図 4.16 は，図 4.15 と同様に提案手法 NLC によって (3,3)-関係多様化したデータに対して，関係の共起頻度を算出した結果とその誤差を示している．誤差は平均で 2.24%，最大で 20.9%である．図 4.15 と比べると，要求されている関係多様性が大きいため，誤差が大きくなっているが，提案手法によってノイズの出現が抑制されているため，大きな誤差は生じていない．

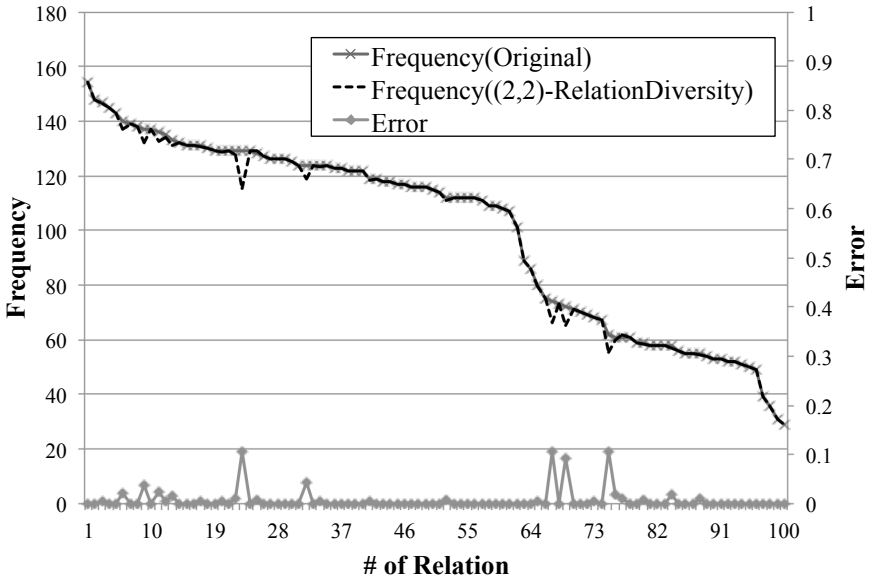


図 4.15: 相関関係の出現頻度 ((2,2)-関係多様化)

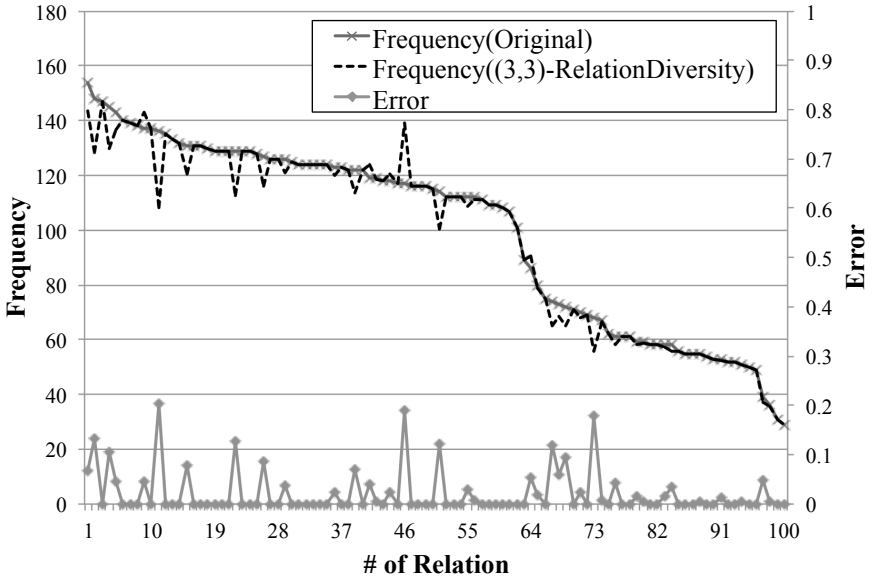


図 4.16: 相関関係の出現頻度 ((3,3)-関係多様化)

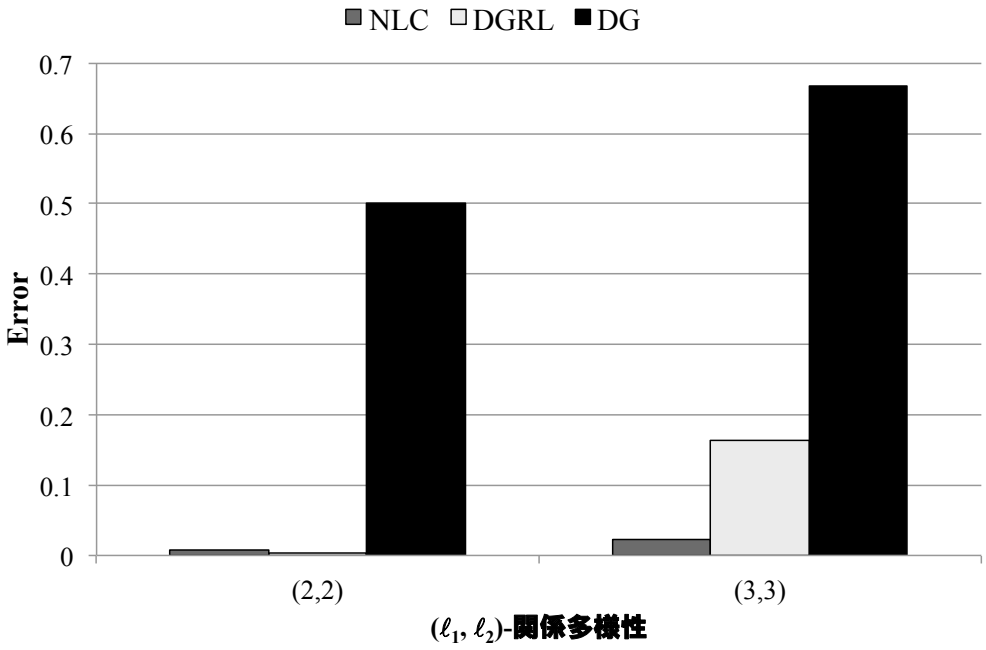


図 4.17: 共起頻度算出の誤差

図 4.17 は，提案手法の NLC と，クラスタリングによる手法 DGRL，DG によって関係多様化した場合の，共起頻度の算出の平均誤差を示している．(2,2)-関係多様化の際には，NLC と DGRL の誤差が非常に小さく，共に 1% 未満である．(3,3)-関係多様化の際には，NLC は約 2%，DGRL は約 16% の誤差である．関係ノイズ比を考慮していない DG は，極めて大きな誤差である．これらの結果より，提案手法 NLC が関係ノイズ比を低減することで，共起分析や，それらを応用した相関分析等の分析を，小さな誤差で実現できることが確認できた．

以上より，関係多様化したデータセットは，データ操作を工夫することで共起頻度を概算でき，それによって頻出な関係の抽出や相関関係の分析に用いることができると言える．また，その際に誤差が生じる可能性があるが，提案手法によって関係多様化した場合には誤差が小さく，分析結果に与える影響が小さいことが確認できた．

4.9 関連研究

同一のセンシティブ属性ではあるが，同一データ主体の複数のセンシティブ属性値を扱うデータに時系列データがある．時系列データに対する匿名化方式は，主に移動軌跡に対する匿名化方式が研究されている [1][70][57][68]．しかしながら，いずれも k-匿名性を拡張させた手法であり，センシティブ属性間の多様性を保証するものではない．

テーブルを分割することによって，関係を曖昧化し，属性値の推定を困難にする技術が提案されている [3][35]．Aggarwal ら [3] は，センシティブな関係が分割されるようにテーブル分割を行い，論理的に異なる 2 つ以上のサーバに配置して，関係の再構築を抑止する手法を提案している．Jiang ら [35] は，関係従属性のある属性間の関係をテーブル分割によって曖昧化し，ℓ-多様性を充足させる手法を提案している．両手法ともに，属性間の関係をテーブル分割によって曖昧化することについては本稿と同じモチベーションであるが，本稿のように，関係の曖昧化を抑止する仕組みが導入されていない．Jiang ら [35] の手法では，2 つ以上の属性間の関係の ℓ-多様化を提案している．本稿の手法も二項関係を繋ぎ合わせていくことで 2 つ以上の属性間に対応することが可能であるが，ノイズの発生を抑止できるのは二項関係のみであるという制限がある．

トランザクションデータに対して，アイテムの出現パターンをグラフ化し，グラフ中のサイクル等の特徴からアイテム集合をビット列に変換することで匿名化を行う手法が提案されている [84]．センシティブな属性値の出現パターンから生成したグラフを活用して匿名化を行う点は，本章の提案手法と類似するが，生成されるデータの形式や，保証する匿名性の指標が異なる．

4.10 まとめ

本研究では，2 つのセンシティブ属性を持つパーソナルデータを対象として，関係多様性を関係の曖昧化の度合いを抑制しながら実現するデータ匿名化を扱った．本研究では，系列データに対する関係多様化を導入し，関係多様化データの関係多様性指標として (ℓ_1, ℓ_2)-関係多様性の定義した．さらに，関係多様化に伴いセンシティブ属性間の関係が曖昧化される問題に対して，関係の曖昧性を抑止可能な関係多様化を高い効率性を有しながら実現する手法を提案した．

評価実験では，提案手法が関係の曖昧性を抑止しつつ効率よく (ℓ_1, ℓ_2)-関係多様化を実現でき，ナイーブな手法と比較して大幅に関係の曖昧性を抑止でき，10 倍から 100 倍の高い効率性を有していることを確認した．また，データ分析時にデータの操作に工夫を行うことで，誤差の小さいデータ分析が実現できることを確認した．

本研究の成果によって，系列データを活用する際に問題となる属性値の過度な汎化，もしくは属性間の関係の過度な曖昧化の両方を抑止したデータ匿名化が実現できる．

第5章　結言

本稿では，系列データの匿名化技術の内，移動軌跡ストリームの連続的匿名化手法と，センシティブ属性間の関係多様化手法を提案した．

移動軌跡ストリームの連続的匿名化手法では，移動軌跡のリアルタイムな利活用をプライバシを考慮しながら実現するために，移動軌跡ストリームの連続的匿名化手法を提案した．提案手法では，移動軌跡ストリームを匿名化する際に，位置情報を一定の精度を保ちながら連続的に匿名化を行うことができる．評価実験では，10万件程度の移動軌跡ストリームを毎分リアルタイムに匿名化できることを確認した．提案手法によって，人々の移動パターンの即時解析が実現できるようになる，また，移動のパターンに合わせた情報配信等も可能になると考えられる．今後は，移動軌跡ストリームの連続的匿名化手法は，位置情報の精度 (解像度) だけでなく，トレース可能時間についても一定の時間を保証するような手法の実現が望まれる．

センシティブ属性間の関係多様化では，センシティブ属性間である属性から他の属性がどの程度推測できるかを表す (ℓ_1, ℓ_2)-関係多様性という新たな匿名性の指標を定義し，これを実現する関係多様化を導入した．関係多様化では，センシティブ属性の属性値を汎化せずに，属性間の関係を曖昧化することで属性間の関係多様性の保証を実現した．さらに関係多様化を実現する手法として，関係多様化によって生じる属性間の関係の曖昧化を抑止する手法を提案した．評価実験では，提案手法がセンシティブ属性間の関係多様性を小さな曖昧化で実現でき，ナイーブな手法との比較におい

て 10 倍から 100 倍という高い効率性を有することが示された．また，データ分析時にデータの操作に工夫を行うことで，関係多様化によって元のデータセットと異なるデータ形式に成ってしまったとしても，属性値間の共起頻度の計算を小さな誤差で実現できることを確認した．本研究の成果によって，系列データを活用する際に問題となる属性値の過度な汎化，もしくは属性間の関係の過度な曖昧化の両方を抑止したデータ匿名化が実現できる．今後は，提案方式の改良を継続していく．特に，ノイズレスな関係多様化を実現可能なセンシティブ属性値の間には，2 部クリークが張られていることがわかっており，この性質を活用した手法の実現を目指す．また，本研究で導入した関係ノイズ比 RNR は, クラス単位の局所的な視点に基づくものであり，テーブル全体の関係とノイズの比率を全域的な視点での評価したものではない．全域的な視点での関係ノイズ比の定義も今後の課題の一つである．

本研究の提案手法を用いることで，様々な種類の系列データの活用を安全に実現することができ，特にリアルタイムな活用や，複数のセンシティブ属性を持つデータをオリジナルに近い状態での活用が実現できる．これによってリアルタイムな移動パターン検出や，精度が維持された月次集計・パターン分析が実現され，パーソナルデータの活用シーンの拡大が期待される．

謝辞

本研究の遂行ならびに論文の作成にあたり，筑波大学大学院システム情報工学研究科の北川博之教授には懇切なるご指導を賜りました．ここに謹んで感謝の意を表します．また，筑波大学大学院システム情報工学研究科の岡本栄司教授，加藤和彦教授，佐久間淳准教授，天笠俊之准教授には，本論文審査におきまして副査を快く引き受けて下さり，本論文の内容についてご指導とご助言を賜りましたことを深く御礼申し上げます．

本研究は日本電気株式会社において，多くの方々のご指導とご協力を得て行った成果に基づいています．日本電気株式会社に在籍したまま社会人博士課程への就学を許可頂き，また便宜を図って頂いた，クラウドシステム研究所 所長 西原基夫氏，情報・ナレッジ研究所 所長 野口誠氏，クラウドシステム研究所 研究部長 宮内幸司 氏，クラウドシステム研究所 主任研究員 森拓也氏に心から御礼申し上げます．また，本研究を進めるにあたり，日本電気株式会社のグリーンプラットフォーム研究所 宮川伸也氏，クラウドシステム研究所 側高幸治氏，竹之内隆夫氏には，様々なご助言を頂きました．研究活動を進める上でご指導頂き，様々なご支援を頂いた日本電気株式会社およびNECシステムテクノロジー株式会社の皆様にも御礼申し上げます．

投稿した論文等に対して，国内外の多くの査読者から様々なコメントを頂きました．深く感謝いたします．さらに，学会や勉強会等で議論を交わして下さった研究者の皆

様，特に，データベースコミュニティの皆様と情報セキュリティコミュニティの皆様に御礼申し上げます．

北川データ工学研究室に配属以来，川島英之講師，早瀬康裕助教，渡辺知恵美助教より多くの有益なご助言を賜りました．ここに厚く御礼申し上げます．北川データ工学研究室の学生の皆様や，OB/OG の皆様には多大なご支援を頂きました．誠にありがとうございます．特に，研究室の先輩として様々なアドバイスやご支援をいただきました東京工業大学の渡辺陽介助教に心から感謝いたします．

最後に，日々の研究活動を心身両面に渡って支えてくれた家族や友人に心から感謝します．

参考文献

[1] Osman Abul, Francesco Bonchi, and Mirco Nanni. Never walk alone: Uncertainty for anonymity in moving objects databases. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pp. 376–385. IEEE, 2008.

[2] Charu C. Aggarwal and Philip S. Yu. *Privacy-Preserving Data Mining: Models and Algorithms*. Springer-Verlag, 2008.

[3] Gagan Aggarwal, Mayank Bawa, Prasanna Ganesan, Hector Garcia-Molina, Krishnaram Kenthapadi, Rajeev Motwani, Utkarsh Srivastava, Dilys Thomas, and Ying Xu. Two can keep a secret: A distributed architecture for secure database services. *CIDR 2005*, 2005.

[4] Gagan Aggarwal, Nina Mishra, and Benny Pinkas. Secure computation of the kth-ranked element. In Avdances in Cryptology - Proc. of Eurocyrpt ' 04, pp. 40–55. Springer-Verlag, 2004.

[5] Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. In *Proc. SIGMOD'00*, pp. 439–450. ACM, 2000.

[6] Mário S Alvim, Miguel E Andrés, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi. Differential privacy: on the trade-off between utility and information leakage. In *Formal Aspects of Security and Trust*, pp. 39–54. Springer, 2012.

[7] Gennady Andrienko, Natalia Andrienko, Fosca Giannotti, Anna Monreale, and Dino Pedreschi. Movement data anonymity through generalization. In *Proceedings of the 2nd SIGSPATIAL ACM GIS 2009 International Workshop on Security and Privacy in GIS and LBS*, pp. 27–31. ACM, 2009.

[8] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[9] Gilad Asharov and Yehuda Lindell. A full proof of the bgw protocol for perfectly-secure multiparty computation. *IACR Cryptology ePrint Archive*, pp. 136–136, 2011.

[10] Michael Ben-Or, Shafi Goldwasser, and Avi Wigderson. Completeness theorems for non-cryptographic fault-tolerant distributed computation. In *Proc. STOC '88*, pp. 1–10. ACM, 1988.

[11] Alastair R Beresford and Frank Stajano. Mix zones: User privacy in location-aware services. In *Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second IEEE Annual Conference on*, pp. 127–131. IEEE, 2004.

[12] Francesco Bonchi, Laks VS Lakshmanan, and Hui Wendy Wang. Trajectory anonymity in publishing personal mobility data. *ACM SIGKDD Explorations Newsletter*, Vol. 13, No. 1, pp. 30–42, 2011.

[13] Jianneng Cao, Panagiotis Karras, Chedy Raïssi, and Kian-Lee Tan. ρ-uncertainty: inference-proof transaction anonymization. *Proceedings of the VLDB Endowment*, Vol. 3, No. 1-2, pp. 1033–1044, 2010.

[14] University of Tokyo Center for Spatial Information Science. People flow project (pflow). `http://pflow.csis.u-tokyo.ac.jp/index.html`.

[15] Rui Chen, Noman Mohammed, Benjamin CM Fung, Bipin C Desai, and Li Xiong. Publishing set-valued data via differential privacy. *Proceedings of the VLDB Endowment*, Vol. 4, No. 11, pp. 1087–1098, 2011.

[16] Benny Chor, Eyal Kushilevitz, Oded Goldreich, and Madhu Sudan. Private information retrieval. *Journal of the ACM (JACM)*, Vol. 45, No. 6, pp. 965–981, 1998.

[17] Chi-Yin Chow and Mohamed F Mokbel. Trajectory privacy in location-based services and data publication. *ACM SIGKDD Explorations Newsletter*, Vol. 13, No. 1, pp. 19–29, 2011.

[18] Chris Clifton. Privately computing a distributed knn classifier. In *In Proceedings of the Eighth European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2004.

[19] William HE Day and Herbert Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, Vol. 1, No. 1, pp. 7–24, 1984.

[20] Josep Domingo-Ferrer, Maria Bras-Amorós, Qianhong Wu, and Jesús Manjón. User-private information retrieval based on a peer-to-peer community. *Data & Knowledge Engineering*, Vol. 68, No. 11, pp. 1237–1252, 2009.

[21] Cynthia Dwork. Differential privacy. In *Automata, languages and programming*, pp. 1–12. Springer, 2006.

[22] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the 41st annual ACM symposium on Theory of computing*, pp. 371–380. ACM, 2009.

[23] Cynthia Dwork and Kobbi Nissim. Privacy-preserving datamining on vertically partitioned databases. In *In CRYPTO*, pp. 528–544. Springer, 2004.

[24] B.C.M. Fung, K. Wang, A.W.C. Fu, and P.S. Yu. *Privacy-Preserving Data Publishing: Concepts and Techniques*, chapter 11–12. CRC Press, 2010.

[25] Benjamin CM Fung, Ke Wang, and Philip S Yu. Top-down specialization for information and privacy preservation. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, pp. 205–216. IEEE, 2005.

[26] Benjamin Fung, Ke Wang, Rui Chen, and Philip S Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (CSUR)*, Vol. 42, No. 4, p. 14, 2010.

[27] Benjamin Fung, Ke Wang, Ada Wai-Chee Fu, and Jian Pei. Anonymity for continuous data publishing. In *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*, pp. 264–275. ACM, 2008.

[28] Ryo Furukawa, Takao Takenouchi, and Takuya Mori. Behavioral tendency obfuscation framework for personalization services. In *Database and Expert Systems Applications*, pp. 289–303. Springer, 2013.

[29] O. Goldreich. Secure multi-party computation, working draft, version 1.3, 2001.

[30] O. Goldreich, S. Micali, and A. Wigderson. How to play any mental game or a completeness theorem for protocols with honest majority. In *Proc. STOC'87*, pp. 218–229. ACM, 1987.

[31] Marco Gruteser and Dirk Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of the 1st international conference on Mobile systems, applications and services*, pp. 31–42. ACM, 2003.

[32] Michael Hay, Vibhor Rastogi, Gerome Miklau, and Dan Suciu. Boosting the accuracy of differentially private histograms through consistency. *Proceedings of the VLDB Endowment*, Vol. 3, No. 1-2, pp. 1021–1032, 2010.

[33] Yeye He and Jeffrey F Naughton. Anonymization of set-valued data via top-down, local generalization. *Proceedings of the VLDB Endowment*, Vol. 2, No. 1, pp. 934–945, 2009.

[34] Wei Jiang and Chris Clifton. Privacy-preserving distributed k-anonymity. In *Proc. DBSec'05*, pp. 166–177. Springer, 2005.

[35] Xiao Jiang, Jun Gao, Tengjiao Wang, and Dongqing Yang. Multiple sensitive association protection in the outsourced database. In *Database Systems for Advanced Applications*, pp. 123–137. Springer, 2010.

[36] Pawel Jurczyk and Li Xiong. Distributed anonymization: Achieving privacy for both data subjects and data providers. In *Proc. DBSec'09*, pp. 191–207. Springer, 2009.

[37] Richard M Karp. *Reducibility among combinatorial problems*. Springer, 1972.

[38] KDDI 株式会社, 株式会社コロプラ. 位置情報ビッグデータを活用した観光動態調査レポートの提供開始について. `http://www.kddi.com/corporate/news_release/2013/1029a/`.

[39] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 49–60. ACM, 2005.

[40] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*, pp. 25–25. IEEE, 2006.

[41] Chao Li, Michael Hay, Vibhor Rastogi, Gerome Miklau, and Andrew McGregor. Optimizing linear counting queries under differential privacy. In *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 123–134. ACM, 2010.

[42] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pp. 106–115. IEEE, 2007.

[43] Yehuda Lindell and Benny Pinkas. Privacy preserving data mining. In *JOURNAL OF CRYPTOLOGY*, pp. 36–54. Springer-Verlag, 2000.

[44] Yehuda Lindell and Benny Pinkas. Secure multiparty computation for privacy-preserving data mining. *Journal of Privacy and Confidentiality*, Vol. 1, pp. 59–98, 2009.

[45] Junqiang Liu and Ke Wang. Anonymizing transaction data by integrating suppression and generalization. In *Advances in Knowledge Discovery and Data Mining*, pp. 171–180. Springer, 2010.

[46] Xinxin Liu, Han Zhao, Miao Pan, Hao Yue, Xiaolin Li, and Yuguang Fang. Traffic-aware multiple mix zone placement for protecting location privacy. In *INFOCOM, 2012 Proceedings IEEE*, pp. 972–980. IEEE, 2012.

[47] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Vol. 1, No. 1, p. 3, 2007.

[48] Frank D McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pp. 19–30. ACM, 2009.

[49] Adam Meyerson and Ryan Williams. On the complexity of optimal k-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 223–228. ACM, 2004.

[50] Shinya Miyakawa, Nobuyuki Saji, and Takuya Mori. Location l-diversity against multifarious inference attacks. In *Applications and the Internet (SAINT), 2012 IEEE/IPSJ 12th International Symposium on*, pp. 1–10. IEEE, 2012.

[51] N. Mohammed, B. C. M. Fung, K. Wang, and P. C. K. Hung. Privacy-preserving data mashup. In *Proc. EDBT'09*, pp. 228–239. ACM, 2009.

[52] Noman Mohammed, Rui Chen, Benjamin Fung, and Philip S Yu. Differentially private data release for data mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 493–501. ACM, 2011.

[53] Noman Mohammed, Benjamin Fung, and Mourad Debbabi. Walking in the crowd: anonymizing trajectory data for pattern analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 1441–1444. ACM, 2009.

[54] Noman Mohammed, Benjamin Fung, Patrick CK Hung, and Cheuk-kwong Lee. Anonymizing healthcare data: a case study on the blood transfusion service. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1285–1294. ACM, 2009.

[55] Toshikazu Nakamura, Yoshihide Sekimoto, Tomotaka Usui, and Ryosuke Shibasaki. A study on data assimilation of people flow in kanto urban area. *Proc. of Asia GIS*, 2010.

[56] Mehmet Ercan Nergiz, Maurizio Atzori, and Chris Clifton. Hiding the presence of individuals from shared databases. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pp. 665–676. ACM, 2007.

[57] Mehmet Ercan Nergiz, Maurizio Atzori, and Yucel Saygin. Towards trajectory anonymization: a generalization-based approach. In *Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS*, pp. 52–61. ACM, 2008.

[58] Mehmet Ercan Nergiz, Chris Clifton, and Ahmet Erhan Nergiz. Multirelational k-anonymity. *Knowledge and Data Engineering, IEEE Transactions on*, Vol. 21, No. 8, pp. 1104–1117, 2009.

[59] Aleksandar Nikolov, Kunal Talwar, and Li Zhang. The geometry of differential privacy: the sparse and approximate cases. In *Proceedings of the 45th annual ACM symposium on Symposium on theory of computing*, pp. 351–360. ACM, 2013.

[60] Clark F Olson. Parallel algorithms for hierarchical clustering. *Parallel computing*, Vol. 21, No. 8, pp. 1313–1325, 1995.

[61] Femi Olumofin and Ian Goldberg. Revisiting the computational practicality of private information retrieval. In *Financial Cryptography and Data Security*, pp. 158–172. Springer, 2012.

[62] Balaji Palanisamy and Ling Liu. Mobimix: Protecting location privacy with mix-zones over road networks. In *Data Engineering (ICDE), 2011 IEEE 27th International Conference on*, pp. 494–505. IEEE, 2011.

[63] Hwee Hwa Pang, Xiaokui Xiao, and Jialie Shen. Obfuscating the topical intention in enterprise text search. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pp. 1168–1179. IEEE, 2012.

[64] Giorgos Poulis, Grigorios Loukides, Aris Gkoulalas-Divanis, and Spiros Skiadopoulos. Anonymizing data with relational and transaction attributes. In *Machine Learning and Knowledge Discovery in Databases*, pp. 353–369. Springer, 2013.

[65] Pierangela Samarati. Protecting respondents identities in microdata release. *Knowledge and Data Engineering, IEEE Transactions on*, Vol. 13, No. 6, pp. 1010–1027, 2001.

[66] Latanya Sweeney. Guaranteeing anonymity when sharing medical data, the datafly system. In *Proceedings of the AMIA Annual Fall Symposium*, p. 51. American Medical Informatics Association, 1997.

[67] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 05, pp. 557–570, 2002.

[68] Tsubasa Takahashi and Shinya Miyakawa. Cmoa: continuous moving object anonymization. In *Proceedings of the 16th International Database Engineering & Applications Sysmposium*, pp. 81–90. ACM, 2012.

[69] Tsubasa Takahashi, Koji Sobataka, Takao Takenouchi, Yuki Toyoda, Takuya Mori, and Takahide Kohro. Top-down itemset recoding for releasing private complex data. In *Privacy, Security and Trust (PST), 2013 Eleventh Annual International Conference on*, pp. 373–376. IEEE, 2013.

[70] Manolis Terrovitis and Nikos Mamoulis. Privacy preservation in the publication of trajectories. In *Mobile Data Management, 2008. MDM'08. 9th International Conference on*, pp. 65–72. IEEE, 2008.

[71] Manolis Terrovitis, Nikos Mamoulis, and Panos Kalnis. Privacy-preserving anonymization of set-valued data. *Proceedings of the VLDB Endowment*, Vol. 1, No. 1, pp. 115–125, 2008.

[72] U.S. National Archives and Records Administration. Standards for privacy of individually identifiable health information. *Federal Register*, Vol. 67, No. 157, pp. 53182–53273, 2002.

[73] K. Wang, B. C. M. Fung, and G. Dong. Integrating private databases for data analysis. In *Proc. of the 2005 IEEE International Conference on Intelligence and Security Informatics (ISI)*, Vol. 3495, pp. 171–182, 2005.

[74] Ke Wang and Benjamin Fung. Anonymizing sequential releases. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 414–423. ACM, 2006.

[75] Ke Wang, Benjamin CM Fung, and S Yu Philip. Handicapping attacker's confidence: an alternative to k-anonymization. *Knowledge and Information Systems*, Vol. 11, No. 3, pp. 345–368, 2007.

[76] Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai-Chee Fu, and Ke Wang. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 754–759. ACM, 2006.

[77] Xiaokui Xiao and Yufei Tao. Anatomy: Simple and effective privacy preservation. In *Proceedings of the 32nd international conference on Very large data bases*, pp. 139–150. VLDB Endowment, 2006.

[78] Xiaokui Xiao and Yufei Tao. M-invariance: towards privacy preserving republication of dynamic datasets. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pp. 689–700. ACM, 2007.

[79] Xiaokui Xiao and Yufei Tao. Dynamic anonymization: accurate statistical analysis with privacy preservation. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 107–120. ACM, 2008.

[80] Xiaokui Xiao, Guozhang Wang, and Johannes Gehrke. Differential privacy via wavelet transforms. *Knowledge and Data Engineering, IEEE Transactions on*, Vol. 23, No. 8, pp. 1200–1214, 2011.

[81] Xiaokui Xiao, Guozhang Wang, and Johannes Gehrke. Differential privacy via wavelet transforms. *Knowledge and Data Engineering, IEEE Transactions on*, Vol. 23, No. 8, pp. 1200–1214, 2011.

[82] Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi, and Ada Wai-Chee Fu. Utility-based anonymization using local recoding. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 785–790. ACM, 2006.

[83] Yabo Xu, Ke Wang, Ada Wai-Chee Fu, and Philip S Yu. Anonymizing transaction databases for publication. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 767–775. ACM, 2008.

[84] Mingqiang Xue, Panagiotis Karras, Chedy Raïssi, Jaideep Vaidya, and Kian-Lee Tan. Anonymizing set-valued data by nonreciprocal recoding. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1050–1058. ACM, 2012.

[85] Andrew C. Yao. Protocols for secure computations. In *Proc. SFCS'82*, pp. 160–164. IEEE Computer Society, 1982.

[86] Roman Yarovoy, Francesco Bonchi, Laks VS Lakshmanan, and Wendy Hui Wang. Anonymizing moving objects: How to hide a mob in a crowd? In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pp. 72–83. ACM, 2009.

[87] Ganzhao Yuan, Zhenjie Zhang, Marianne Winslett, Xiaokui Xiao, Yin Yang, and Zhifeng Hao. Low-rank mechanism: optimizing batch queries under differential privacy. *Proceedings of the VLDB Endowment*, Vol. 5, No. 11, pp. 1352–1363, 2012.

[88] Justin Zhan, Liwu Chang, and Stan Matwin. Privacy preserving k-nearest neighbor classification. *International Journal of Network Security*, Vol. 1, No. 1, pp. 46–51, 2005.

[89] Bin Zhou, Yi Han, Jian Pei, Bin Jiang, Yufei Tao, and Yan Jia. Continuous privacy preserving publishing of data streams. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pp. 648–659. ACM, 2009.

[90] 厚生労働省医薬品の安全対策における医療関係データベースの活用方策に関する懇談会. 電子化された医療情報データベースの活用による医薬品等の安全・安心に関する提言（日本のセンチネル・プロジェクト）について, 2010.

[91] 株式会社エヌ・ティ・ティ・ドコモ. モバイル空間統計の作成手順. http://www.nttdocomo.co.jp/corporate/technology/rd/tech/main/mobile_spatial_statistics/how_to_produce/.

[92] 経済産業省. 平成 24 年 6 月 1 日ＩＴ融合フォーラム有識者会議 - 配付資料-「ＩＴ融合新産業の創出に向けて」, 2012.

[93] 五十嵐大, 高橋克巳. 注目のプライバシー differential privacy. コンピュータソフトウェア, Vol. 29, No. 4, pp. 40–49, 2012.

[94] 五十嵐大, 千田浩司, 高橋克巳. k-匿名性の確率的指標への拡張とその適用例. コンピュータセキュリティシンポジウム 2009 (CSS2009) 論文集, pp. 763–768, 2009.

[95] 五十嵐大, 千田浩司, 高橋克巳. Pl-多様性：属性推定に対する再構築法のプライバシーの定量化. コンピュータセキュリティシンポジウム 2010 (CSS2010) 論文集, pp. 813–818, 2010.

[96] 五十嵐大, 長谷川聡, 納竜也, 菊池亮, 千田浩司. 数値属性に適用可能な, ランダム化により k-匿名性を保証するプライバシー保護クロス集計. コンピュータセキュリティシンポジウム 2012 論文集, pp. 639–646, oct 2012.

[97] 菊池浩明. データマイニングと個人情報保護. 情報科学技術フォーラム FIT2004 講演論文集, 2004.

[98] 高橋翼, 側高幸治, 豊田由起, 竹之内隆夫, 森拓也. 大規模レセプトに対する匿名化システムの開発. 第 33 回医療情報学連合大会抄録集, 2013.

[99] 社会保険診療報酬支払基金. レセプト電算処理システム. http://www.ssk.or.jp/rezept/index.html.

[100] 佐久間淳, 小林重信. プライバシ保護データマイニング. 人工知能学会誌, Vol. 24, No. 2, pp. 283–294, 2009.

[101] 千田浩司, 五十嵐大, 高橋克巳, 濱田浩気, 菊池亮, 冨士仁. 集合匿名化クラウドの課題と対策 (プライバシー保護,< 小特集 > ビッグデータ時代を支えるセキュリティ・プライバシー保護技術論文). 電子情報通信学会論文誌. A, 基礎・境界, Vol. 96, No. 4, pp. 149–156, apr 2013.

[102] 千田浩司, 木村映善, 五十嵐大, 濱田浩気, 菊池亮, 石原謙. 集合匿名化データの多変量解析評価. コンピュータセキュリティシンポジウム 2012 論文集, pp. 647–654, oct 2012.

[103] 千田浩司, 木村映善, 濱田浩気, 五十嵐大, 高木康彦. 撹乱手法を用いたプライバシー保護医療情報分析の実験評価. 第 31 回医療情報学連合大会抄録集, 2011.

[104] 側高幸治, 高橋翼, 豊田由起, 竹之内隆夫, 森拓也. レセプト匿名化システムの実証と評価. 第 32 回医療情報学連合大会抄録集, 2012.

[105] 側高幸治, 高橋翼, 豊田由起, 竹之内隆夫, 森拓也. センシティブ情報からの個人特定を防ぐレセプト匿名化の検討. 第 33 回医療情報学連合大会抄録集, 2013.

[106] 竹之内隆夫, 川村隆浩, 大須賀昭彦. プライバシ保護データマイニングのための分散匿名化プロトコルの提案. 人工知能学会全国大会 (JSAI2012) 論文集, 2012.

[107] 竹之内隆夫, 川村隆浩, 大須賀昭彦. ユーザ存在/不在確率の範囲を限定した分散匿名化手法と医療データによる評価. コンピュータセキュリティシンポジウム 2012 (CSS2012) 論文集, pp. 525–532, 2012.

[108] 竹之内隆夫, 川村隆浩, 大須賀昭彦. ユーザ存在の特定を困難にした分散匿名化の提案 : 2 診療機関のレセプトデータを用いた有効性の評価 (人工知能, データマイニング,< 特集 > 学生論文). 電子情報通信学会論文誌. D, 情報・システム, Vol. 96, No. 3, pp. 596–610, mar 2013.

研究業績

博士論文に関する論文

査読付き論文誌

- 髙橋翼, 側高幸司, 竹之内隆夫, 豊田由起, 森拓也: センシティブ属性間の関係多様化によるプライバシ保護手法. 情報処理学会論文誌データベース, Vol. 6, No. 5 (TOD60), 2013 年 12 月.
- 髙橋翼, 宮川伸也, 伊東直子: 移動軌跡ストリームに対するリアルタイム k-匿名化手法の提案. 日本データベース学会論文誌, Vol. 10, No. 1, pp. 37–43, 2011 年 6 月.

査読付き国際会議

- Tsubasa Takahashi and Shinya Miyakawa: CMOA: Continuous Moving Object Anonymization. Proceedings of the International Database Engineering & Applications Symposium (IDEAS2012), pp. 81–90, 2012.

研究会発表

- 髙橋翼, 竹之内隆夫, 側高幸治: 時系列データに対する l-多様化方式の提案. 第 4 回データ工学と情報マネジメントに関するフォーラム (DEIM 2012), 2012.
- 髙橋翼, 宮川伸也, 伊東直子: 移動軌跡ストリームに対するリアルタイム k 匿名化手法の提案. 第 3 回データ工学と情報マネジメントに関するフォーラム (DEIM 2011), 2011.

その他論文

査読付き論文誌

- Tsubasa Takahashi, Hiroyuki Kitagawa and Keita Watanabe: Social Bookmarking Induced Active Page Ranking. IEICE Transactions Vol.E93-D, No.6, pp.1403–1413, 2010.
- 山口祐人, 天笠俊之, 髙橋翼, 北川博之: 情報伝搬を考慮したグラフ分析による Twitter ユーザランキング手法, 情報処理学会論文誌データベース， Vol. 4, No. 2 (TOD50), pp.142–157, 2011.
- 竹之内隆夫, 側高幸治, 豊田由起, 髙橋翼, 森拓也: 部分データセットとの突合に対する耐性を有するレセプト匿名化方式. 医療情報学, Vol. 33, No. 3 pp. 127–138, (2013).

査読付き国際会議

- Tsubasa Takahashi, Koji Sobataka, Takao Takenouchi, Yuki Toyoda, Takuya Mori and Takahide Kohro: Top-down Itemset Recoding for Releasing Private Complex Data. Proceedings of the Eleventh Annual Conference on Privacy, Security and Trust (PST2013), pp.373–376, 2013.
- Tsubasa Takahashi and Hiroyuki Kitagawa: A Ranking Method for Web Search using Social Bookmarks. Proceedings of the 14th International Conference on Database Systems for Advanced Applications (DASFAA2009), pp.585–589, 2009.
- Tsubasa Takahashi and Hiroyuki Kitagawa: S-BITS: Social Bookmarking Induced Topic Search. Proceedings of the Ninth International Conference on Web-Age Information Management (WAIM2008), pp.25–30, 2008.
- Yuto Yamaguchi, Tsubasa Takahashi, Toshiyuki Amagasa and Hiroyuki Kitagawa: TURank: Twitter user ranking based on user-tweet graph analysis, Proceedings of the Eleventh International Conference on Web Information Systems Engineering (WISE2010), pp. 240–253, 2010.

査読付き国内会議

- 髙橋翼, 北川博之: ソーシャルブックマークによる情報の鮮度を考慮した Web ページ評価手法. Web とデータベースに関するフォーラム (WebDB Forum 2008), 2008.

研究会発表

- 髙橋翼, 側高幸治, 豊田由起, 竹之内隆夫, 森拓也: 大規模レセプトに対する匿名化システムの開発. 第 33 回医療情報学連合大会, 2013.
- 髙橋翼, 側高幸治, 豊田由起, 竹之内隆夫, 森拓也: 効率的な集合値再符号化手法による複合データの k-匿名化. コンピュータセキュリティシンポジウム 2013 (CSS2013), 2013.
- 髙橋翼, 側高幸治, 豊田由起, 竹之内隆夫, 森拓也, 興梠貴英: 集合値の一般化階層なし再符号化による複合データの k-匿名化. 第 5 回データ工学と情報マネジメントに関するフォーラム (DEIM2013), 2013.
- 髙橋翼, 側高幸治, 豊田由起, 竹之内隆夫, 森拓也, 興梠貴英: 患者識別子の突合による匿名性破綻を解消した匿名化手法. 第 32 回医療情報学連合大会, 2012.
- 髙橋翼, 側高幸治, 豊田由起, 竹之内隆夫, 森拓也, 興梠貴英: 共通識別子を持つレコード群のデータ匿名化手法. コンピュータセキュリティシンポジウム 2012 (CSS2012), 2012.
- 髙橋翼, 北川博之, 渡辺桂太: ソーシャルブックマークにおけるトピック分析と活性度推定に基づく Web ページのランキング. 第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM2010), 2010.
- 髙橋翼, 北川博之: ソーシャルブックマークにおけるブックマークの活性度を考慮した Web ページのランキング. データ工学と情報マネジメントに関するフォーラム (DEIM2009), A4-1, 2009.
- 髙橋翼, 北川博之: ソーシャルブックマークを利用したユーザ嗜好に基づくページの抽出. 情報処理学会 第 70 回全国大会講演論文集 (1), pp.651–652, 2008.(学生奨励賞受賞)

- 髙橋翼, 北川博之: ソーシャルブックマークを利用したユーザ嗜好に基づくページの評価. 電子情報通信学会 第 19 回データ工学ワークショップ (DEWS2008), B8-6, 2008.
- 山口祐人, 髙橋翼, 天笠俊之, 北川博之: リンク構造解析による Twitter ユーザのランキング手法. 情報処理学会創立 50 周年記念 (第 72 回) 全国大会, 2010.
- 渡邊桂太, 髙橋翼, 北川博之: ソーシャルブックマークにおけるユーザ間の類似度を考慮したスパマー検出. データ工学と情報マネジメントに関するフォーラム (DEIM2009), A1-1, 2009.
- 渡邊桂太, 髙橋翼, 天笠俊之, 北川博之: グラフ構造解析を用いたソーシャルブックマークにおけるスパマー検出. 第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM2010), 2010.
- 渡邊桂太, 髙橋翼, 天笠俊之, 北川博之: グラフ解析に基づくソーシャルブックマークにおけるスパマー検出. 情報処理学会創立 50 周年記念（第 72 回）全国大会, 2010.
- 側高幸治, 髙橋翼, 豊田由起, 竹之内隆夫, 森拓也: センシティブ情報からの個人特定を防ぐレセプト匿名化の検討. 第 33 回医療情報学連合大会, 2013.
- 側高幸治, 髙橋翼, 豊田由起, 竹之内隆夫, 森拓也, 興梠貴英: レセプト匿名化システムの実証と評価. 第 32 回医療情報学連合大会, 2012.
- 豊田由起, 側高幸治, 髙橋翼, 竹之内隆夫, 森拓也, 興梠貴英: 制約と優先度を考慮したレセプト匿名化方式. 第 32 回医療情報学連合大会, 2012.
- 竹之内隆夫, 側高幸治, 豊田由起, 髙橋翼, 森拓也: 部分データセットとの突合に対する耐性を有するレセプト匿名化方式. 第 32 回医療情報学連合大会, 2012.