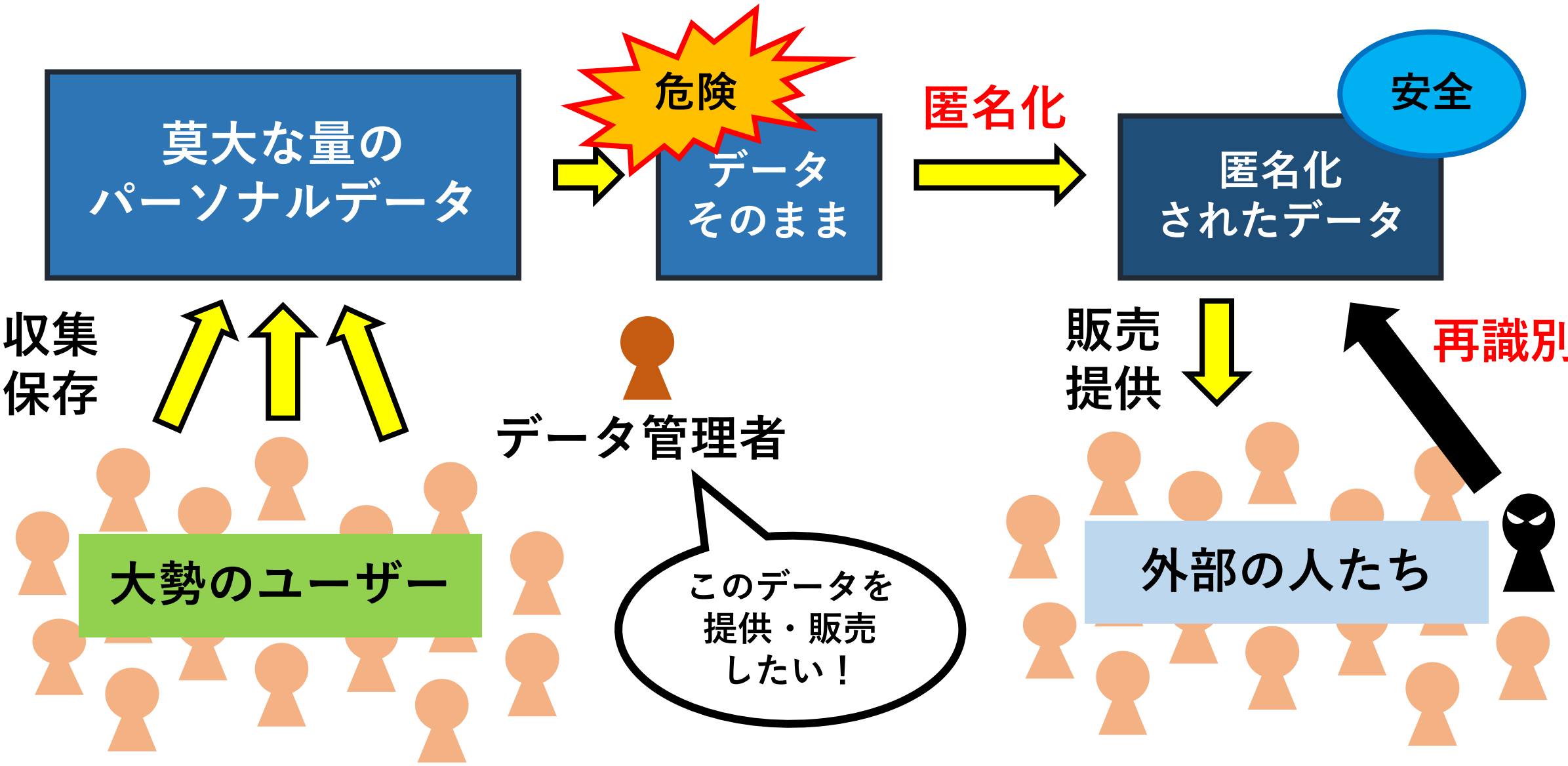
2021年度　明治大学大学院
博士学位請求論文

個人情報の識別リスク評価に基づいた匿名化に関する研究

先端数理科学研究科　先端メディアサイエンス専攻
伊藤　聡志

匿名化とは？


莫大な量の
パーソナルデータ
危険
データ
そのまま
匿名化
安全
匿名化
されたデータ
収集
保存
販売
提供
再識別
データ管理者
このデータを
提供・販売
したい！
大勢のユーザー
外部の人たち

本研究の目的

元のデータ

名前	年齢	郵便番号
Alice	30	10055
Bob	21	10055
Carol	21	10023

全ての値を削除

加工データ１

名前	年齢	郵便番号
-	-	-
-	-	-
-	-	-

安全性：高
有用性：低

名前を仮名に置き換える

加工データ２

仮名	年齢	郵便番号
A	30	10055
B	21	10055
C	21	10023

安全性：低
有用性：高

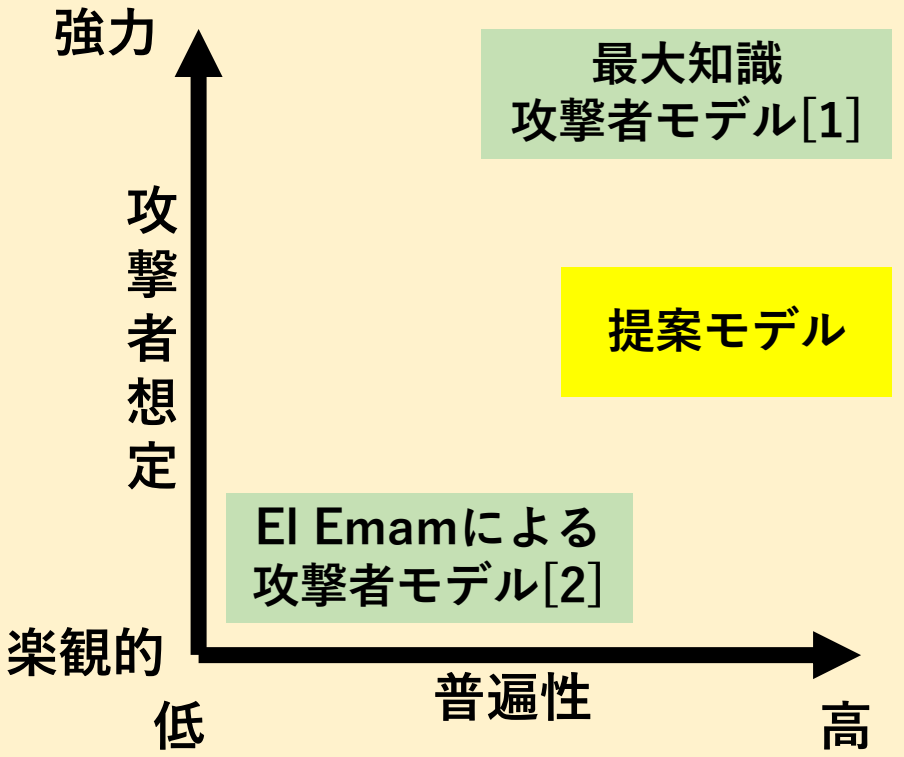
匿名化の影響は
様々な要因によって
大きく変化する

研究目的：データに対する匿名化の影響を明らかにすること

研究内容：目的達成のために，4つの課題を設定し解決する

課題１：既存の攻撃者モデル

課題：既存の攻撃者モデルは強力すぎる/楽観的すぎる



[1] J. Domingo-Ferrer, S. Ricci and J. Soria-Comas, “Disclosure risk assessment via record linkage by a maximum-knowledge attacker”, 2015 13th Annual Conference on Privacy, Security and Trust (PST), Izmir, 2015, pp. 28-35 (2015).
[2] Khaled El Emam, Luk Arbuckle, “Anonymizing Health Data Case Studies and Methods to Get You Started”, O'Reilly, 2013.

最大知識攻撃者モデル[1]

2015年にDomingo-Ferrerらが提案

元データを全て背景知識として持つ攻撃者モデル

→あまりにも強力な攻撃者想定であるためデータの安全性を過度に低く見積もってしまう

Dumber数モデル[2]

2013年にEl Emamらが提案

攻撃者が友人の中に存在することを想定したモデル

平均的な人の友人数であるDumber数を用いる

→攻撃者が知り合いの中にしかいない想定は楽観的

解決：部分的な背景知識を有する新たな攻撃者モデルを提案する

課題２：履歴データの識別リスク

[3] Hiroaki Kikuchi, Katsumi Takahashi, “Ziph Distribution Model for Quantifying Risk of Reidentification from Trajectory Data”, Journal of Information Processing, Vol. 24, No. 5, pp. 816–823, 2016.

課題：動的に変化する履歴データから個人が識別されるリスクの研究は不十分

履歴データ

名前	日付	商品
Alice	12/1	2
Bob	12/2	3
Alice	12/2	1
Carol	12/2	2
Bob	12/3	1
Bob	12/4	4

個人数=3, レコード数=6

履歴データから個人は特定されないという誤解

2013年にJR東日本は，交通系ICカードSuicaの利用履歴データを個人が特定できる情報ではないとみなし，そのデータを他社に提供した.

→2016年の菊池らの研究[3]により，高々3駅分の履歴から98%の個人が識別されることが判明した.

動的に変化するイベントの正確な定式化の難しさが履歴データのリスク研究の妨げになっている.

解決：いくつかの仮定を置いて，履歴データのふるまいを数理モデル化する

課題 3：k-anonymityの問題点

[4] L. Sweeny, “k-anonymity: a model for protecting privacy”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), pp.557--570. (2006)

k-anonymity：Sweeneyによって提案された匿名性指標[4]．データ中の最低でもk人の区別がつかないとき，そのデータはk-anonymityを満たす．
k-anonymityを満たすようにデータを加工することは**k-匿名化**と呼ばれている．

元データ

名前	年齢	郵便番号
Alice	30	10055
Bob	25	10055
Carol	21	10023
David	55	10165
Eve	47	10224

2-匿名化された加工データ

仮名	年齢	郵便番号
1	21-30	100**
2	21-30	100**
3	21-30	100**
4	47-55	10***
5	47-55	10***

3人の区別がつかないグループに属する個人よりも

2人の区別がつかないグループに属する個人の方が危険(不公平)

「最低でも2人の区別がつかない」状態(2-匿名化)のためには，3人のグループは過度な加工ではないか？
データ中の個人によって安全性に差があるのは不公平ではないか？

課題：既存のk-匿名化では過度な加工や安全性の不公平さが生じてしまう

解決：k-concealmentという指標に注目した匿名化手法を提案する

課題 4：実験データへの依存性

課題：データに対する匿名化の影響は，対象となるデータに大きく依存する

解決：多種多様なデータに対して実験的評価を行う

ID	内容	個人数	レコード数 (行)	属性数 (列)	扱う章	データの種類
1	購買履歴	400	38,087	7	3,4,5,9	オープンデータ
2	健康診断	198,740	964,636	49	6	匿名加工情報
3	交通 IC カード	31	584	10	7	個人データ (同意取得)
4	世帯支出	8,333	8,333	25	8	合成データ
5	糖尿病患者	71,518	101,766	50	3	オープンデータ
6	世帯収入	32,561	32,561	16	3,10	オープンデータ
7	ローン借入	42,538	42,538	145	3	オープンデータ
8	疑似人流	6,432	901,465	9	9,10	合成データ
9	傷病レセプト	288,568	39,363,878	15	6	匿名加工情報
10	医薬品レセプト	279,199	31,465,504	21	6	匿名加工情報

本研究の概要

研究目的：データに対する匿名化の影響を明らかにすること

課題１：既存の攻撃者モデル

課題２：履歴データの識別リスク

課題３：k-anonymityの問題点

課題４：実験データへの依存性

解決１：部分的な背景知識を有する新たな攻撃者モデルを提案する

解決２：いくつかの仮定を置いて，履歴データのふるまいを数理モデル化する

解決３：k-concealmentという指標に注目した匿名化手法を提案する

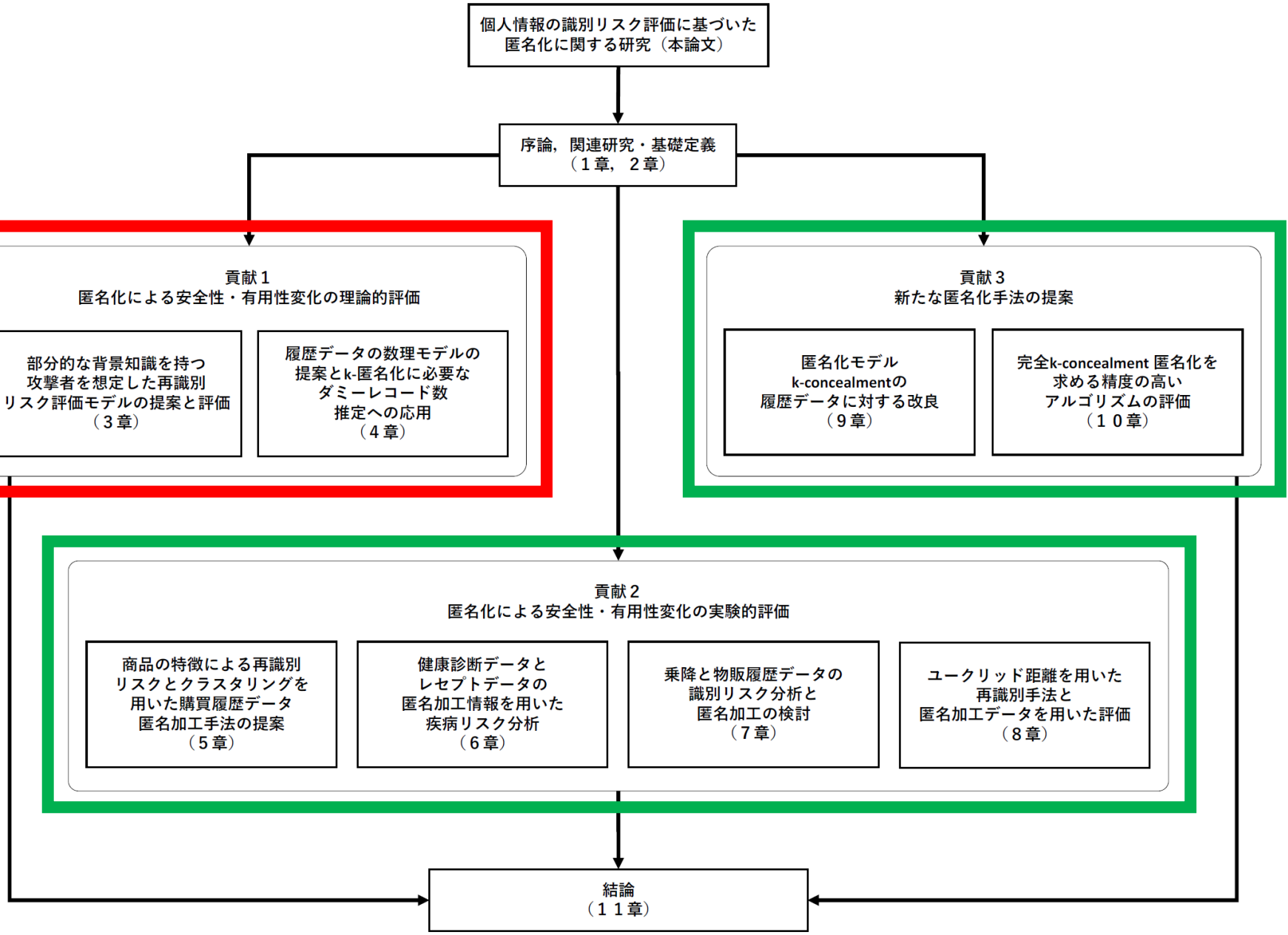
解決４：多種多様なデータに対して実験的評価を行う

貢献１：匿名化による安全性・有用性変化の理論的評価
貢献２：匿名化による安全性・有用性変化の実験的評価
貢献３：新たな匿名化手法の提案

博士論文の構成と本発表

本発表では
博士論文のうち
貢献 1 の部分を
メインに説明する

貢献 2, 3 の部分は
簡単に紹介する



目次

1. 部分的な背景知識を持つ攻撃者を想定した
   再識別リスク評価モデルの提案と評価（3 章）

2. 履歴データの数理モデルの提案とk-匿名化に必要な
   ダミーレコード数推定への応用（4 章）

3. 貢献 2 の紹介（5~8 章）

4. 貢献 3 の紹介（9~10 章）

目次

1. 部分的な背景知識を持つ攻撃者を想定した
   再識別リスク評価モデルの提案と評価（3 章）

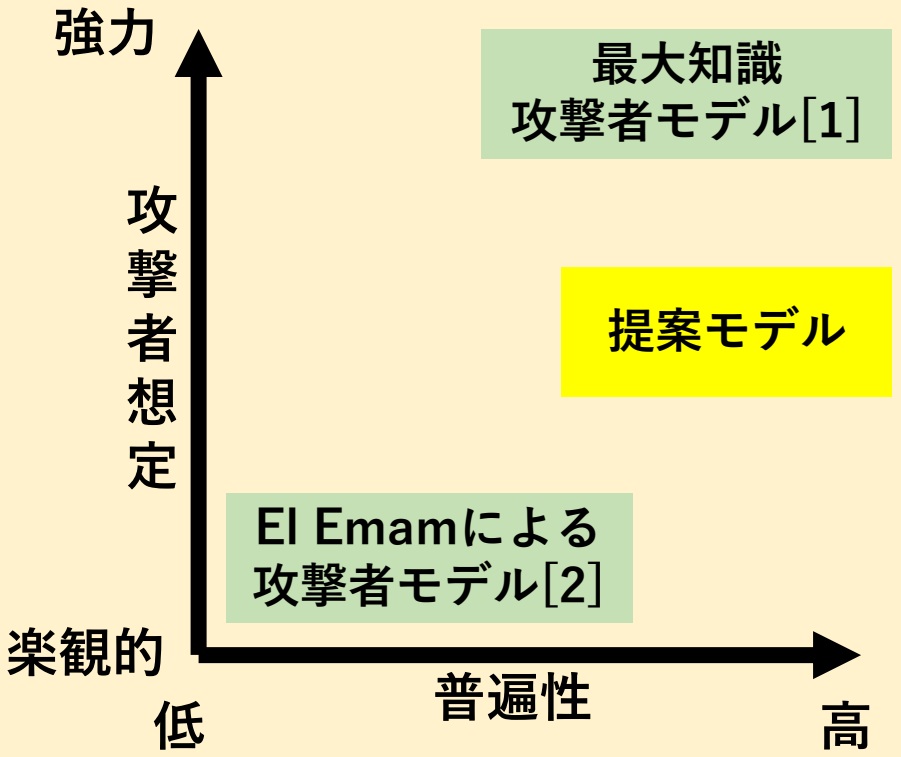
2. 履歴データの数理モデルの提案とk-匿名化に必要な
   ダミーレコード数推定への応用（4 章）

3. 貢献 2 の紹介（5~8 章）

4. 貢献 3 の紹介（9~10 章）

課題１：既存の攻撃者モデル(再掲)

課題：既存の攻撃者モデルは強力すぎる/楽観的すぎる



[1] J. Domingo-Ferrer, S. Ricci and J. Soria-Comas, “Disclosure risk assessment via record linkage by a maximum-knowledge attacker”, 2015 13th Annual Conference on Privacy, Security and Trust (PST), Izmir, 2015, pp. 28-35 (2015).
[2] Khaled El Emam, Luk Arbuckle, “Anonymizing Health Data Case Studies and Methods to Get You Started”, O'Reilly, 2013.

最大知識攻撃者モデル[1]

2015年にDomingo-Ferrerらが提案

元データを全て背景知識として持つ攻撃者モデル

→あまりにも強力な攻撃者想定であるためデータの安全性を過度に低く見積もってしまう

Dumber数モデル[2]

2013年にEl Emamらが提案

攻撃者が友人の中に存在することを想定したモデル

平均的な人の友人数であるDumber数を用いる

→攻撃者が知り合いの中にしかいない想定は楽観的

解決：部分的な背景知識を有する新たな攻撃者モデルを提案する

部分的な背景知識を持つ攻撃者

元データ

真名	属性A
伊藤	2019/2/1
伊藤	2019/2/1
山田	2019/2/2
岡本	2019/2/2
岡本	2019/2/3

確率$Pr(\mathrm{idf}|a)$で
再識別に成功する

加工データ

仮名	属性A
100	2019/2/1
100	2019/2/1
200	2019/2/2
300	2019/2/2
300	2019/2/3

攻撃者

確率$Pr(a)$である顧客の
属性Aについての
部分的な背景知識aを得る

背景知識 a の危険度

$$Pr(\mathrm{idf}, a) = Pr(a)Pr(\mathrm{idf}|a)$$

例：背景知識 a =“2019/2/2” の場合

元データ

真名	購買日
伊藤	2019/2/1
伊藤	2019/2/1
山田	2019/2/2
岡本	2019/2/2
岡本	2019/2/3

加工データ

仮名	購買日
100	2019/2/1
100	2019/2/1
200	2019/2/2
300	2019/2/2
300	2019/2/3

$Pr(\mathrm{idf}|a) = 1/2$の
確率であるユーザの
再識別に成功する

攻撃者

$Pr(a) = 2/5$の確率で
「あるユーザが2019/2/2に
買い物をした」
という背景知識を得る

背景知識 a の危険度

$$Pr(\mathrm{idf}, a) = \frac{1}{2} \cdot \frac{2}{5} = \frac{1}{5}$$

部分的な背景知識を持つ攻撃者の危険度

元データ

真名	属性A
伊藤	2019/2/1
伊藤	2019/2/1
山田	2019/2/2
岡本	2019/2/2
岡本	2019/2/3

確率$Pr(\mathrm{idf}|a)$で
再識別に成功する

加工データ

仮名	属性A
100	2019/2/1
100	2019/2/1
200	2019/2/2
300	2019/2/2
300	2019/2/3

攻撃者

確率$Pr(a)$である顧客の
属性Aについての
部分的な背景知識xを得る

本章では，　$Pr(\mathrm{idf}|a)$の
期待値（平均識別確率）を
攻撃者の危険度と定義する

$$Pr(\mathrm{idf}, A) = \sum Pr(\mathrm{idf}, a)$$

平均識別確率と平均レコード数 α_a (1)

α_a：aについての平均レコード数

m：データセットのレコード数

$$\alpha_a = \frac{a\text{を満たすレコードの数}}{a\text{を満たすユーザの数}}$$

例：購買履歴データ

真名	購買日
伊藤	2019/2/1
伊藤	2019/2/1
山田	2019/2/2
岡本	2019/2/2
岡本	2019/2/3

a =“2019/2/1”のとき $\alpha_a = \frac{2}{1} = 2$

a =“2019/2/2”のとき $\alpha_a = \frac{2}{2} = 1$

a =“2019/2/3”のとき $\alpha_a = \frac{1}{1} = 1$

$m = 5$

平均識別確率と平均レコード数 $\boldsymbol{\alpha_a}$ (2)

このとき，平均識別確率は $Pr(\mathbf{idf}, A) = \sum \frac{\alpha_a}{m}$ と表せる．

例：購買履歴データ

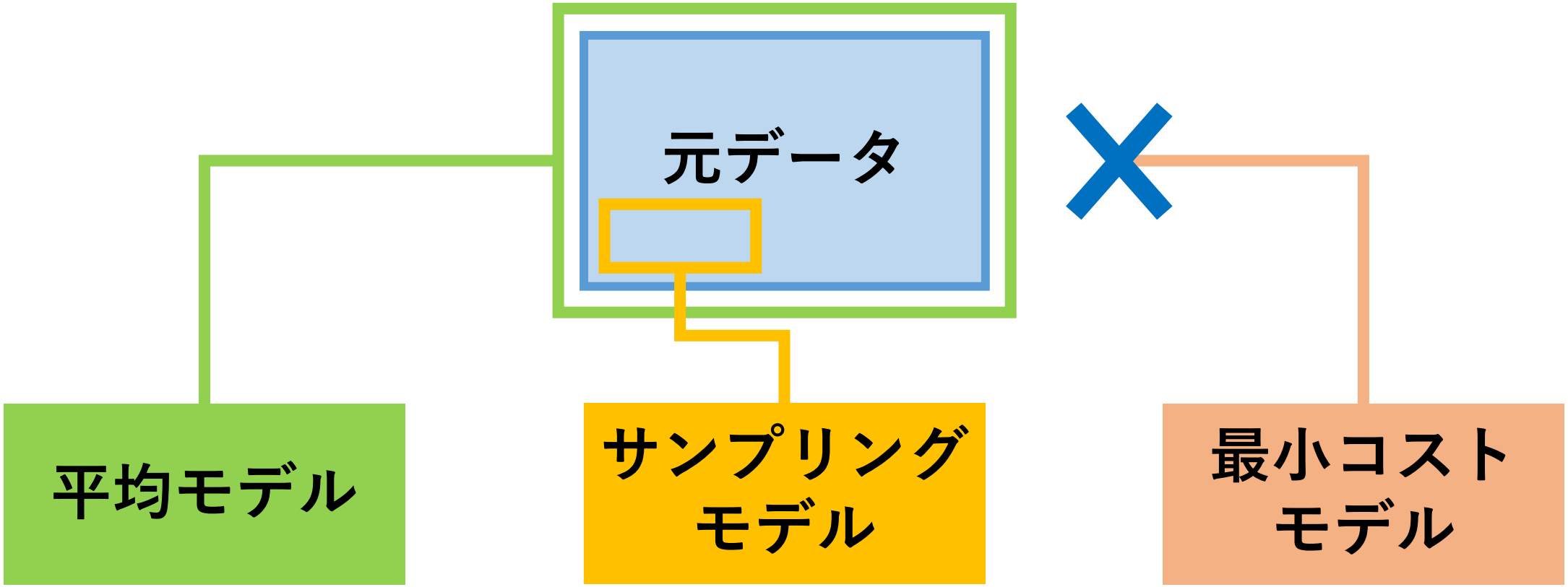
真名	購買日
伊藤	2019/2/1
伊藤	2019/2/1
山田	2019/2/2
岡本	2019/2/2
岡本	2019/2/3

$$Pr(\mathbf{idf}, A) = \sum \frac{\alpha_a}{m} = \frac{2}{5} + \frac{1}{5} + \frac{1}{5} = \frac{4}{5}$$

しかし，ビッグデータについて全てのα_aを計算するのは時間がかかるため，これを近似してリスク評価を行うモデルを提案する．

※38,087レコード，6属性のデータのリスクを計算するのに27.5秒かかる(後述)

平均識別確率を近似する３つのモデル



データの全体から
α_aの平均値α_Aを求め
リスクを計算する

データの一部から
α_aの平均値$\alpha_{a'}$を求め
リスクを計算する

全てのaについて
$\alpha_a = 1$として
リスクを計算する

サンプリングモデル

購買履歴データ
（5レコード, 3種類）

真名	購買日
伊藤	2019/2/1
伊藤	2019/2/1
山田	2019/2/2
岡本	2019/2/2
岡本	2019/2/3

5レコードから
2レコードを
サンプリング
(一様サンプリング)

サンプリング
サイズ 2

3種類の値から
2種類の値を
サンプリング
（層別サンプリング）

本研究

真名	購買日
伊藤	2019/2/1
山田	2019/2/2

真名	購買日
伊藤	2019/2/1
伊藤	2019/2/1
山田	2019/2/2
岡本	2019/2/2

各モデルの式

平均識別確率 $$Pr(\mathrm{idf}, A) = R(A) = \sum \frac{\alpha_a}{m}$$

1. 平均モデル $$\mathrm{R_{mean}}(A) = \sum \frac{\alpha_A}{m}$$

2. 最小コストモデル $$\mathrm{R_{cost}}(A) = \sum \frac{1}{m}$$

3. サンプリングモデル $$\mathrm{R_{sample}}(A) = \sum \frac{\alpha_{a'}}{m}$$

提案モデルの誤差率

定理 3.3.1：最小コストモデルの誤差率は$|1-1/\alpha_A|$である.

$$\frac{|R_{cost}(A)-R(A)|}{R(A)}=\frac{\left|\frac{|D_A|}{m}-\frac{\alpha_A|D_A|}{m}\right|}{\frac{\alpha_A|D_A|}{m}}=\left|1-\frac{1}{\alpha_A}\right|$$

定理 3.3.2：サンプリングモデルの誤差率の最大値は$\frac{\sigma_s m}{\sqrt{s}|D_A|\alpha_A}$である.

$$\frac{\left|R_{sample}(A)-R(A)\right|}{R(A)}<\frac{\frac{\sigma_s}{\sqrt{s}}}{\frac{\alpha_A|D_A|}{m}}=\frac{\sigma_s m}{\sqrt{s}|D_A|\alpha_A}$$

評価実験

提案した３つのモデルで以下の４データのリスク評価をした

T_1 : Online Retail Dataset
英国の１年間の小売データ

T_2 : Diabetes Dataset
糖尿病患者の入院履歴データ

T_3 : Adult Dataset
国勢調査による世帯収入データ

T_4 : LOAN DATA
2007年から2011年の間のローン借り入れデータ

各データの大きさ

	レコード数	ユーザ数	属性数
T_1	38,087	400	7
T_2	101,766	71,518	50
T_3	32,561	32,561	16
T_4	42,538	42,538	145

実験には
一部の属性を
用いる

各モデルの評価結果

※サンプリングサイズは10
※90%信頼区間

データ	属性	平均識別確率	平均モデル	最小コストモデル	サンプリングモデル
T_1	購買時刻	0.3217	0.3217	0.0145	[0.1411, 0.5998]
	購買日	0.1860	0.1860	0.0076	[0.1267, 0.2786]
	購買商品	0.0965	0.0965	0.0730	[0.0718, 0.0982]
	価格	0.0121	0.0121	0.0048	[0.0036, 0.0132]
	個数	0.0080	0.0080	0.0025	[0.0017, 0.0152]
T_2	入院日数	1.45E-04	1.45E-04	1.38E-04	[1.46E-04, 1.52E-04]
	年齢	1.33E-04	1.33E-04	9.83E-05	[1.21E-04, 1.42E-04]
	人種	7.73E-05	7.73E-05	5.90E-05	[6.92E-05, 8.31E-05]
	性別	3.78E-05	3.78E-05	2.95E-05	[3.08E-05, 4.30E-05]
T_3	年齢	2.24E-03	2.24E-03	2.24E-03	[2.24E-03, 2.24E-03]
	職業	4.61E-04	4.61E-04	4.61E-04	[4.61E-04, 4.61E-04]
	婚姻状況	2.15E-04	2.15E-04	2.15E-04	[2.15E-04, 2.15E-04]
	人種	1.54E-04	1.54E-04	1.54E-04	[1.54E-04, 1.54E-04]
T_4	職業	0.7208	0.7208	0.7208	[0.7208, 0.7208]
	収入	0.1316	0.1316	0.1316	[0.1316, 0.1316]
	ローン量	0.0211	0.0211	0.0211	[0.0211, 0.0211]
	ローンクラス	0.0002	0.0002	0.0002	[0.0002, 0.0002]

各モデルの評価結果

※サンプリングサイズは10
※90%信頼区間

データ	属性	平均識別確率	平均モデル	最小コストモデル	サンプリングモデル
T_1	購買時刻	0.3217	0.3217	0.0145	[0.1411, 0.5998]
	購買日	0.1860	0.1860	0.0076	[0.1267, 0.2786]
	購買商品	0.0965	0.0965	0.0	[0.0718, 0.0982]
	価格	0.0121	0.0121	0.0028	[0.0068, 0.0202]
	個数	0.0080	0.0080	0.0025	[0.0068, 0.0202]
T_2	入院日数	1.45E-04	1.45E-04	1.38E-04	[1.46E-04, 1.52E-04]
	年齢	1.33E-04	1.33E-04	9.83E-05	[1.21E-04, 1.42E-04]
	人種	7.73E-05	7.73E-05	5.90E-05	[6.92E-05, 9.53E-05]
	性別	3.78E-05	3.78E-05	2.95E-05	[3.07E-05, 4.15E-05]
T_3	年齢	2.24E-03	2.24E-03	2.24E-03	[2.24E-03, 2.24E-03]
	職業	4.61E-04	4.61E-04	4.61E-04	[4.61E-04, 4.61E-04]
	婚姻状況	2.15E-04	2.15E-04	2.15E-04	[2.15E-04, 2.15E-04]
	人種	1.54E-04	1.54E-04	1.54E-04	[1.54E-04, 1.54E-04]
T_4	職業	0.7208	0.7208	0.7208	[0.7208, 0.7208]
	収入	0.1316	0.1316	0.1316	[0.1316, 0.1316]
	ローン量	0.0211	0.0211	0.0211	[0.0211, 0.0211]
	ローンクラス	0.0002	0.0002	0.0002	[0.0002, 0.0002]

平均識別確率を用いることにより
どの属性が危険であるかを
判断することができる

T_1：購買時刻，T_2：入院日数
T_3：年齢，T_4：職業

匿名化をする際に
どの属性を優先的に加工するか
決めることができる

各モデルの評価結果

※サンプリングサイズは10
※90%信頼区間

データ	属性	平均識別確率	平均モデル	最小コストモデル	サンプリングモデル
	購買時刻	0.3217	0.3217	0.0145	[0.1411, 0.5998]
	購買日	0.1860	0.1860	0.0076	[0.1267, 0.2786]
T_1	購買商品	0.0965	0.0965	0.0730	[0.0718, 0.0982]
	価格	0.0121	0.0121	0.0048	[0.0036, 0.0132]
	個数	0.0080	0.0080	0.0025	[0.0017, 0.0152]
	入院日数	1.45E-04	1.45E-04	1.38E-04	[1.46E-04, 1.52E-04]
	年齢	1.03E-04	1.03E-04	0.83E-05	[1.21E-04, 1.42E-04]
T_2	人種	7.73E-05	7.73E-05	5.90E-05	[6.92E-05, 8.31E-05]
	性別	3.69E-05	3.69E-05	2.95E-05	[3.08E-05, 4.30E-05]
	年齢	2.24E-03	2.24E-03	2.24E-03	[2.24E-03, 2.24E-03]
	職業	4.61E-04	4.61E-04	4.61E-04	[4.61E-04, 4.61E-04]
T_3	婚姻状況	2.15E-04	2.15E-04	2.15E-04	[2.15E-04, 2.15E-04]
	人種	1.54E-04	1.54E-04	1.54E-04	[1.54E-04, 1.54E-04]
	性別	0.7208	0.7208	0.7208	[0.7208, 0.7208]
	収入	0.1316	0.1316	0.1316	[0.1316, 0.1316]
T_4	ローン量	0.0211	0.0211	0.0211	[0.0211, 0.0211]
	ローンクラス	0.0002	0.0002	0.0002	[0.0002, 0.0002]

評価値はサンプリングの結果によって変化する

ここではサンプリングサイズ10のときの90%信頼区間を示している（10レコードではなく，10種類のa）

データの一部しか用いていないが属性の危険度を精度よく評価できる

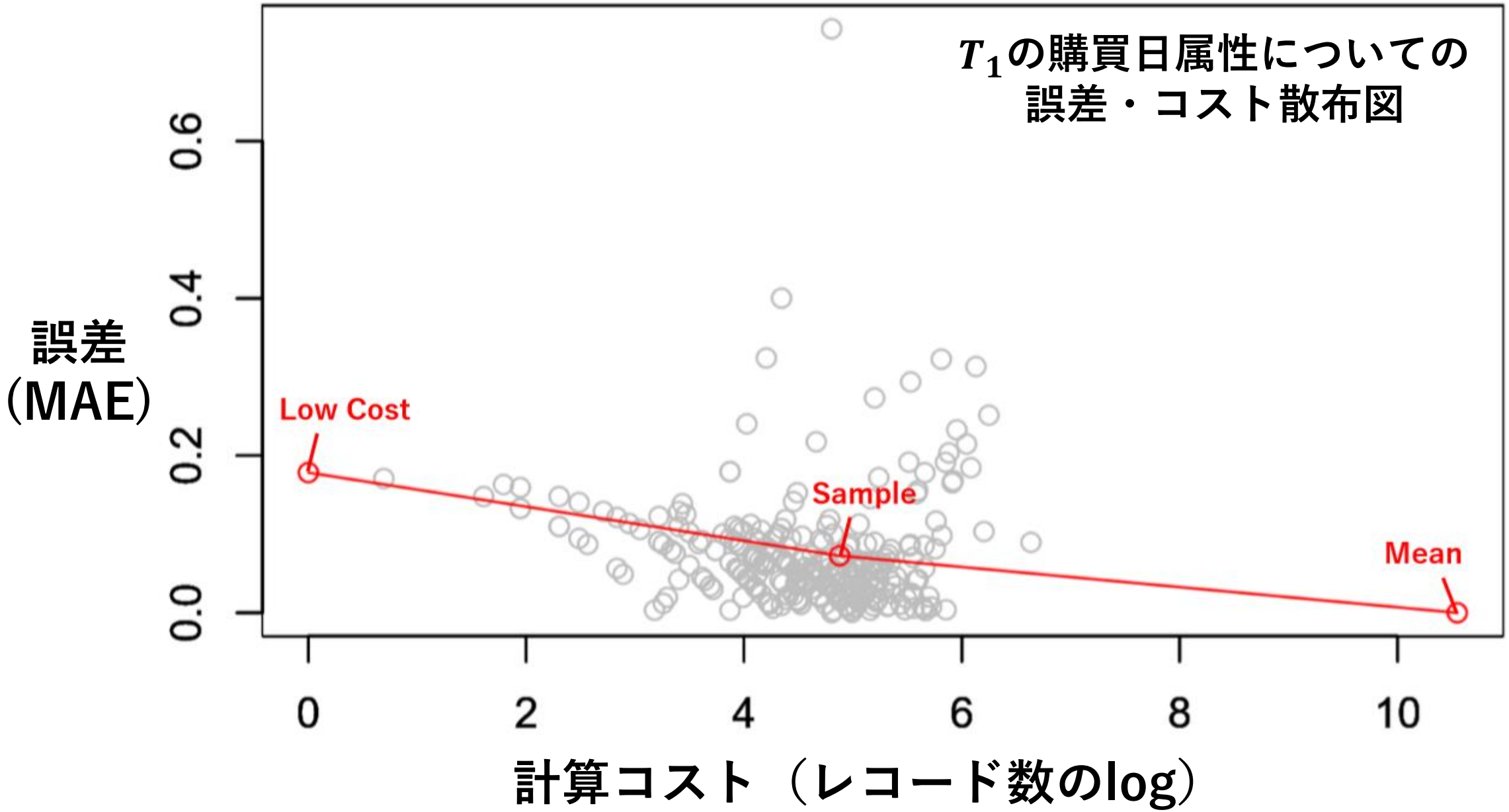
各モデルの評価結果

※サンプリングサイズは10
※90%信頼区間

データ	属性	平均識別確率	平均モデル	最小コストモデル	サンプリングモデル
T_1	購買時刻	0.3217	0.3217	0.0145	[0.1411, 0.5998]
	購買日	0.1860	0.1860	0.0076	[0.1267, 0.2786]
	購買商品	0.0965	0.0965	0.0730	[0.0718, 0.0982]
	価格	0.0121	0.0121	0.0048	[0.0036, 0.0132]
	個数	0.0080	0.0080	0.0025	[0.0017, 0.0152]
T_2	入院日数	1.45E-04	1.45E-04	1.38E-04	[1.46E-04, 1.52E-04]
	年齢	1.33E-04	1.33E-04	9.83E-05	[1.21E-04, 1.42E-04]
	人種	7.73E-05	7.73E-05	5.90E-05	[6.92E-05, 8.31E-05]
	性別	3.78E-05	3.78E-05	2.95E-05	[3.08E-05, 4.30E-05]
T_3	年齢	2.24E-03	2.24E-03	2.24E-03	[2.24E-03, 2.24E-03]
	職業	4.61E-04	4.61E-04	4.61E-04	[4.61E-04, 4.61E-04]
	婚姻状況	2.15E-04	2.15E-04	2.15E-04	[2.15E-04, 2.15E-04]
	人種	1.54E-04	1.54E-04	1.54E-04	[1.54E-04, 1.54E-04]
T_4	職業	0.7208	0.7208	0.7208	[0.7208, 0.7208]
	収入	0.1316	0.1316	0.1316	[0.1316, 0.1316]
	ローン量	0.0211	0.0211	0.0211	[0.0211, 0.0211]
	ローンクラス	0.0002	0.0002	0.0002	[0.0002, 0.0002]

データによっては
最小コストモデルでも
属性の危険度を
精度よく評価できる

各モデルの比較 (誤差・計算コスト)



各モデルの比較 (計算にかかる時間)

T_1の各属性を3つのモデルで評価した際の計算時間

属性名	平均モデル（厳密解）[s]	サンプリングモデル [s]	最小コストモデル [s]
伝票 ID	19.14	0.13	0.01
購買日	0.22	0.02	0
購買時刻	0.47	0.03	0.02
購買商品	1.95	0.03	0.03
価格	5.57	0.34	0.04
数量	0.14	0.03	0.02
合計	27.49	0.58	0.12

データのレコード数や属性数などが増えるほど，この計算時間は増加する
サンプリングモデルを用いることにより，厳密解の47分の1の計算時間で危険度を求められる

3章まとめ

- 匿名化の研究には攻撃者の想定が不可欠である
- 部分的な背景知識を持つ新たな攻撃者モデルを提案し，
  その危険度を平均識別確率によって評価した
- 提案したモデルを用いて4つの実データを評価し，
  データ中の危険な属性を明らかにした
- 平均識別確率を近似する3つのモデルを提案し，
  それらの精度とコストを評価した

目次

1. 部分的な背景知識を持つ攻撃者を想定した
   再識別リスク評価モデルの提案と評価（3 章）

2. 履歴データの数理モデルの提案とk-匿名化に必要な
ダミーレコード数推定への応用（4 章）

3. 貢献 2 の紹介（5~8 章）

4. 貢献 3 の紹介（9~10 章）

課題２：履歴データの識別リスク(再掲)

[3] Hiroaki Kikuchi, Katsumi Takahashi, “Ziph Distribution Model for Quantifying Risk of Reidentification from Trajectory Data”, Journal of Information Processing, Vol. 24, No. 5, pp. 816–823, 2016.

課題：動的に変化する履歴データから個人が識別されるリスクの研究は不十分

履歴データ

名前	日付	商品
Alice	12/1	2
Bob	12/2	3
Alice	12/2	1
Carol	12/2	2
Bob	12/3	1
Bob	12/4	4

個人数=3, レコード数=6

履歴データから個人は特定されないという誤解

2013年にJR東日本は，交通系ICカードSuicaの利用履歴データを個人が特定できる情報ではないとみなし，そのデータを他社に提供した.

→2016年の菊池らの研究[3]により，高々3駅分の履歴から98%の個人が識別されることが判明した.

動的に変化するイベントの正確な定式化の難しさが履歴データのリスク研究の妨げになっている.

解決：いくつかの仮定を置いて，履歴データのふるまいを数理モデル化する

提案モデルで定式化するもの

xレコードの履歴データが y種類の商品(全 l 商品)を持つ確率	$Pr(y\|x)$ =???
xレコードの履歴データが持つ 商品種類数(全 l 商品)の期待値	$E[y\|x,l]$ = ???
加工に必要な疑似レコード数 Δmの期待値	$E(\Delta m)$ = ???

匿名化と疑似レコード

履歴データを匿名化する手法の一つとして，疑似レコードの追加がある

元データ

User ID	Goods
1	Apple
2	Apple
2	Book
3	Book

購買商品から
個人が簡単に識別される

疑似レコード
追加

k-匿名化

加工データ

User ID	Goods
1	Apple
1	Book
2	Apple
2	Book
3	Apple
3	Book

3人の区別がつかない
(3-匿名化)

加工コストと疑似レコード数

疑似レコード数は加工データの有用性評価値として用いられる

- **原田玲央, 伊藤聡志, 菊池浩明,「商品の特徴による再識別リスクとクラスタリングを用いた購買履歴データ匿名加工手法の提案」, SCIS-2017**

  顧客400人，38,087レコードの購買履歴データをJaccard距離を用いて5-匿名化する際に，疑似レコードが約180,000必要になる

- **匿名加工・再識別コンテストPWSCUP-2018**

  レコード数の多い上位2顧客の区別をつかなくするために2,688レコードの追加/削除が必要である

基礎定義 1

n：顧客数，m：レコード数，l：商品種類数

元データ

User ID	Goods
ITO	Apple
YAMA	Apple
YAMA	Book
MORI	Book

顧客集合 $U = \{u_1, u_2, \dots, u_n\}$

商品集合 $I(U) = \{g_1, g_2, \dots, g_l\}$

顧客u_iの購買商品 $I(u_i) = \{g_1^i, g_2^i, \dots, g_{l_i}^i\}$

例)←の場合

$n = 3, m = 4, l = 2$

$U = \{\mathrm{ITO}, \mathrm{YAMA}, \mathrm{MORI}\}$

$I(U) = \{\mathrm{Apple}, \mathrm{Book}\}$

$I(\mathrm{ITO}) = \{\mathrm{Apple}\}, I(\mathrm{YAMA}) = \{\mathrm{Apple}, \mathrm{Book}\}$

$I(\mathrm{MORI}) = \{\mathrm{Book}\}$

基礎定義２

c：クラスタ数，Δm：疑似レコード数

加工データ

Pseudonym	Goods
1	Apple
1	Book
2	Apple
2	Book
3	Apple
3	Book

クラスタ$U_i = \{u_1^i, u_2^i, \ldots, u_{s_i}^i\}$

$U_1 \cup \cdots \cup U_c = U$

U_iの大きさ $s_i = |U_i|$

U_iの購買商品 $I(U_i) = I(u_1^i) \cup \cdots \cup I(u_{s_i}^i)$

例）元データを仮名化・3-匿名化した場合

$c = 1, \Delta m = 2$

$U_1 = \{1, 2, 3\}$ ， $s_1 = |U_1| = 3$

$I(U_1) = \{\mathrm{Apple}, \mathrm{Book}\}$

疑似レコード数の厳密解

疑似レコード数は$\Delta m = \sum_{i=1}^{c} s_i |I(U_i)| - \sum_{i=1}^{n} |I(u_i)|$で求められる

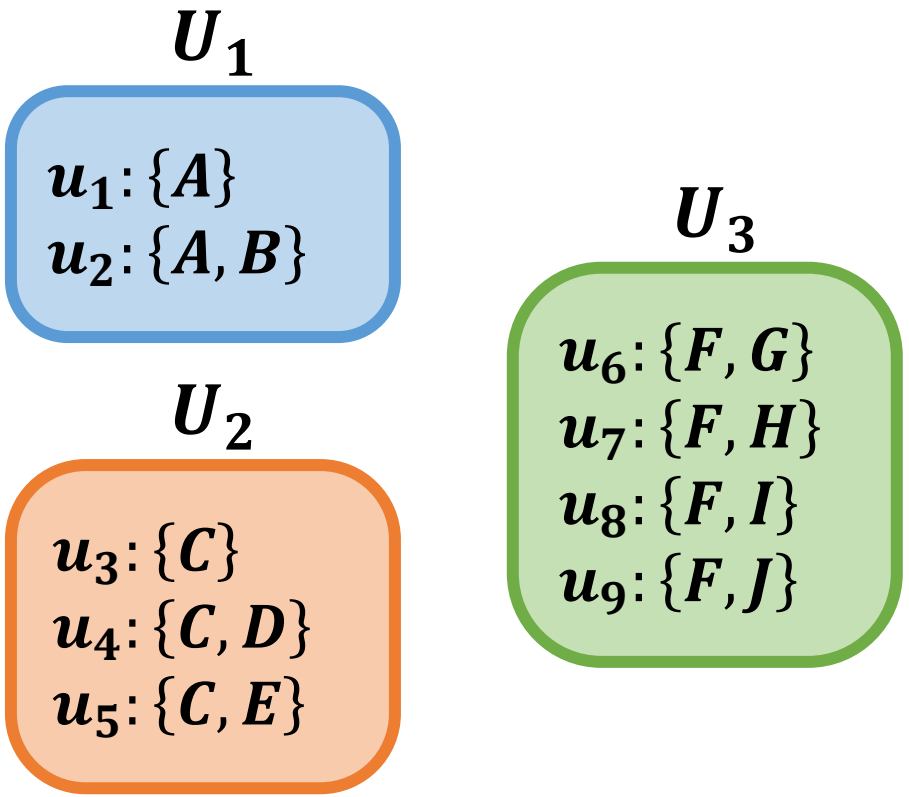
(例)右図の3クラスタ内の顧客の区別がつかないようにするとき

$$\Delta m = \sum_{i=1}^{3} s_i |I(U_i)| - \sum_{i=1}^{9} |I(u_i)|$$

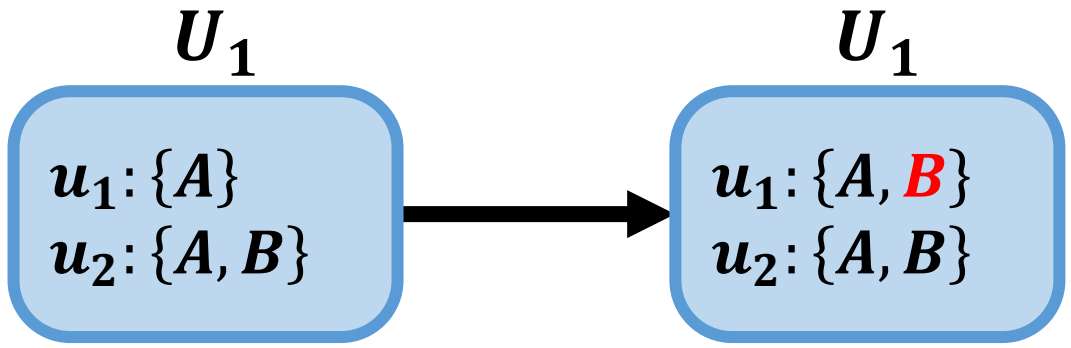
$$= s_1|I(U_1)| + s_2|I(U_2)| + s_3|I(U_3)| \\ |I(u_1)| + \cdots + |I(u_9)|)$$

$$= 2 * 2 + 3 * 3 + 4 * 5 - (16)$$

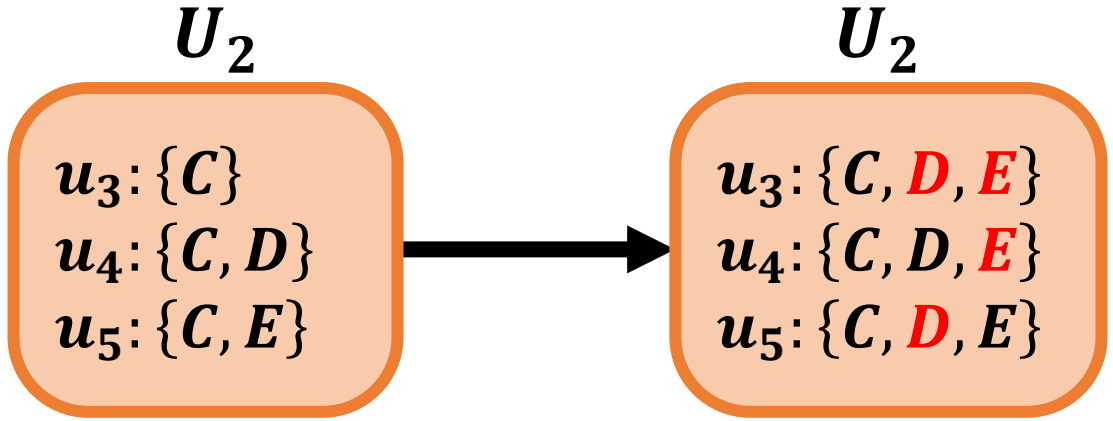
$$= 17$$ レコード追加すればよい！



計算例

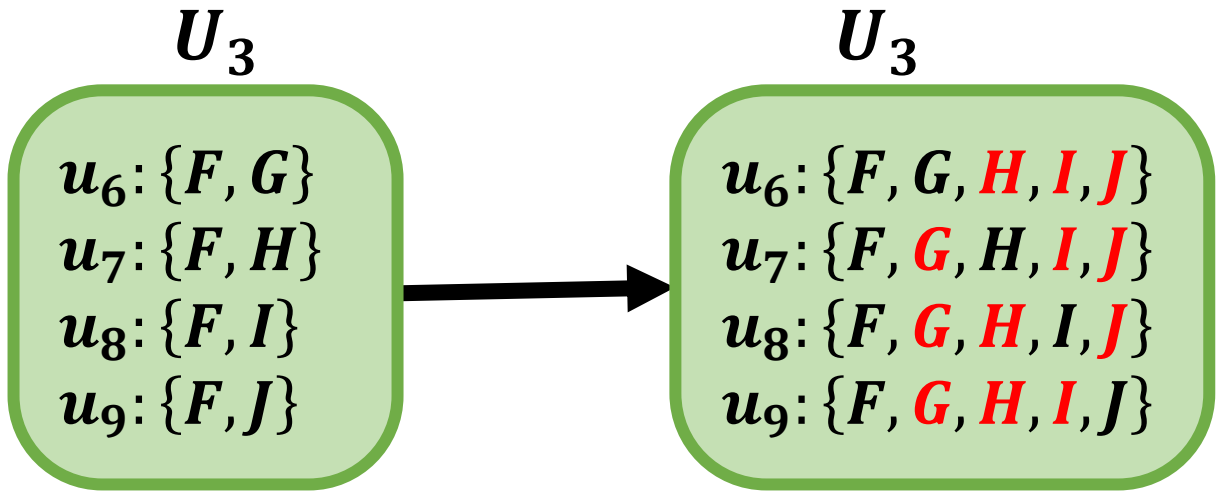


1レコード追加



4レコード追加

計17レコード



12レコード追加

計算例

U_1

$u_1: \{A\}$
$u_2: \{A, B\}$

$s_1 = 2$
$I(U_1) = \{A, B\}$

U_2

$u_3: \{C\}$
$u_4: \{C, D\}$
$u_5: \{C, E\}$

$s_2 = 3$
$I(U_2) = \{C, D, E\}$

U_3

$u_6: \{F, G\}$
$u_7: \{F, H\}$
$u_8: \{F, I\}$
$u_9: \{F, J\}$

$s_3 = 4$
$I(U_3) = \{F, G, H, I, J\}$

$$\Delta m = \sum_{i=1}^{c} s_i |I(U_i)| - \sum_{i=1}^{n} |I(u_i)|$$

$$= \sum_{i=1}^{3} s_i |I(U_i)| - \sum_{i=1}^{9} |I(u_i)|$$

$$= s_1|I(U_1)| + s_2|I(U_2)| + s_3|I(U_3)| - (|I(u_1)| + \cdots + |I(u_9)|)$$

$$= 2 * 2 + 3 * 3 + 4 * 5 - (16)$$

$$= 17$$

Δmは加工しないと求められない！

$$\Delta m = \sum_{i=1}^{c} s_i |I(U_i)| - \sum_{i=1}^{n} |I(u_i)|$$

s_i： クラスタU_iの大きさ(人数)
$|I(U_i)|$：U_iの購買商品の種類数
$|I(u_i)|$：顧客u_iの購買商品の種類数

これらの値は実際に加工をしないと手に入らない
しかし，これらを**加工前に手に入る数**で置き換えれば
加工をしなくてもΔmの(おおまかな)値を計算できる！

3つの仮定と商品種類数問題

目的を達成するために，本研究では3つの仮定を置く

1. l 種類の属性の値は独立で一様な確率$1/l$で生起する（$1/l$仮定）
2. n人の顧客のc個のクラスタの大きさはすべて等しくn/cである（n/c仮定）
3. n人の各顧客が持つ計m個のレコード数はすべて等しくm/nである（m/n仮定）

s_i：n/c仮定より，$s_i = n/c$

$|I(U_i)|$：m/cレコードあるときの商品種類数

$|I(u_i)|$：m/nレコードあるときの商品種類数

これの期待値を求めることができれば
$|I(U_i)|$, $|I(u_i)|$を置き換えることができ，
Δmの期待値を求めることができる！

[商品種類数問題]

履歴データがxレコードあるとき，商品(全 l 種類)の種類数yはいくらか？

求めたいもの(再掲)

x: レコード数，y: 種類数，l: 全種類数，n: 顧客数
m: 全レコード数，c: クラスタ数，Δm: 疑似レコード数

提案モデル

xレコードの履歴データが y種類の商品(全 l 商品)を持つ確率	$Pr(y\|x)$ =???
xレコードの履歴データが持つ 商品種類数(全 l 商品)の期待値	$E[y\|x,l]$ = ???
加工に必要な疑似レコード数 Δmの期待値	$E(\Delta m)$ = ???

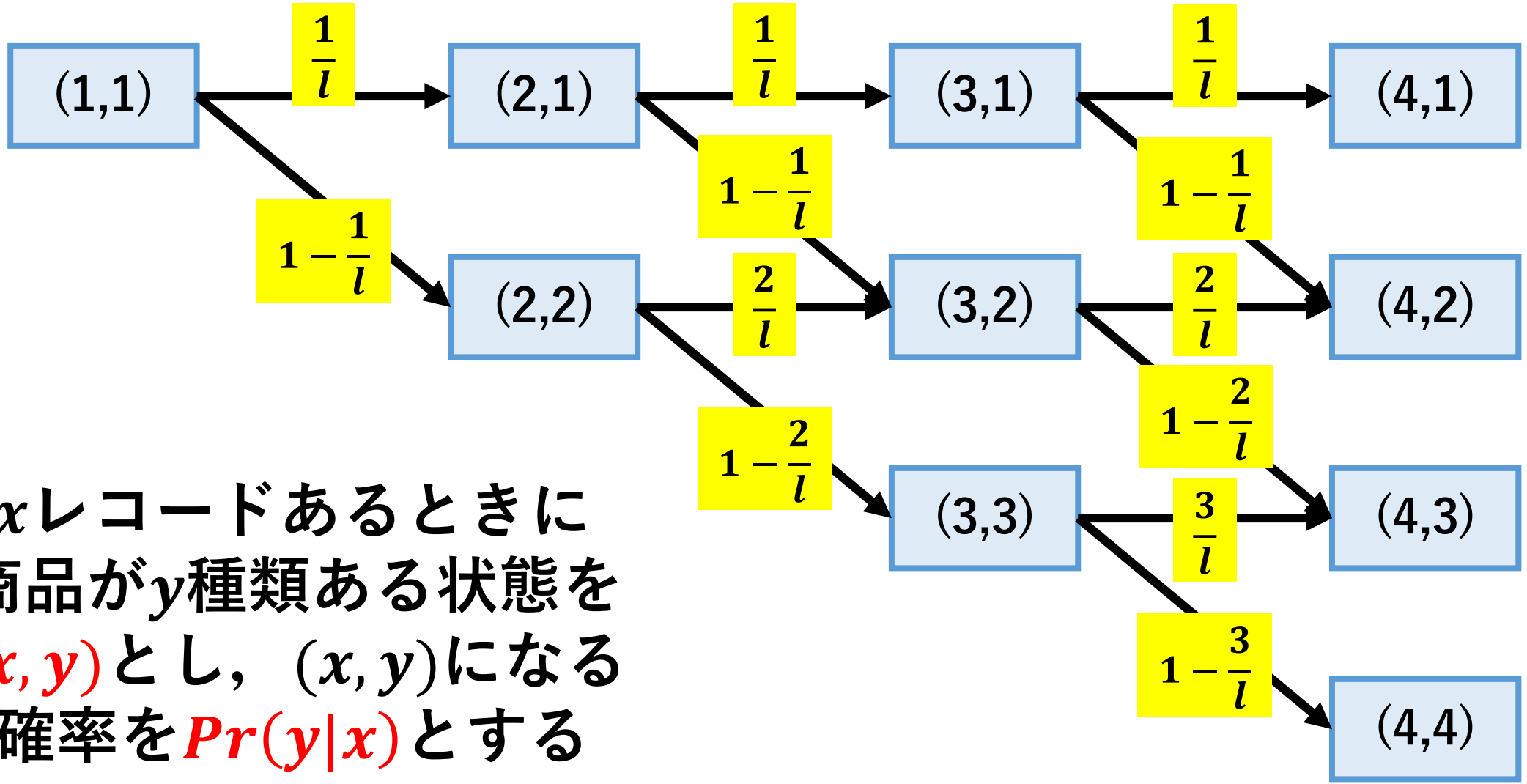
$1/l$仮定

m/n仮定

n/c仮定

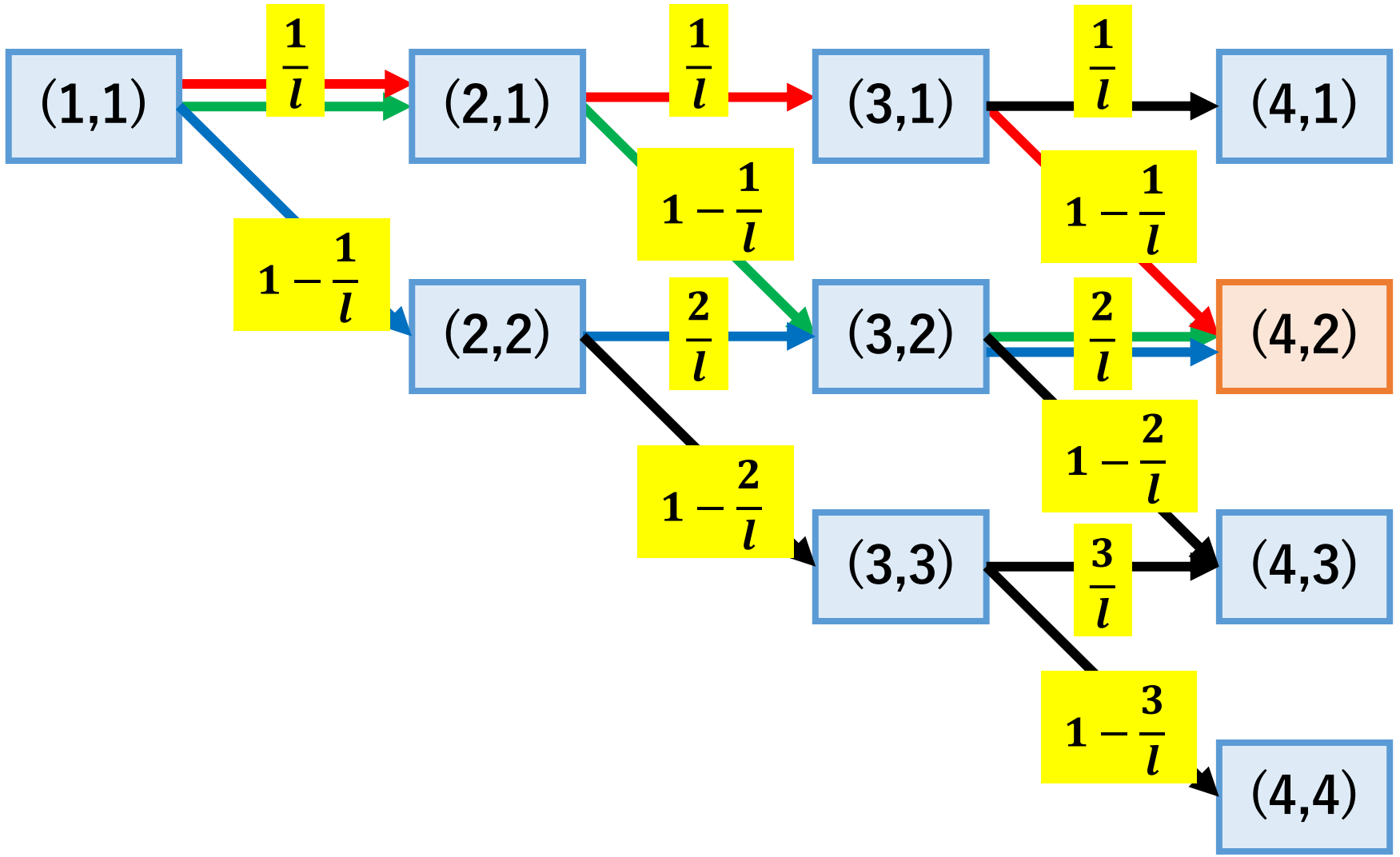
提案モデルの応用

レコード数と商品種類数の状態遷移図



xレコードあるときに
商品がy種類ある状態を
(x, y)とし，(x, y)になる
確率を$Pr(y|x)$とする

例：4レコードのデータが2種類の商品を持つ確率



$$Pr(2|4)$$

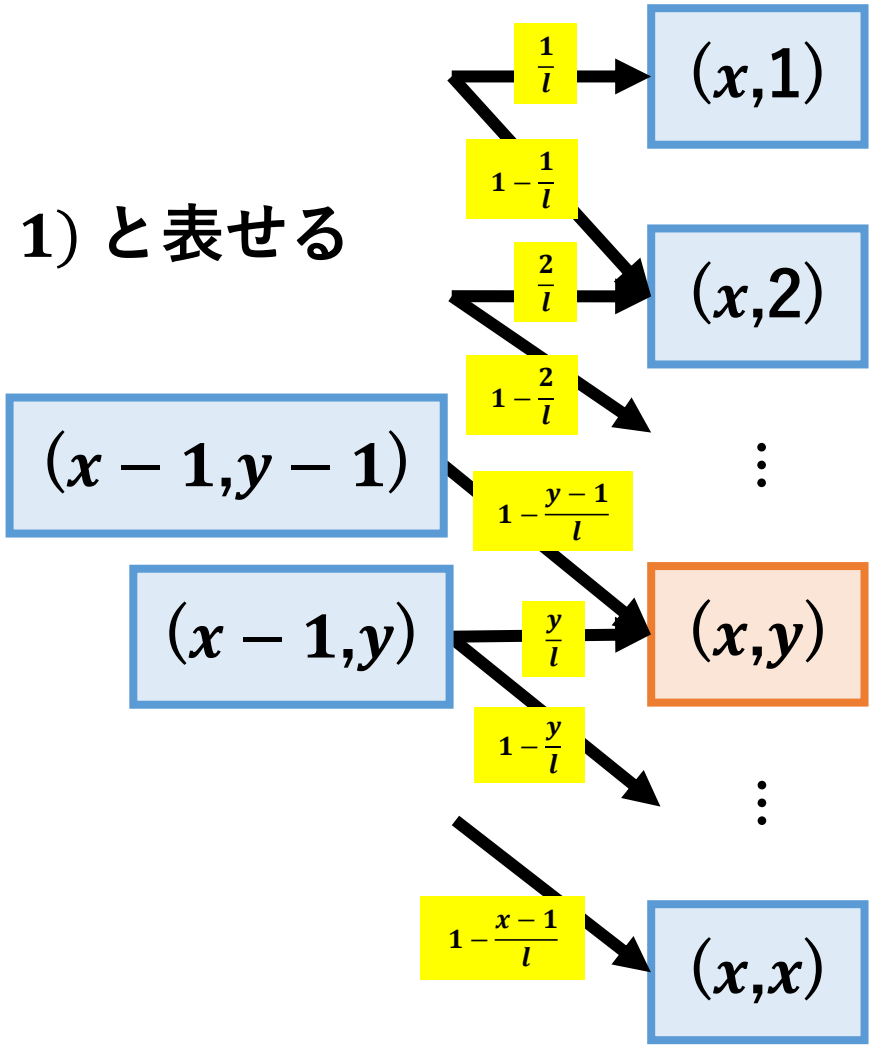
$$= \frac{1}{l} * \frac{1}{l} * \left(1 - \frac{1}{l}\right) + \frac{1}{l} * \left(1 - \frac{1}{l}\right) * \frac{2}{l} + \left(1 - \frac{1}{l}\right) * \frac{2}{l} * \frac{2}{l}$$

$$= \frac{7}{l^2}\left(1 - \frac{1}{l}\right)$$

$Pr(y|x)$の一般式

右図より，

$Pr(y|x) = \left(1 - \frac{y-1}{l}\right) Pr(y-1|x-1) + \left(\frac{y}{l}\right) Pr(y|x-1)$ と表せる



定理 4.2.1

l 種類の値をとるxレコードのデータの属性中に
y種類の値がある生じる確率$Pr(y|x)$は

$$Pr(y|x) = \prod_{j=0}^{y-1}\left(1 - \frac{j}{l}\right) \cdot \sum_{m_1+\cdots+m_y=x-y} \left(\frac{1}{l}\right)^{m_1} \cdots \left(\frac{y}{l}\right)^{m_y}$$

である．ただし，$m_1, \ldots, m_y \geq 0,\ x \geq y \geq 1$ である．

$E[y|x,l]$の一般式

定理 4.2.2　　$E[y|x,l] = (-l)\left(1-\frac{1}{l}\right)^{x} + l$ である.

(証明)
$1/l$ 仮定の下で，x レコードのデータで全 l 種類の値が少なくとも
1回生起する確率は $1-\left(1-\frac{1}{l}\right)^{x}$ である.

各商品が出現するかしないかの期待値はそれぞれ $1-\left(1-\frac{1}{l}\right)^{x}$ であり，
期待値の線形性より，商品数の期待値はその l 倍で求められる.

疑似レコード数の期待値 $E(\Delta m)$

$\Delta m = \sum_{i=1}^{c} s_i |I(U_i)| - \sum_{i=1}^{n} |I(u_i)|$ と3つの仮定より，

$E(\Delta m) = \sum_{i=1}^{c} \frac{n}{c} E\left[y|\frac{m}{c}, l\right] - \sum_{i=1}^{n} E\left[y|\frac{m}{n}, l\right]$ とかける．

定理 4.3.1

3つの仮定の下，疑似レコード数の期待値は，

$E(\Delta m) = nl\left\{\left(1-\frac{1}{l}\right)^{\frac{m}{n}} - \left(1-\frac{1}{l}\right)^{\frac{m}{c}}\right\}$ である．

$n = 400, m = 38000,$
$l = 2700, c = 50$のとき
$E(\Delta m) = 227658.4$

提案モデルまとめ

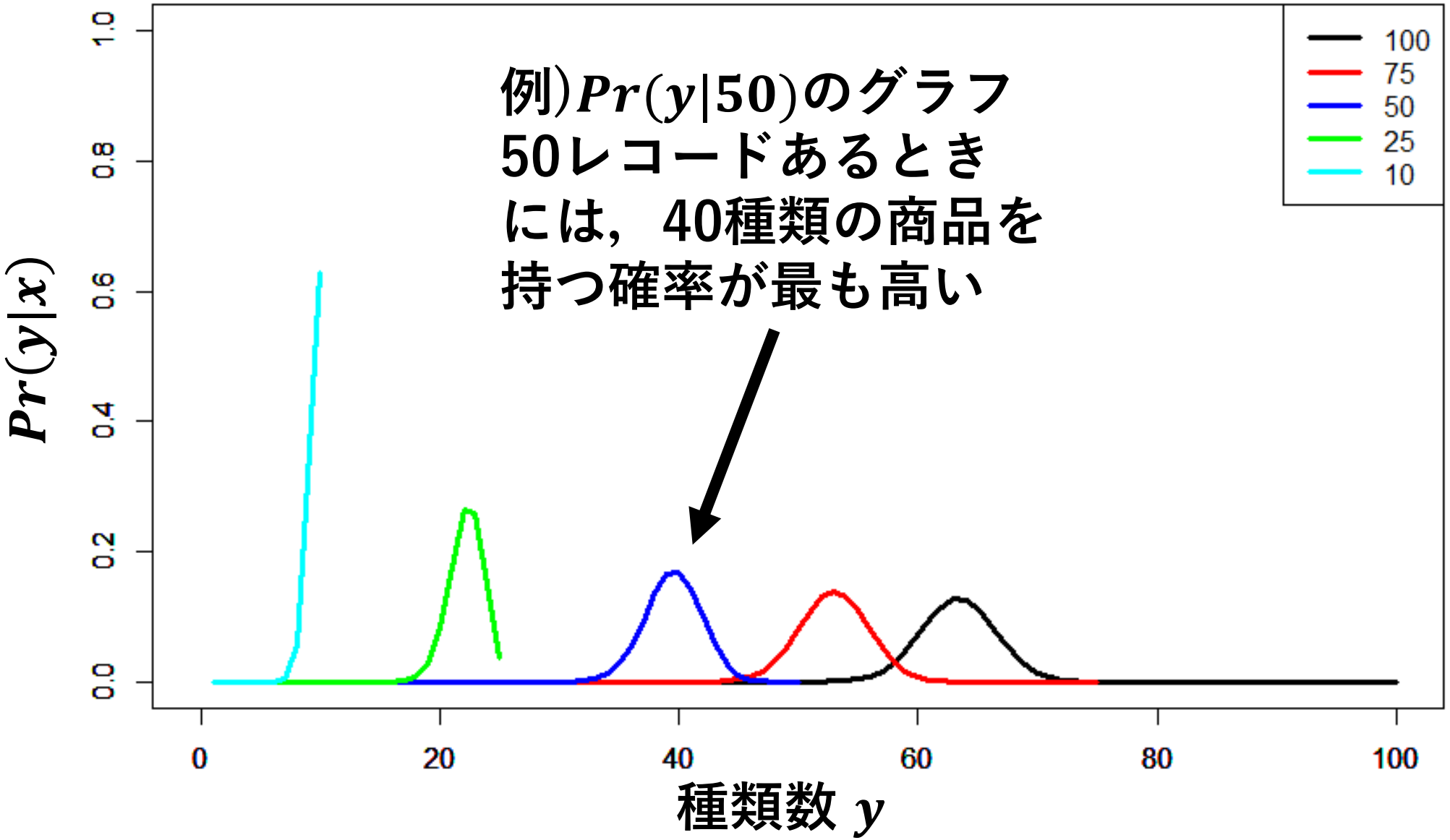
x: レコード数，y: 種類数，l: 全商品種類数
n: 顧客数，m: 全レコード数，c: クラスタ数
Δm: 疑似レコード数

提案モデル

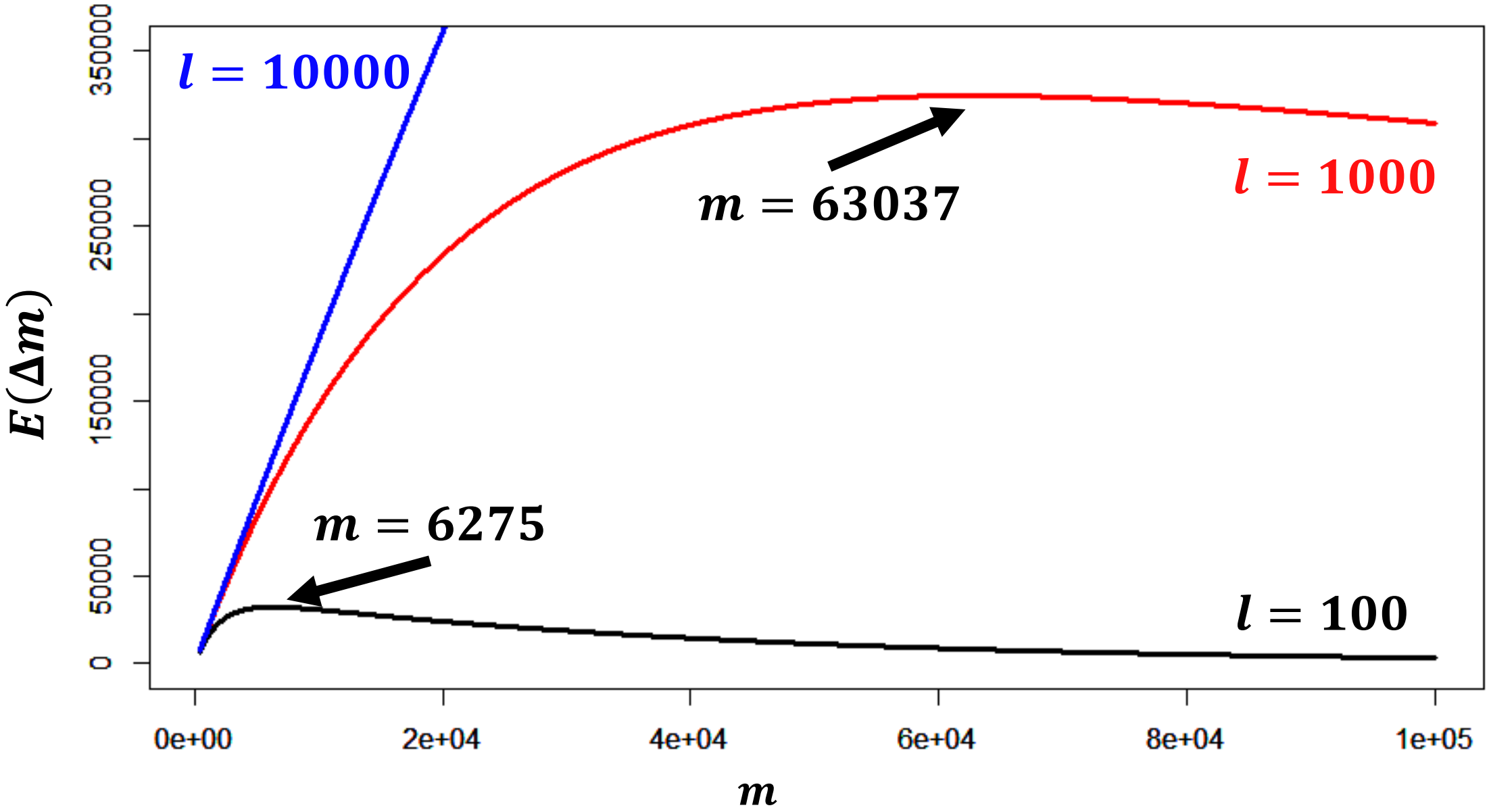
xレコードの履歴データが y種類の商品(全 l 商品)を持つ確率	$Pr(y\|x) = \prod_{j=0}^{y-1}\left(1-\frac{j}{l}\right) * \sum_{m_1+\cdots+m_y=x-y}\left(\frac{1}{l}\right)^{m_1}\cdots\left(\frac{y}{l}\right)^{m_y}$
xレコードの履歴データが持つ 商品種類数(全 l 商品)の期待値	$E[y\|x,l] = (-l)\left(1-\frac{1}{l}\right)^{x} + l$
加工に必要な疑似レコード数 Δmの期待値	$E(\Delta m) = nl\left\{\left(1-\frac{1}{l}\right)^{\frac{m}{n}} - \left(1-\frac{1}{l}\right)^{\frac{m}{c}}\right\}$

提案モデルの応用

$Pr(y|x)$の分布$(l = 100)$



$E(\Delta m)$の分析$(n = 400, c = 20)$



$1/l, m/n$ 仮定の影響

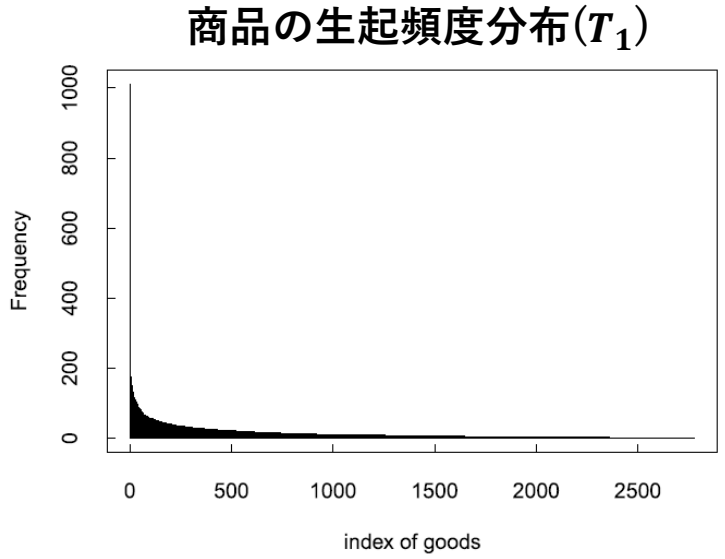
p_j：l 種類中 j 番目の値が生起する確率

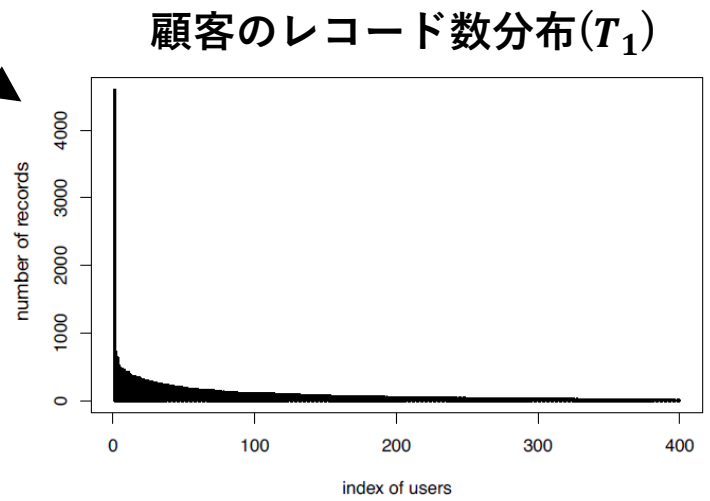
b_i：n 人中 i 番目の顧客のレコード数

$E(\Delta m_2)$：$1/l$ 仮定を外したときの$E(\Delta m)$

$E(\Delta m_3)$：m/n, $1/l$ 仮定を外したときの$E(\Delta m)$

$E(\Delta m_4)$：m/n 仮定を外したときの$E(\Delta m)$





系 4.3.2, 4.3.3, 4.3.4

$$E(\Delta m_2) = n\sum_{j=1}^{l}\left\{\left(1-p_j\right)^{m/n} - \left(1-p_j\right)^{m/c}\right\}$$

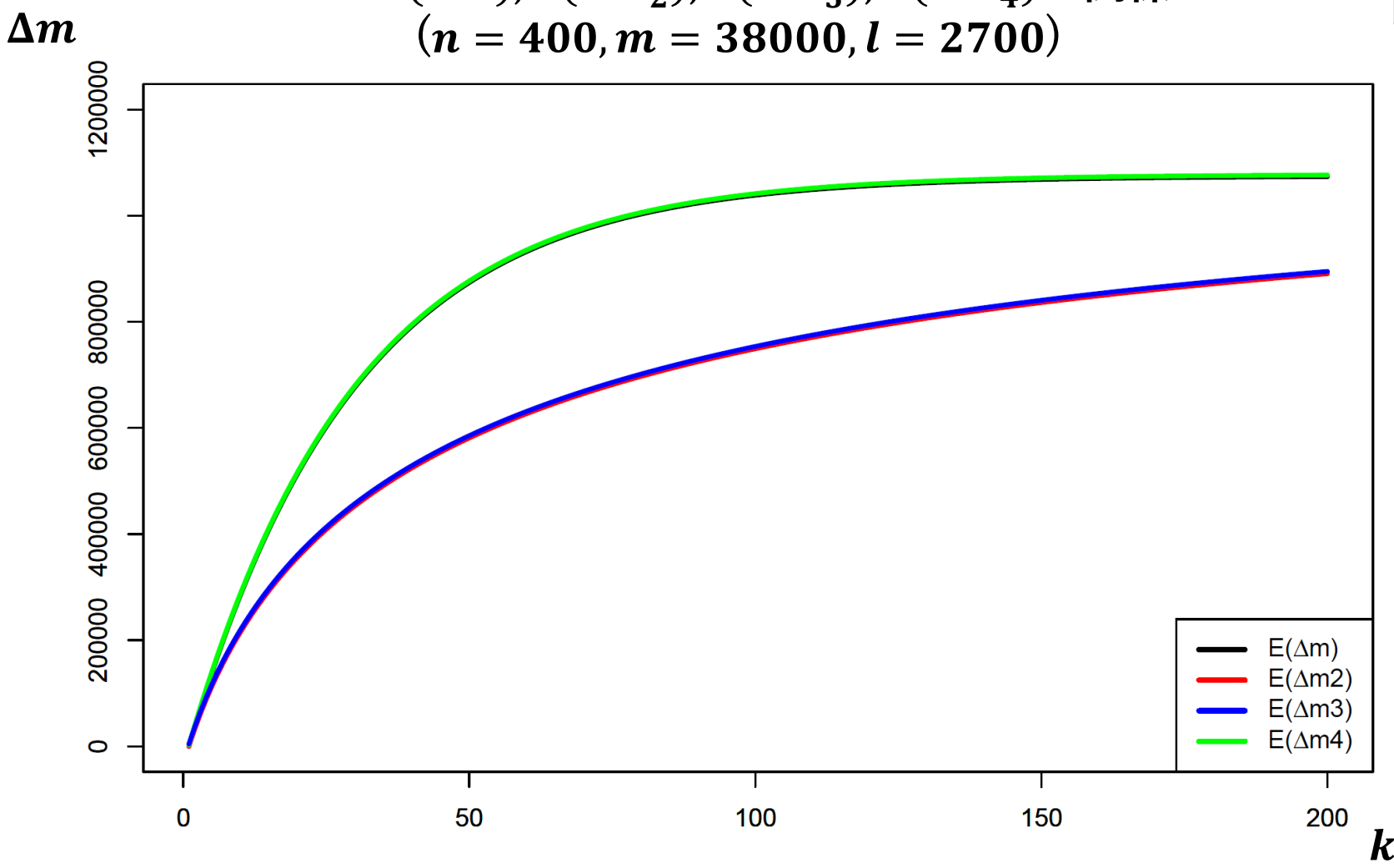
$$E(\Delta m_3) = n\sum_{j=1}^{l}\left\{1-\left(1-p_j\right)^{m/c}\right\} - \sum_{i=1}^{n}\sum_{j=1}^{l}\left\{1-\left(1-p_j\right)^{b_i}\right\}$$

$$E(\Delta m_4) = l\sum\left\{(1-1/l)^{b_i} - (1-1/l)^{m/c}\right\}$$

$1/l, m/n$ 仮定の影響

		$1/l$ 仮定	
		有	無
m/n 仮定	有	$E(\Delta m)$	$E(\Delta m_2)$
	無	$E(\Delta m_4)$	$E(\Delta m_3)$

kと$E(\Delta m), E(\Delta m_2), E(\Delta m_3), E(\Delta m_4)$の関係
$(n = 400, m = 38000, l = 2700)$



黒線と緑線，赤線と青線がほぼ重なっている
→m/n 仮定による加工コストへの影響は殆ど無い

黒線と緑線が，赤線と青線より上部にある
→$1/l$ 仮定によって加工コストが多く推定されている

赤線，青線，緑線は p_j, b_i が無いと求めることができないため，加工コスト推定には黒線$(E(\Delta m))$が有用である

4章まとめ

- 履歴データから個人が識別されるリスクの研究は十分にされておらず，そのためには動的なイベントの定式化が必要である

- 3つの仮定($1/l$, m/n, n/c 仮定)をおくことにより，履歴データの振る舞いを定式化する数理モデルを提案した

- xレコードの履歴データがy種類の値を持つ確率$Pr(y|x)$と種類数の期待値$E[y|x,l]$を与える数理モデルを提案した

- 提案モデルを応用することにより，履歴データのk-匿名化に必要な疑似レコード数の期待値$E(\Delta m)$を，加工前に見積ることを実現した

目次

1. 部分的な背景知識を持つ攻撃者を想定した
   再識別リスク評価モデルの提案と評価（3 章）

2. 履歴データの数理モデルの提案とk-匿名化に必要な
   ダミーレコード数推定への応用（4 章）

3. 貢献 2 の紹介（5~8 章）

4. 貢献 3 の紹介（9~10 章）

課題 4：実験データへの依存性 (再掲)

課題：データに対する匿名化の影響は，対象となるデータに大きく依存する

解決：多種多様なデータに対して実験的評価を行う

ID	内容	個人数	レコード数 (行)	属性数 (列)	扱う章	データの種類
1	購買履歴	400	38,087	7	3,4,5,9	オープンデータ
2	健康診断	198,740	964,636	49	6	匿名加工情報
3	交通 IC カード	31	584	10	7	実データ
4	世帯支出	8,333	8,333	25	8	合成データ
5	糖尿病患者	71,518	101,766	50	3	オープンデータ
6	世帯収入	32,561	32,561	16	3,10	オープンデータ
7	ローン借入	42,538	42,538	145	3	オープンデータ
8	疑似人流	6,432	901,465	9	9,10	合成データ
9	傷病レセプト	288,568	39,363,878	15	6	匿名加工情報
10	医薬品レセプト	279,199	31,465,504	21	6	匿名加工情報

5～8章で用いるデータセット

章	用いるデータ	実験目的	評価対象
5	購買履歴 (動的データ)	動的なデータから個人が 識別されるリスクを評価する	安全性 (Jaccard 係数を用いた 再識別攻撃)
6	健康診断 (静的) 傷病/医薬品 レセプト (動的)	データから得られる有益な 情報が匿名化によってどれだけ 変化するかを評価する	有用性 (F 値や 相対リスク による評価)
7	交通 IC カード (動的データ)	複数の用途からなる複雑な データから個人が識別される リスクを評価する	安全性 (エント ロピーによる評価)
8	世帯支出 (静的データ)	静的なデータから個人が 識別されるリスクを評価する	安全性 (ユーク リッド距離による 再識別攻撃)

4つのタイプのデータセットを想定

1. **動的なデータ**
   (個人数 < レコード数)
2. **静的+動的なデータ**
3. **複数の用途が含まれる動的データ**
4. **静的データ**
   (個人数 = レコード数)

これらのデータを評価するのに
最適な指標を検討し,
5～8章で実験的に評価した

5 章：購買履歴データを用いた実験的評価

Jaccard 係数： $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$

元データ

顧客名	商品
Alice	Apple
Bob	Apple
Bob	Book
Carol	Book

加工データ

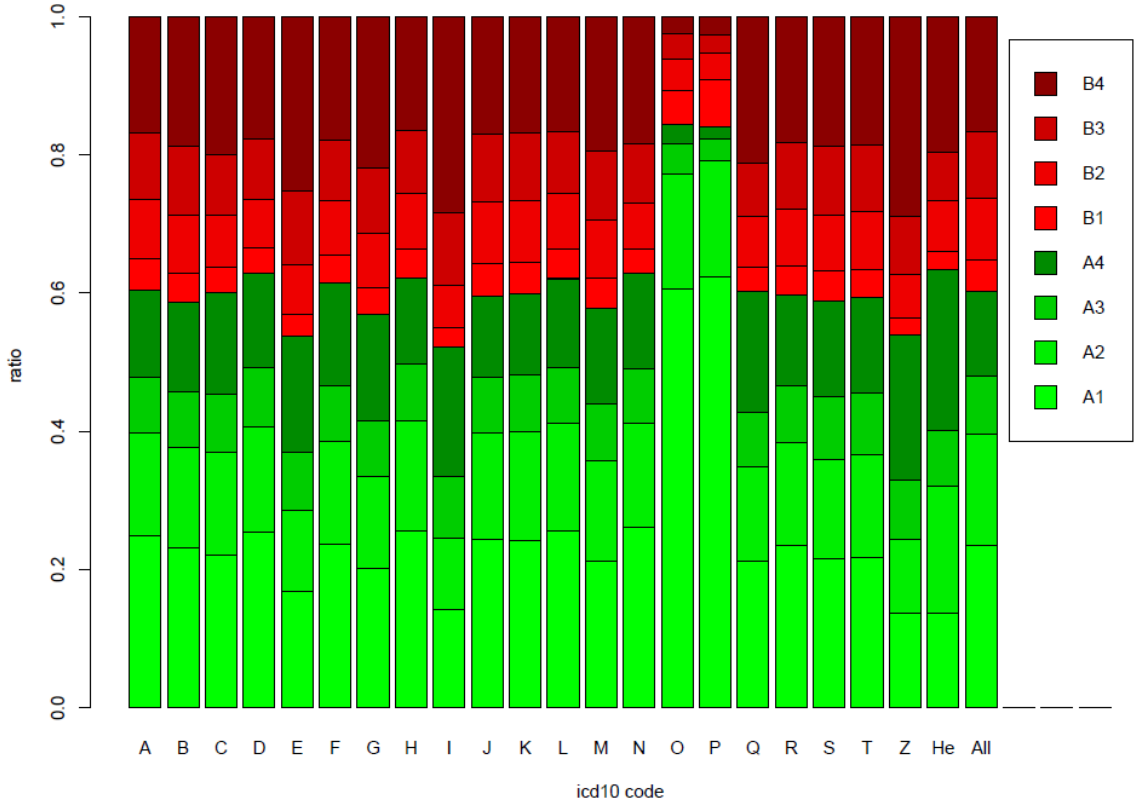
仮名	商品
1	Apple
2	Book
3	Book

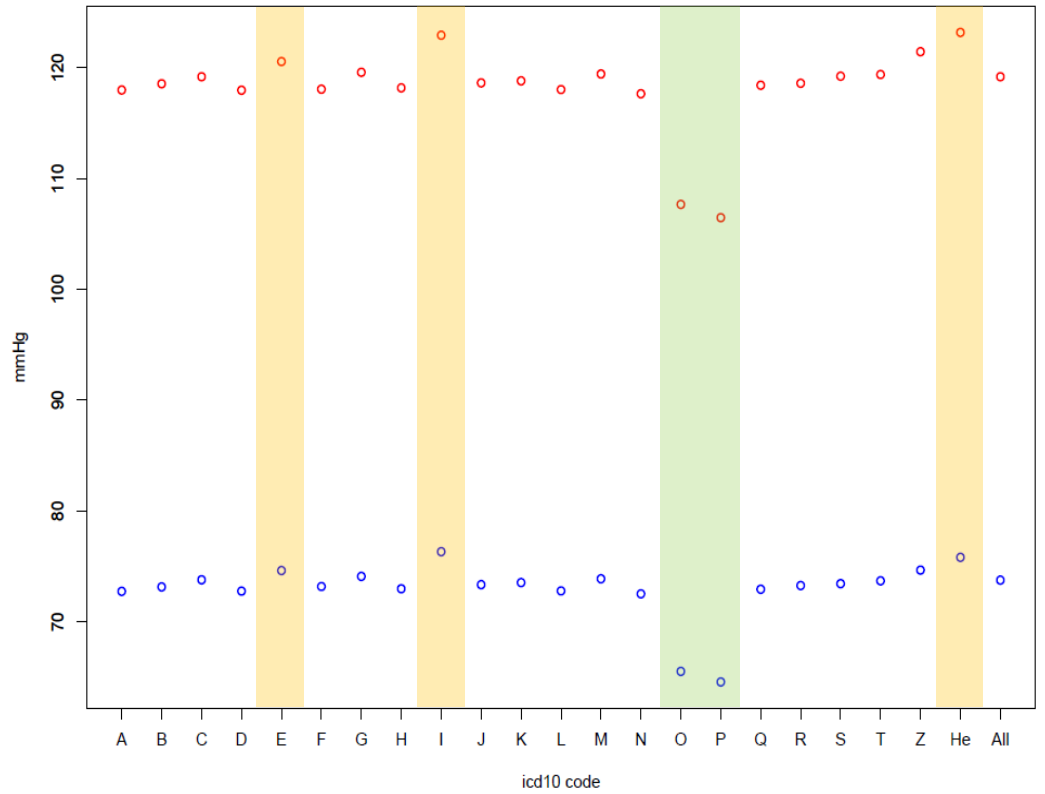
仮名1の購買商品集合は
Aliceに最も似ているので
仮名1 = Alice だ！

$J(A,1) = 1$
$J(B,1) = 1/2$
$J(C,1) = 0$

主な結果：38,000レコードの購買履歴データを50個のグループに分割してk-匿名化をするためには，約180,000のダミーレコードが必要

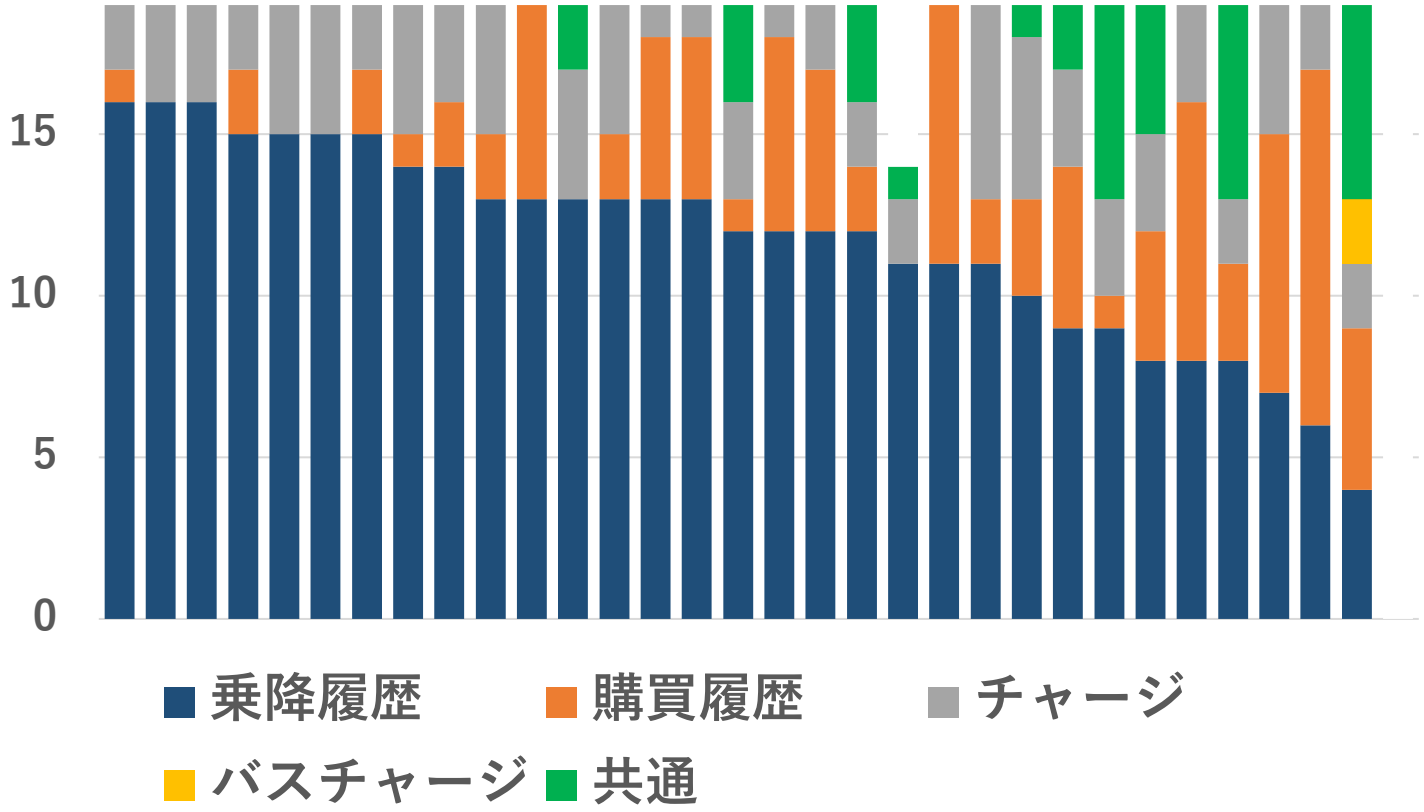
6 章：健康診断データを用いた実験的評価

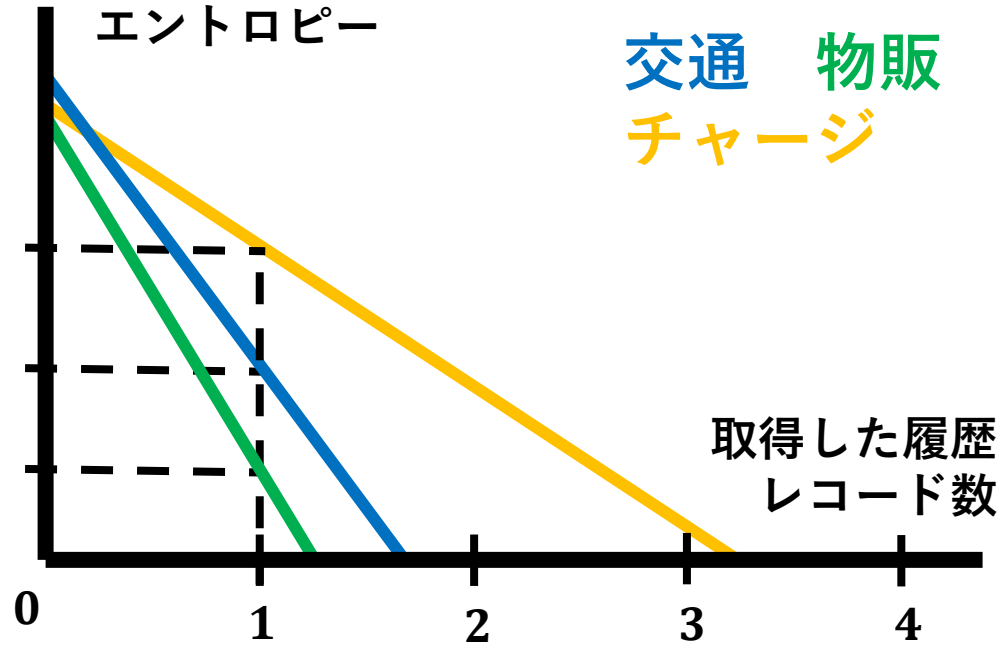




主な結果： 29万人分の病歴データを10-匿名化することによって,
識別される人数の割合は24.9%から2.9%まで減少するが,
高血圧に対する相対リスクは相対誤差で0.073しか変化しなかった

7章：交通ICカードデータを用いた実験的評価





主な結果：1つの乗降履歴(利用駅)が攻撃者に知られた場合，31人分の交通ICカードデータから個人が識別される確率は3.3%から28.4%まで上がり，購買履歴と乗降履歴を1つずつ知られた場合は，88.1%まで上がる

8 章：世帯支出データを用いた実験的評価

元データ

QI1	QI2	QI3	SA1	SA2
2	1	1	100	100
2	1	1	110	300
1	1	2	300	200
1	1	2	400	500

50

203.96

加工データ

QI1	QI2	QI3	SA1	SA2
2	1	1	150	100
2	1	1	160	300
1	1	2	350	200
1	1	2	450	500

ユークリッド距離を用いた再識別攻撃は
どのような加工をすれば防ぐことができるか？

	D_A	D_F	D_{10}
Method	K-anony	Averaging	K-ano + Ave
U1	-	-	-
U2	negative	-	negative
U3	negative	-	negative
U4	-	negative	negative
U5	-	negative	slightly
U6	-	-	-
S1	positive	negative	positive
S2	positive	negative	positive
E1	slightly	negative	slightly
E2	slightly	negative	slightly
E3	negative	positive	positive
E4	negative	positive	positive
EUC1	slightly	negative	positive

主な結果：8 つの加工手法を用いてデータを匿名化した結果,
ノイズ付加や平均化のような単純な摂動化では再識別を防げないことや,
k-匿名化によって再識別率を17%まで下げられることが明らかになった

目次

1. 部分的な背景知識を持つ攻撃者を想定した
   再識別リスク評価モデルの提案と評価（3 章）

2. 履歴データの数理モデルの提案とk-匿名化に必要な
   ダミーレコード数推定への応用（4 章）

3. 貢献 2 の紹介（5~8 章）

4. 貢献 3 の紹介（9~10 章）

課題３：k-anonymityの問題点(再掲)

[4] L. Sweeny, “k-anonymity: a model for protecting privacy”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), pp.557--570. (2006)

k-anonymity：Sweeneyによって提案された匿名性指標[4]．データ中の最低でもk人の区別がつかないとき，そのデータはk-anonymityを満たす．
k-anonymityを満たすようにデータを加工することは**k-匿名化**と呼ばれている．

元データ

名前	年齢	郵便番号
Alice	30	10055
Bob	25	10055
Carol	21	10023
David	55	10165
Eve	47	10224

2-匿名化された加工データ

仮名	年齢	郵便番号
1	21-30	100**
2	21-30	100**
3	21-30	100**
4	47-55	10***
5	47-55	10***

3人の区別がつかないグループに属する個人よりも

2人の区別がつかないグループに属する個人の方が危険(不公平)

「最低でも2人の区別がつかない」状態(2-匿名化)のためには，3人のグループは過度な加工ではないか？
データ中の個人によって安全性に差があるのは不公平ではないか？

課題：既存のk-匿名化では過度な加工や安全性の不公平さが生じてしまう

解決：k-concealmentという指標に注目した匿名化手法を提案する

k-concealment

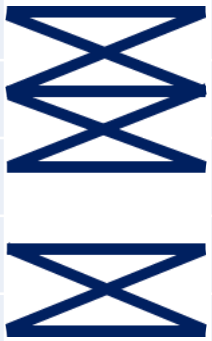
[5] Tamir Tassa, Arnon Mazza, Aristides Gionis, “k-Concealment: An Alternative Model of k-Type Anonymity”, TRANSACTIONS ON DATA PRIVACY 5, pp. 189--222. (2012)

k-concealment：

2012年にTamirらによって提案された匿名性指標[5]．元データと加工データの
レコード関係を2部グラフに表し，元データの各レコードが，加工データとの間に
少なくともk種類の完全マッチングの辺を持つとき，加工データは
k-concealmentを満たす．（2部グラフ＝加工の設計図，完全マッチング＝攻撃者の回答）
データをk-concealmentを満たすように加工することをk-concealment化とする．

元データ

名前	年齢	郵便番号
Alice	30	10055
Bob	25	10055
Carol	21	10023
David	55	10165
Eve	47	10224



2-concealment化された
加工データ

仮名	年齢	郵便番号
1	25-30	10055
2	21-30	100**
3	21-25	100**
4	47-55	10***
5	47-55	10***

攻撃者の回答パターン

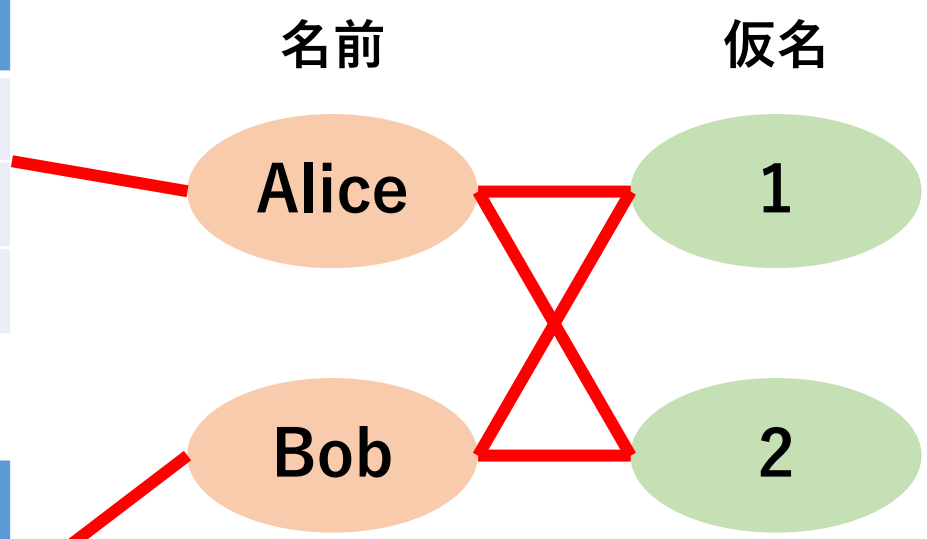
Ans1	Ans2	Ans3	Ans4	Ans5	Ans6
Alice	Bob	Alice	Alice	Bob	Alice
Bob	Alice	Carol	Bob	Alice	Carol
Carol	Carol	Bob	Carol	Carol	Bob
David	David	David	Eve	Eve	Eve
Eve	Eve	Eve	David	David	David

9 章：履歴データに対するk-concealment化手法の提案

k-concealmentは履歴データ(レコード数>顧客数)には適用できない

名前	日付	商品
Alice	12/2	3
Alice	12/3	1
Alice	12/4	4

名前	日付	商品
Bob	12/1	2
Bob	12/2	5



個人間 2-concealment
(少なくとも個人2人の
区別がつかない)

履歴データでは,
顧客ごとにレコード数が
異なる場合があるので,
個人のk-concealment化
だけでは不十分

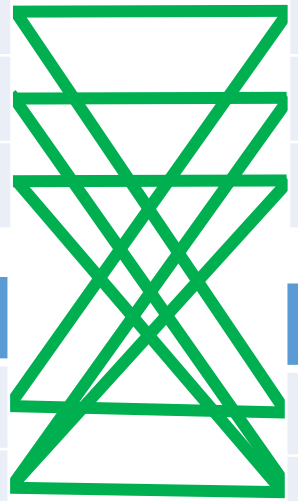
複数人の個人の区別を
つかないようにするには
レコード削除/追加が必要

9 章：履歴データに対するk-concealment化手法の提案

レコード間 k-concealmentと仮名の一般化を用いる手法を提案した

名前	日付	商品
Alice	12/2	3
Alice	12/3	1
Alice	12/4	4

名前	日付	商品
Bob	12/1	2
Bob	12/2	5



仮名	日付	商品
1	12/1-12/2	{2,3}
1	12/2-12/3	{1,2}
1,2	12/2-12/4	{4,5}

仮名	日付	商品
2	12/1-12/2	{2,3}
2	12/2-12/4	{1,4,5}

レコード間 2-concealment
(少なくともレコード2つの区別がつかない)

個人間だけでなく,
レコード間でも
k-concealment化すれば
履歴データを
k-concealment化できる

複数レコードの区別を
つかなくするために,
仮名の一般化も行う

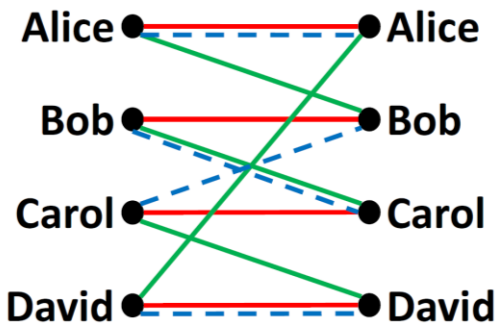
10 章：完全k-concealment化手法の提案

k-匿名化の過度な加工や安全性の不公平さを完全に解消する
加工アルゴリズム（完全k-concealment化）は，いまだ提案されていない

元データ

名前	年齢	性別
Alice	10	女
Bob	20	男
Carol	40	男
David	50	女

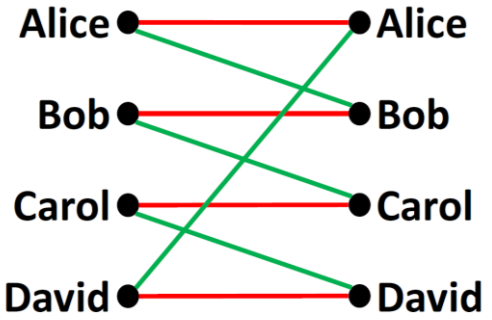
2部グラフ作成
（設計図作成）



2-concealment化

仮名	年齢	性別
1	10-50	女
2	10-40	男,女
3	20-40	男
4	40-50	男,女

データ作成



完全
2-concealment化

仮名	年齢	性別
1	10-50	女
2	10-20	男,女
3	20-40	男
4	40-50	男,女

2部グラフの辺の距離の総和を
加工コストと定義する

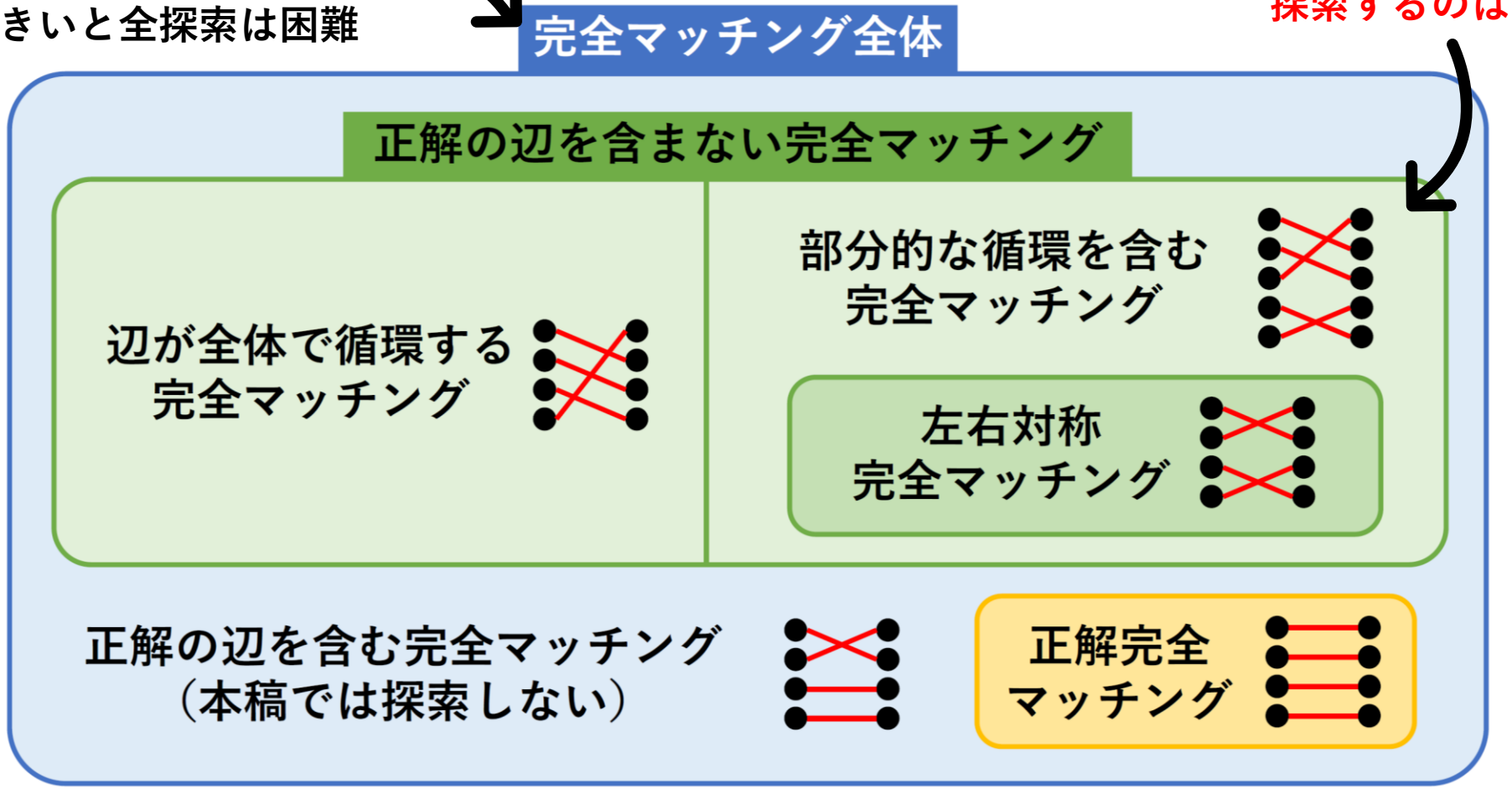
匿名化のコストを抑えるためには
距離の総和が小さい二部グラフを
見つければよい

10 章：完全k-concealment化手法の提案

完全マッチング全体を探索するのは膨大な時間がかかる

完全マッチングは全部で $n!$ 種類
あるため，nが大きいと全探索は困難

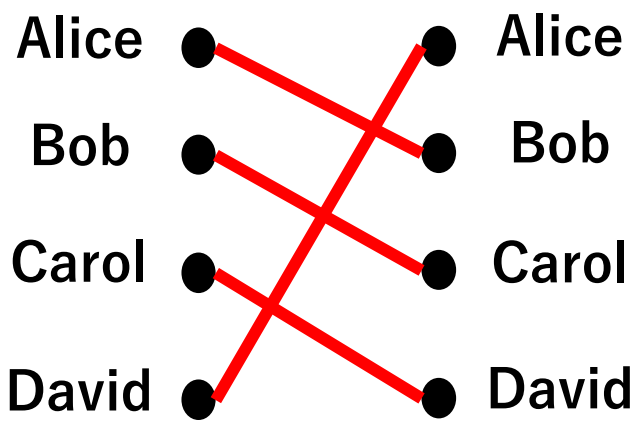
本研究の提案手法で
探索するのはこの部分のみ

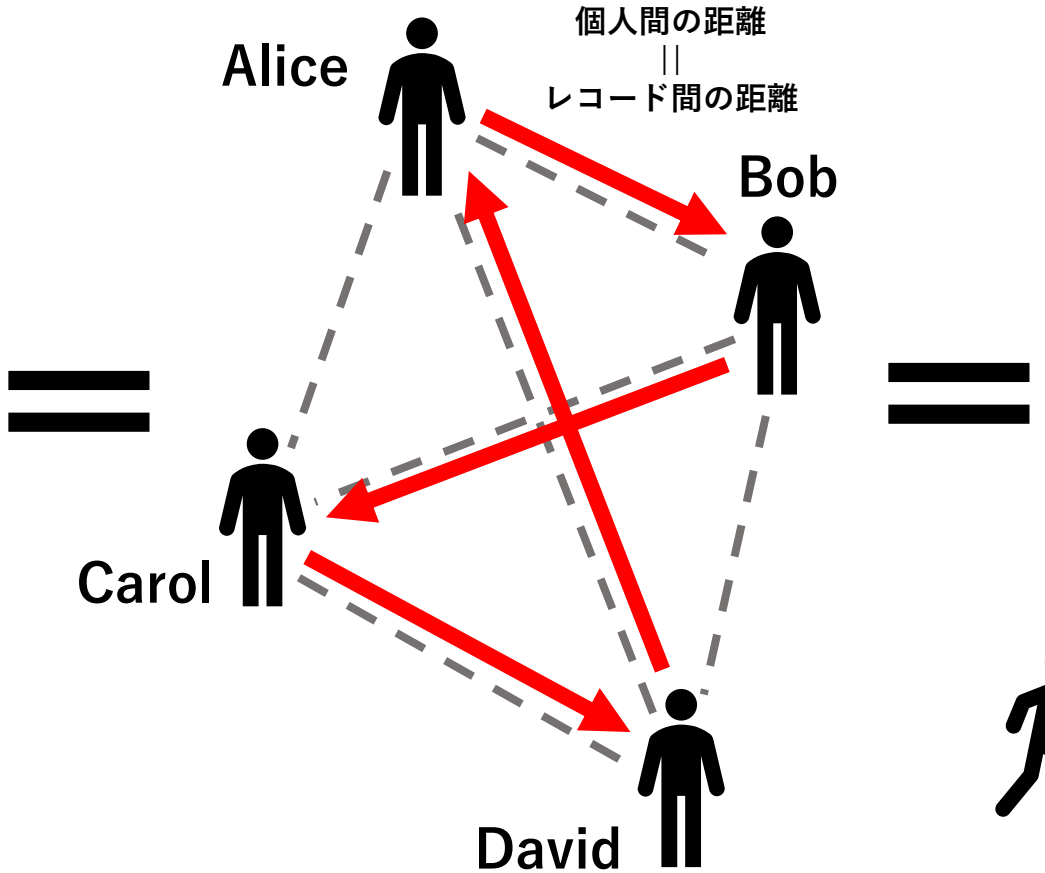


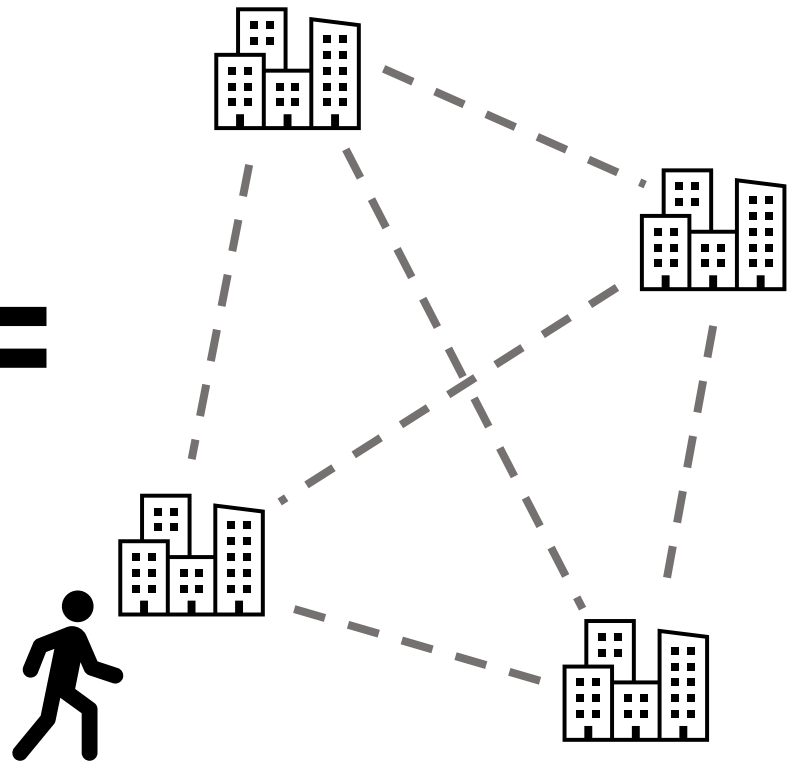
10 章：完全k-concealment化手法の提案

個人間の循環する完全マッチングならば,
都市間の巡回経路に置き換えることができるので
TSPの近似解法で(高速に)求められるのではないか？

巡回セールスマン問題 (TSP)
セールスマンが全ての都市を1回ずつ
巡回する場合の最短経路を求める問題







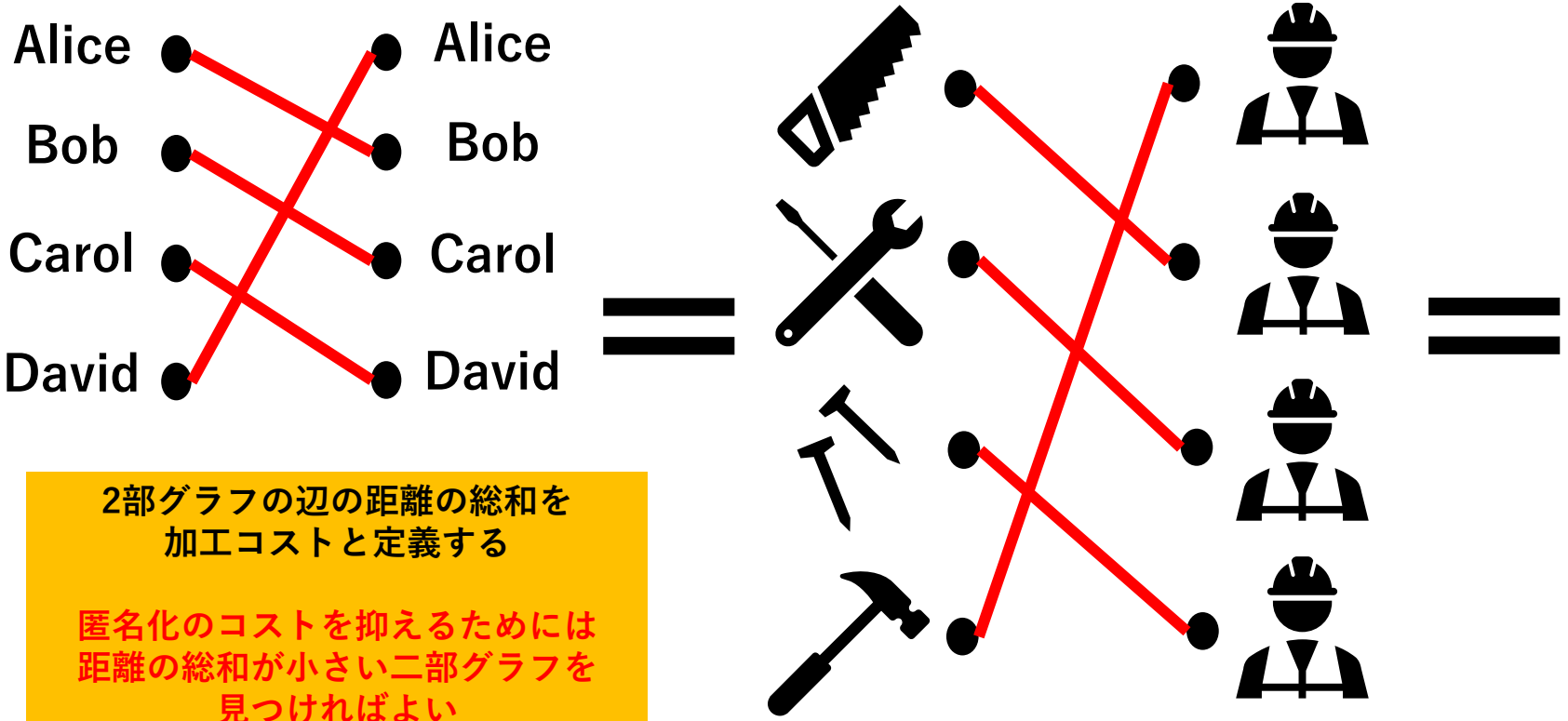
2部グラフの辺の距離の総和を
加工コストと定義する

匿名化のコストを抑えるためには
距離の総和が小さい二部グラフを
見つければよい

10 章：完全k-concealment化手法の提案

個人間のコストが低い**完全マッチング**ならば，
線形和割り当て問題（LASP）の近似解法で
求められるのではないか？

線形和割り当て問題（LASP）
どの仕事をどの作業員に割り当てれば
最も効率よく仕事が終わるか？



	○	◎	×	△
	×	△	○	×
	○	△	◎	○
	△	○	×	△

2部グラフの辺の距離の総和を
加工コストと定義する

匿名化のコストを抑えるためには
距離の総和が小さい二部グラフを
見つければよい

博士論文まとめ

- データに対する匿名化の影響を明らかにするため，
  4つの課題を設定してそれらを解決した

- 攻撃者や履歴データをモデル化することにより，
  データの安全性や有用性を理論的に評価した（貢献１）

- 多種多様なデータセットに対する識別リスクを想定し，それらを
  匿名化した際の影響を実験的に評価した（貢献２）

- k-匿名化の際に生じる過度な加工や安全性の不公平さを解決する
  k-concealment指標に注目し，これを満たすための新たな
  匿名化手法を提案した（貢献３）