資料１－３

個人特定性低減データの考察
(1) 識別子に関する脅威

菊池浩明（明治大学）

事務局案(4月16日，資料2)

	情報の種類	措置	備考
0	個人情報 （基本4情報，身長体重，会社名）	現行法の文言は維持	容易照合性は，情報取扱い事業者を基準とする
1	デバイス （パスポート番号，免許証番号，IPアドレス，スマホID)	必要な一項目についてのみ，(1)元の番号と不可逆で他と共有できない番号，かつ，(2)他の事業者と共有することが出来ない番号に置換	・個人情報の登録がないニックネームは除外か． ・Google ID(?) ・クッキー，ウェブビーコン？
2	普遍性を要する生体情報 （顔認識データ，指紋，DNA）	削除	普遍性でないもの（脈拍，血圧などの動的なもの）は除外か
3	移動情報，購買履歴	精度を落とす（匿名化措置）	

HIPAA Safe harbor ruleとの比較

HIPPA削除項目	準個人情報分類
1.氏名，2. 住所(郵便番号)，3. 生月日(年はよい)	（現行法）個人情報
4. 電話番号，5. FAX番号，6. メールアドレス， 10. 口座番号，11. 免許証番号， 12. 車両番号，13. 装置の識別番号(Device)，14. URL, 15. IPアドレス，	1. パスポート番号、免許証番号、IPアドレス、携帯端末ID等の個人または個人の情報通信端末（携帯電話端末、PC端末等）
7. SSN, 8. カルテ番号，9. 保険番号，	（該当なし？）
16. 生体情報，17. 顔写真，	2. 顔認識データ、遺伝子情報、声紋並びに指紋等
18. その他の識別子	
（該当なし）	3. 移動履歴、購買履歴等の**特徴的な行動の履歴**

普遍性のある生体情報

- 「観察容易性」と「普遍性（＝静的）」

	外部から観察が容易	内部でないと観察できない	
静的	性別，肌の色，人種， 生体情報（顔，指紋），身体的特性（筆跡，歩行，声紋）	家族構成，DNA，血液型， 生体情報（静脈，虹彩）	普遍性あり， 変更が効かない
中間	身長，体重		
動的	髪の色，服装	血圧，脈拍，血液型， 診療履歴	本人特定には繋がらない

個人情報保護法による定義

- 個人情報（第2条①）

「個人情報」とは，生存する個人に関する情報であって，当該情報に含まれる氏名，生年月日その他の記述等により特定の個人を識別することができるものをいう

  - 映像，声，指紋，筆跡等により本人を識別しうる場合も「その他の記述」に含まれる(宇賀克也「個人情報保護法の逐条解説第4版」, p. 28）
  - そもそも（準でない）個人情報だった

購買データの例

個人データ

顧客ID	氏名	住所	身長	IPアドレス	店舗	日時	商品	数量
121	菊池一郎	東京都中野区	175cm	13.1.1.3	中野通り店	2014年4月21日17:09	メロンパン	1
240	佐藤克巳	東京都	180cm	7.8.8.8	新宿中央店	2014年4月21日17:10	ドリップコーヒー	1
121	菊池一郎	東京都中野区	175cm	13.1.1.10	中野通り店	2014年4月21日17:15	てりたまドッグ	1
355	伊藤浩明	東京都八王子	211cm	7.8.8.8	新宿中央店	2014年4月21日17:18	クロックドーナツ	50

同定性低減データ

仮名ID	（削除）	一般化住所	（削除）	仮名ID	店舗	日時	商品	数量
88		東京都		101	中野通り店	2014年4月21日17:09	メロンパン	1
89		東京都		102	新宿中央店	2014年4月21日17:10	ドリップコーヒー	1
88		東京都		103	中野通り店	2014年4月21日17:15	てりたまドッグ	1
90		東京都		102	新宿中央店	2014年4月21日17:18	クロックドーナツ	50

個人特定性低減データのリスク

- これで十分か？

仮名ID	（削除）	一般化住所	（削除）	仮名IP	店舗	日時	商品	数量
88		東京都		101	中野通り店	2014年4月21日17:09	メロンパン	1
89		東京都		102	新宿中央店	2014年4月21日17:10	ドリップコーヒー	1
88		東京都		103	中野通り店	2014年4月21日17:15	てりたまドッグ	1
90		東京都		102	新宿中央店	2014年4月21日17:18	クロックドーナツ	50

同一利用者は同一仮ID

他人が同一仮ID

詳細すぎる時刻

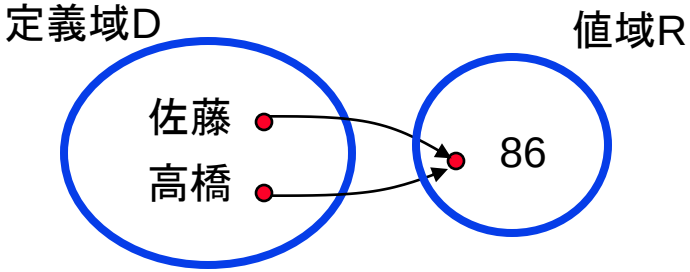
特徴的な履歴

「不可逆な番号(識別子)」の脅威

- 1. 総当り攻撃
  - 計算量的な一方向性に対する脆弱性
- 2. 決め打ち攻撃・辞書攻撃
  - 入力値を予測して，試行を繰り返す
- 3. 値域の大きさ，偏差からの漏えい
- 4. 特徴的な記録による攻撃
  - 著しい履歴，組合せによる識別性

不可逆性(一方向性)の種類

- 計算量的一方向性
  - 与えられたh(x) から, h(x) = h(y)となるyを見つけることが難しい
  - 定義域Dの全ての要素xについて探せば確実に見つかる.（総当り攻撃）
  - 実用上は十分. SHA1 (160 bit), SHA2 (512 bit)
- 情報論的不可逆性
  - 単射でない写像h



  - 理論的に安全
- 確率的不可逆性

  h(佐藤) = 86

  h(佐藤) = 92
  - 不確定的(最も安全)

変換しても特定が容易な例

- 値の種類が少ない．分布に偏りがある

時刻	注文
16:01	激辛担担麺
16:02	激辛担担麺
16:07	つけ麺
16:08	激辛担担麺
16:09	激辛担担麺
16:13	つけ麺

時刻	注文
16:01	A5e2efe09f28873
16:02	A5e2efe09f28873
16:07	E442e2151f8e4b4
16:08	A5e2efe09f28873
16:09	A5e2efe09f28873
16:13	E442e2151f8e4b4

ランダムな識別子に置換えても，元の情報の種類が少ないのでどちらか容易に推測できる．

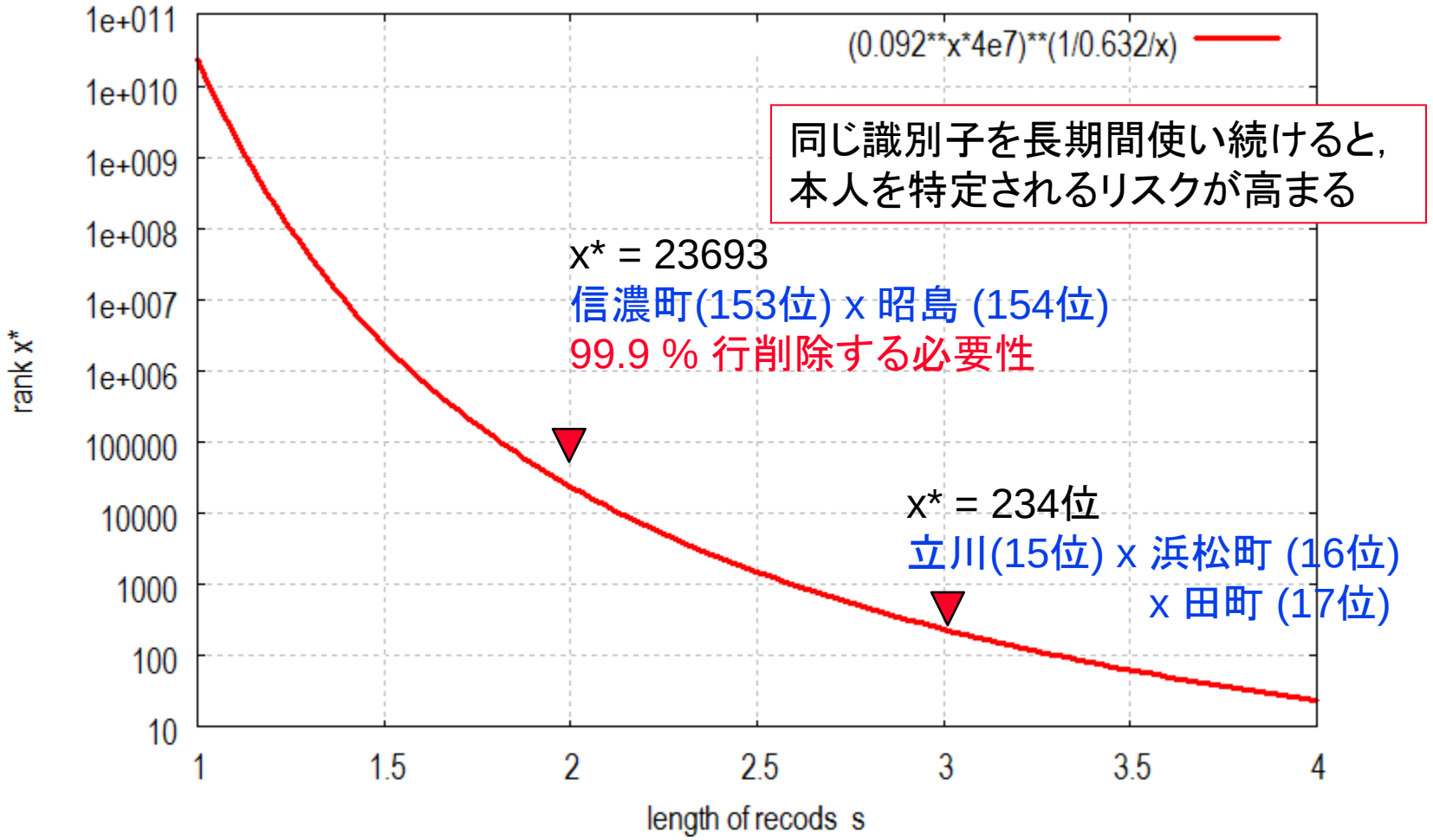
決め打ち攻撃(辞書攻撃)

- 一方向性ハッシュ関数による仮名匿名化
  - H(菊池，男性，4月1日生) = Y
- 辞書攻撃
  - 入力の値を推測して，ハッシュ値を特定
  - H(菊池，男性，1月1日生) ≠ Y,
    H(菊池，男性，1月2日生) ≠ Y, …..
    H(菊池，男性，4月1日生) = Y

定義域が限られるとき，識別子から個人名などの属性が漏えいする

確定的ID生成	確率的ID生成
QI+ハッシュ関数	ソルト(乱数)+ハッシュ関数
教科書RSA暗号	RSA+OAEP

履歴の長さ(s)に対する識別性



割当期間の長さ

- デバイスIDの有効期間と識別リスク

情報	有効期間	漏洩リスク
IPアドレス	数時間～数日	小さい
AD Truth	?	
クレジットカード	3年～6年	大きい（経済的被害）
パスポート番号	1年，5年，10年	？
メール，電話番号	数十年（転職，引越しがないまで）	小さい
口座番号	数十年（生涯？）	大きい（経済的被害）

- 漏洩時の被害を考慮して，「仮名ID」の有効期限を定める必要がある

「他の事業者と共有できない番号」

- 事業者A

	仮ID	購買	数
菊池	121	メロンパン	53
高橋	89	ホットドック	1

- 事業者B

	仮ID	購買	数
菊池	2100	コーヒー	53
高橋	2200	紅茶	1

問題点
(1) 時刻などの詳細な情報や特徴的な履歴から，121=2100 の対応が分かる危険性が残る.
(2)Aにとって121=菊池の対応が分かれば，（準でない）個人データのままではないか.

個人情報保護法による定義

- 個人データ（第2条②）

「個人情報データベース等」とは以下に挙げるものをいう
1. 特定の個人情報を電子計算機を用いて検索できるように体系的に構成したもの
2. （…)として政令で定めるもの

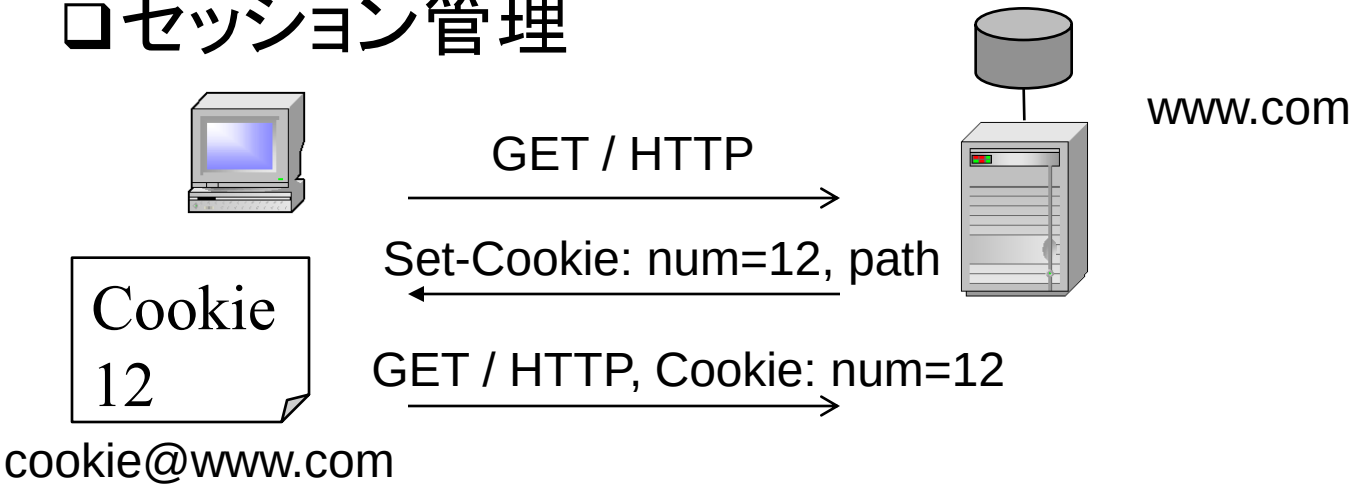
- 検索エンジンは，キーワードと同一文字列であれば法人名や地名を含めて検索するので「個人情報データベース等」ではない.
- 通信販売業者のデータベースで特定の個人に関する情報も抽出可能なシステムになっていれば「個人情報データベース等」である.（宇賀「個人情報保護法の逐条解説」p.30)

仮名IDについてまとめ

- 「不可逆な」には，計算量的，情報理論的など色々な種類がある．
- 様々な脅威（総当り，辞書攻撃，値域の大きさや精度によっては推測が容易，履歴を繋げることによる識別）
- そもそも保護法では，仮名IDがあるだけで「個人データ」に含まれる

HTTP Cookie について

- 仕様 (RFC 6265)
  - セッション管理



  - 他のドメインからはアクセスできない
  - クッキーファイルはローカルに残る．受け付けない設定もできる（オプトアウト）

準個人情報としてのHTTP cookie

- IPアドレスとの比較

	IPアドレス	クッキー
普遍性	動的（DHCP)	動的（更新）
照合容易性	時刻情報があれば ISPは照合可能	ウェブサーバ側の ログから照合可能
有効期限	数時間～数日	数時間～数年
外部観察性	容易	困難

- 結論
  - IPアドレスと同様の準個人情報である.

提案1

- 前提
  - 個人特定性は識別性と比例する（非特定識別ならば危険）
- 提案
  - 準個人情報1,2ともに削除する
  - 準個人情報3（履歴データ）は，識別不能にして，特徴的なデータを落とす．

識別不能性の判断

- 条件
  1. １次提供者（個人情報取扱い事業者）が見ても，レコード間の識別が出来ないこと
  2. 特定の個人を識別する仮名IDは持たない（または確率的，不確定的な仮名IDを振る）
  3. 全ての購買記録について，一意なレコードを削除する（k=2匿名性)
  4. 値域の大きさが十分に大きいこと

検討依頼事項 (4/18)

- 検討依頼事項
  - a)「準個人データ」から「個人特定性低減データ」への最低限の加工は、「準個人データ」で無くすことの妥当性 →「準個人データ」でなくても現行法「個人データ」で扱い可能
  - b)「個人データ」及び「準個人データ」から「個人特定性低減データ」への最低限の加工について、一般化、より適切な方法
  - d) 提供先（受領者）と提供元に課せられる制約について →属性の種類，値域，各レコード数を申告する.
  - e) Cookieは対象外よいか → No.

まとめ

- 番号を「不可逆」に振るだけでは不十分であり，辞書攻撃や統計値から識別されるリスクを考慮しなくてはならない
- 準個人情報の分類に対して，次の観点も必要である.
  - 動的・静的
  - 外部観察容易・困難
  - オプトアウト可能・不能

「普遍性のある生体的・特性的情報」

- 「オプトアウトが出来る」と「出来ない」

	外部から観察が容易	内部でないと観察できない
オプトアウト不能	身長，性別，体重(?)，肌の色，人種， 生体情報（顔，指紋）， 趣味，位置情報	DNA，
オプトアウト可能	髪の色，服装， 会員カードの提示（乗降履歴，購買履歴）， 匿名サービス（閲覧履歴）	生体情報（静脈，虹彩） 家族構成，住所，生年月日，学歴， 年収， 血圧，脈拍，診療履歴
	攻撃容易	

個人特定性低減データの考察その(2)

菊池浩明(明治大学)

この資料で主張したいこと

- 特定性を低減する為には，真面目に各種の「匿名化」処置を施す必要があり，様々な問題があって難しい.
- 技術レベルの異なる事業者がそれぞれ勝手に「低減データ」を作ってしまう危険あり
- 各種の低減データをその方法でなく，処置されたデータの特定性で定量化する一つの方法を導入する.

提案1（最も厳しく，最も安全）

- 事務局案(4月16日資料2)
  - 準個人情報1(所有物）は不可逆で共有不能な番号に置換える.
  - 準個人情報2(生体情報）は削除する.
  - 準個人情報3（履歴）は精度を落とす.
- 提案1.
  - 準個人情報1,2ともに削除する
  - 準個人情報3（履歴データ）は，識別不能にして，特徴的なデータを落とす．(完全にぶつ切り）

低減データの種類とそのリスク

高橋類型		事務局案（不可逆な識別子）	提案1 (識別子削除，ぶつ切り)
特定	1. 対応表	×	○
	2a. 後に登録	×	○
	2b. 顔と結びつき	○（準個人2)	○
	3. マッチング	○(共有禁止）	○
	4a. 知識攻撃	？	？
	4b. 悉皆性	？	？
個人連絡	6. 習慣的履歴	×	○
	7. ターゲット広告	×（対象外）	○（不可能）
プロファイリング	8. 利用期間，範囲の増大	×	○
	9. 風評被害	×	×
	10. 本人知らない	×	×

購買データの例

個人データ

顧客ID	氏名	住所	身長	IPアドレス	店舗	日時	商品	数量
121	菊池一郎	東京都中野区	175cm	13.1.1.3	中野通り店	2014年4月21日17:09	メロンパン	1
240	佐藤克巳	東京都	180cm	7.8.8.8	新宿中央店	2014年4月21日17:10	ドリップコーヒー	1
121	菊池一郎	東京都中野区	175cm	13.1.1.10	中野通り店	2014年4月21日17:15	てりたまドッグ	1
355	伊藤浩明	東京都八王子	211cm	7.8.8.8	新宿中央店	2014年4月21日17:18	クロックドーナツ	50

同定性低減データ

仮名ID	（削除）	一般化住所	（削除）	仮名ID	店舗	日時	商品	数量
88		東京都		101	中野通り店	2014年4月21日17:09	メロンパン	1
89		東京都		102	新宿中央店	2014年4月21日17:10	ドリップコーヒー	1
88		東京都		103	中野通り店	2014年4月21日17:15	てりたまドッグ	1
90		東京都		102	新宿中央店	2014年4月21日17:18	クロックドーナツ	50

基本定義

- レコード
  - ユーザ集合U = $\{i_1, \ldots, i_n\}$, nをユーザ数. アイテム集合をA. 時刻集合をTとする.
  - ユーザ識別子(仮名ID) i, 時刻t, アイテム uの組 r = (i, t, u) をレコードrと呼ぶ.
  - レコードの集合をデータセット S = $\{r_1, \ldots, r_m\}$と呼ぶ. mをレコード数とする.

仮名ID	日時	商品
88	2014年4月21日17:09	メロンパン
89	2014年4月21日17:10	コーヒー
88	2014年4月21日17:15	ドーナツ
90	2014年4月21日17:18	ドーナツ

S = $\{r_1, r_2, r_3, r_4\}$, U={88, 89, 90}
r_1 = (88, 2014年4月21日, メロンパン)
r_2 = (89, 2014年4月21日, コーヒー)
r_3 = (88, 2014年4月21日, ドーナツ)
r_4 = (90, 2014年4月21日, ドーナツ)

完全に安全な低減データ

- 1. 完全に一般化
  - m個のレコードが全て同一.
    - » 時刻とアイテムが同一
    - » たとえ本人でもどのレコードが自分が分からない

仮名ID	日時	商品
88	2014年4月	食品
89	2014年4月	食品
88	2014年4月	食品
90	2014年4月	食品

- 2. 完全に独立
  - m個のレコードの識別子が全て異なる
    - » レコード間の関係がぶつ切り.

仮名ID	日時	商品
1	2014年4月	食品1
2	2014年4月	食品2
3	2014年4月	食品3
4	2014年4月	食品4

- しかし，ここまで低減すると有用性がない，

完全一般化データのリスク

高橋類型		事務局案(不可逆識別子)	提案1 (識別子削除)	完全一般化
特定	1. 対応表	×	○	○
	2a. 後に登録	×	○	○
	2b. 顔と結び付	○(準個人2)	○	○
	3. マッチング	○(共有禁止)	○	○
	4a. 知識攻撃	?	?	○
	4b. 悉皆性	?	?	○
個人連絡	6. 習慣的履歴	×	○	○
	7. ターゲット広告	×(対象外)	○	○
プロファイリング	8. 利用期間	×	○	○
	9. 風評被害	×	×	×
	10. 本人知ら	×	×	×

提案2(現実的な低減の方法)

- 特定性を低減する手法には，列(行，セル)削除，アイテム一般化，トップ(ボトム)コーディング，サンプリング，摂動化などがあり，一般化することは出来ない.
- 代わりに，低減された特定性を定める一般指標(識別度)を定義して，その度合いで「低減」したことを提供者が示す.

提案2の目標

高橋類型		事務局案（不可逆識別子）	提案1 (識別子削除)	提案2(識別度による低減)
特定	1. 対応表	×	○	△
	2a. 後に登録	×	○	△
	2b. 顔と結び付	○（準個人2)	○	○
	3. マッチング	○(共有禁止）	○	○
	4a. 知識攻撃	？	？	？
	4b. 悉皆性	？	？	？
個人連絡	6. 習慣的履歴	×	○	△
	7. ターゲット広告	×（対象外）	○	△
プロファイリング	8. 利用期間	×	○	△
	9. 風評被害	×	×	×
	10. 本人知ら	×	×	×

識別度の定義

- アイテムについての利用者集合
  - アイテムaを含むレコードの識別子の集合U(a)
  - 例）U(メロンパン) = {88, 89}, U(ドーナツ) = {90}
  - アイテムaの利用者数 N(a) = | U(a) |
  - 例）N(メロンパン) = 2, N(ドーナツ) = 1
- データセットSの識別度
  - 最低識別度 $d^*_S = \mathrm{argmix}_{a \text{ in } A} N(a)$
  - 平均識別度 $d_S = 1/m \sum_{i=1,..,m} N(ai)$

例題(1/2)

S1 (オリジナル)		
仮名ID	T	A
88	2014年4月21日	メロンパン
89	2014年4月21日	メロンパン
90	2014年4月21日	ドーナツ
88	2014年4月22日	ドーナツ
91	2014年4月23日	メロンパン
92	2014年4月23日	アンパン

U(4月21日,メロンパン)={88}
U(4月22日,メロンパン)={89}

最小 $k^*_{S1} = 1$
平均 $k_{S1} = 1$

S2 (時刻一般化)		
仮名ID	T	A
88	2014年	メロンパン
89	2014年	メロンパン
90	2014年	ドーナツ
88	2014年	ドーナツ
91	2014年	メロンパン
92	2014年	アンパン

U(メロンパン)={88,89,91}
U(ドーナツ)={88,90}
U(アンパン)={92}
最小 $k^*_{S2} = 1$
平均 $k_{S2} = (3*3+2*2+1)/6 = 2.3$

S3 (行削除)		
仮名ID	T	A
88	2014年	メロンパン
89	2014年	メロンパン
90	2014年	ドーナツ
88	2014年	ドーナツ
91	2014年	メロンパン

U(メロンパン)={88,89,91}
U(ドーナツ)={88,90}

最小 $k^*_{S3} = 2$
平均 $k_{S3} = 13/5=2.6$

例題(2/2)

S4 (アイテム一般化)		
仮名ID	T	A
88	2014年	パン
89	2014年	パン
90	2014年	ドーナツ
88	2014年	ドーナツ
91	2014年	パン
92	2014年	パン

U(パン)={88,89,91,92}
U(ドーナツ)={88,90}

最小 $k^*_{S4} = 2$
平均 $k_{S4} = 10/3 = 3.3$

S5 (完全匿名化)		
仮名ID	T	A
88	2014年	食品
89	2014年	食品
90	2014年	食品
88	2014年	食品
91	2014年	食品
92	2014年	食品

U(食品)=U

最小 $k^*_{S5} = 6 = m$
平均 $k_{S5} = 6 = m$

考察

- 最小値k^*_Sがよいか平均k_Sがよいか.
- レコード平均がよいか，アイテム平均がよいか.
  - アイテム平均 $1/|A| \sum_{a \text{ in } A} N(a)$
  - Aの組み合わせ爆発で網羅するのが困難かも
- 同一利用者の識別子（仮名ID)をある期間で付け替える効果を定量化できないか.
- 低減された識別度のしきい値定めず，個人情報保護委員会が（事後規制で?)判断する.