La lluita contra el *Big Brother*

# Tècniques de reidentificació per atacar la privadesa

Vicenç Torra

Sta Coloma de Queralt, Juny 2007

# Índice

# Introducció

# Privadesa

Les tecnologies per la privadesa de les dades te com a objectiu fer complir el dret a la privadesa en la societat de la informació.

Són drets emparats per:

- Declaració Universal dels Drets Humans (article 12)
- Constitució Española (article 18)
- Llei Orgànica de Protecció de Dades de Caràcter Personal (15/1999)
- Llei de l'Agència Catalana de Protecció de Dades (5/2002)

# Privadesa

Mineria de dades preservant la privadesa i control d'inferència
(Statistical Disclosure Control, SDC)

- Necessitat de publicar (o cedir) les dades

Construcció del fitxer públic (a partir del fitxer original):

- Modificació de l'original amb mètodes d'emmascarament
- Generació de dades sintètiques

# Privadesa

Mineria de dades preservant la privadesa i control d'inferència
(Statistical Disclosure Control, SDC)

- Necessitat de publicar (o cedir) les dades

Construcció del fitxer públic (a partir del fitxer original):

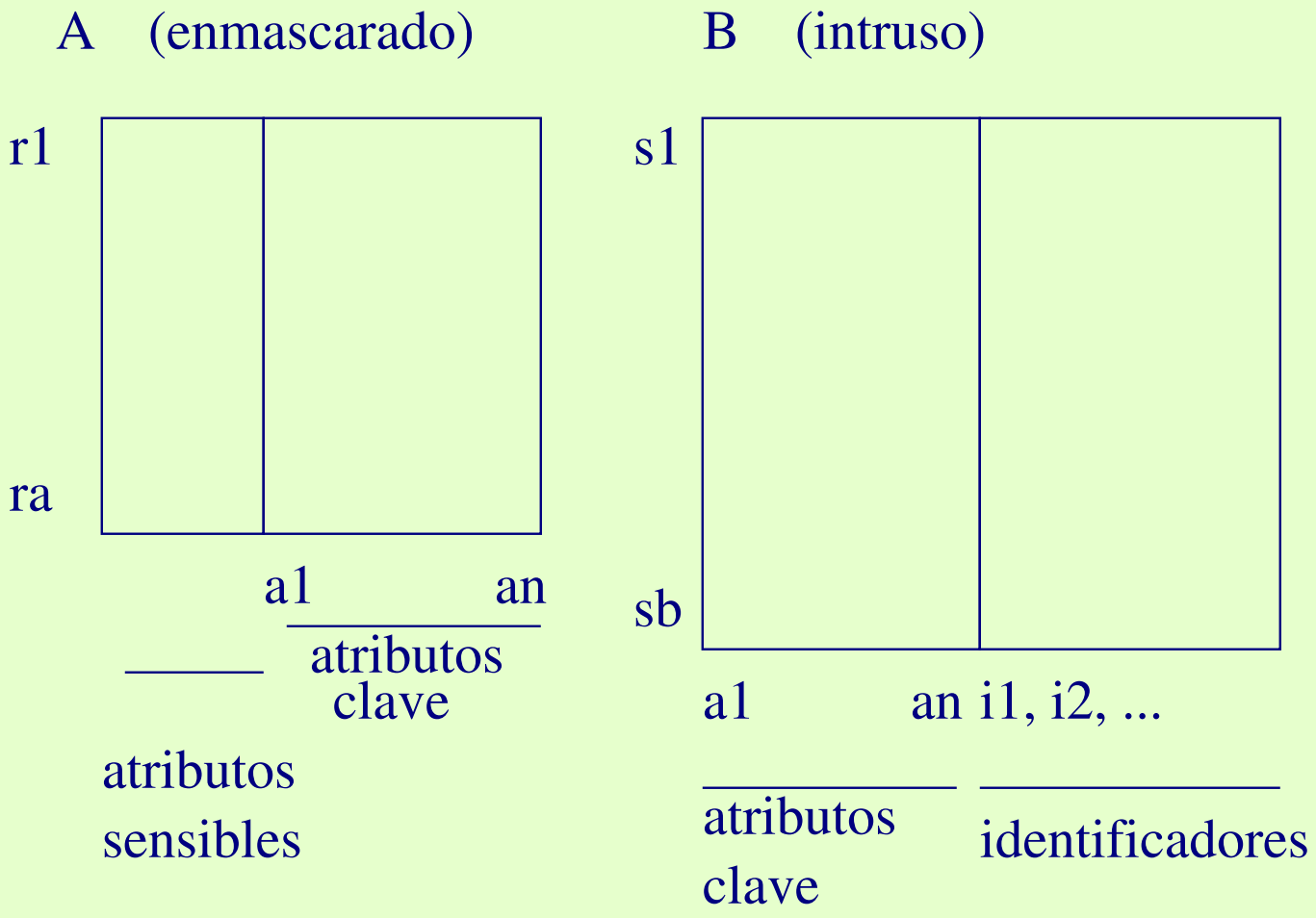
- Modificació de l'original amb mètodes d'emmascarament
- Generació de dades sintètiques

Fitxer publicat que

- Mantingui la validesa:
  resultats semblants pels diversos anàlisis de l'usuari
  $\longrightarrow$ mesures de pèrdua d'informació
- Asseguri la privadesa:
  informació confidencial no deduïble
  $\longrightarrow$ Mesures de risc de revelació

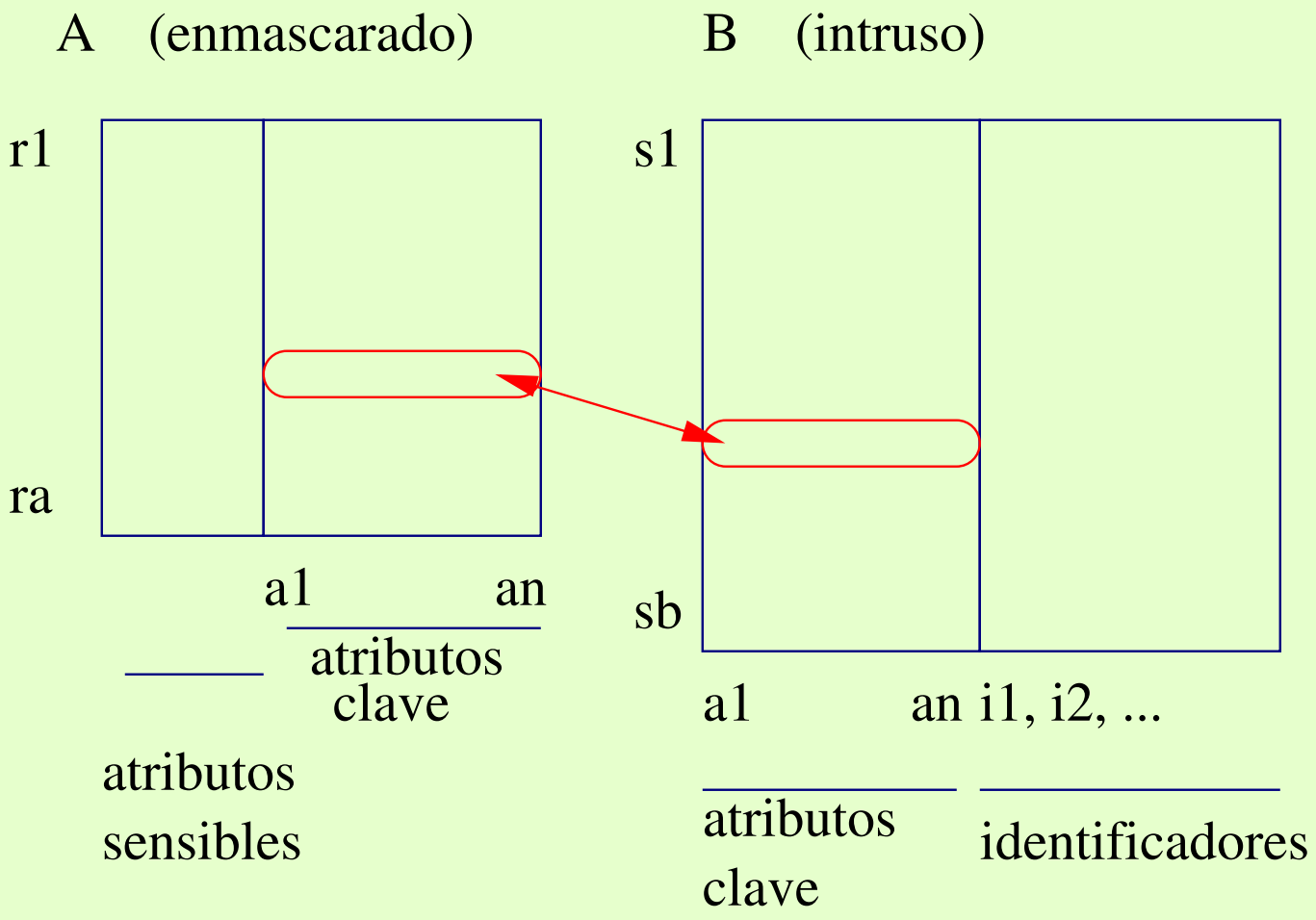
# Risc de revelació: escenari

## Mesura de risc de revelació



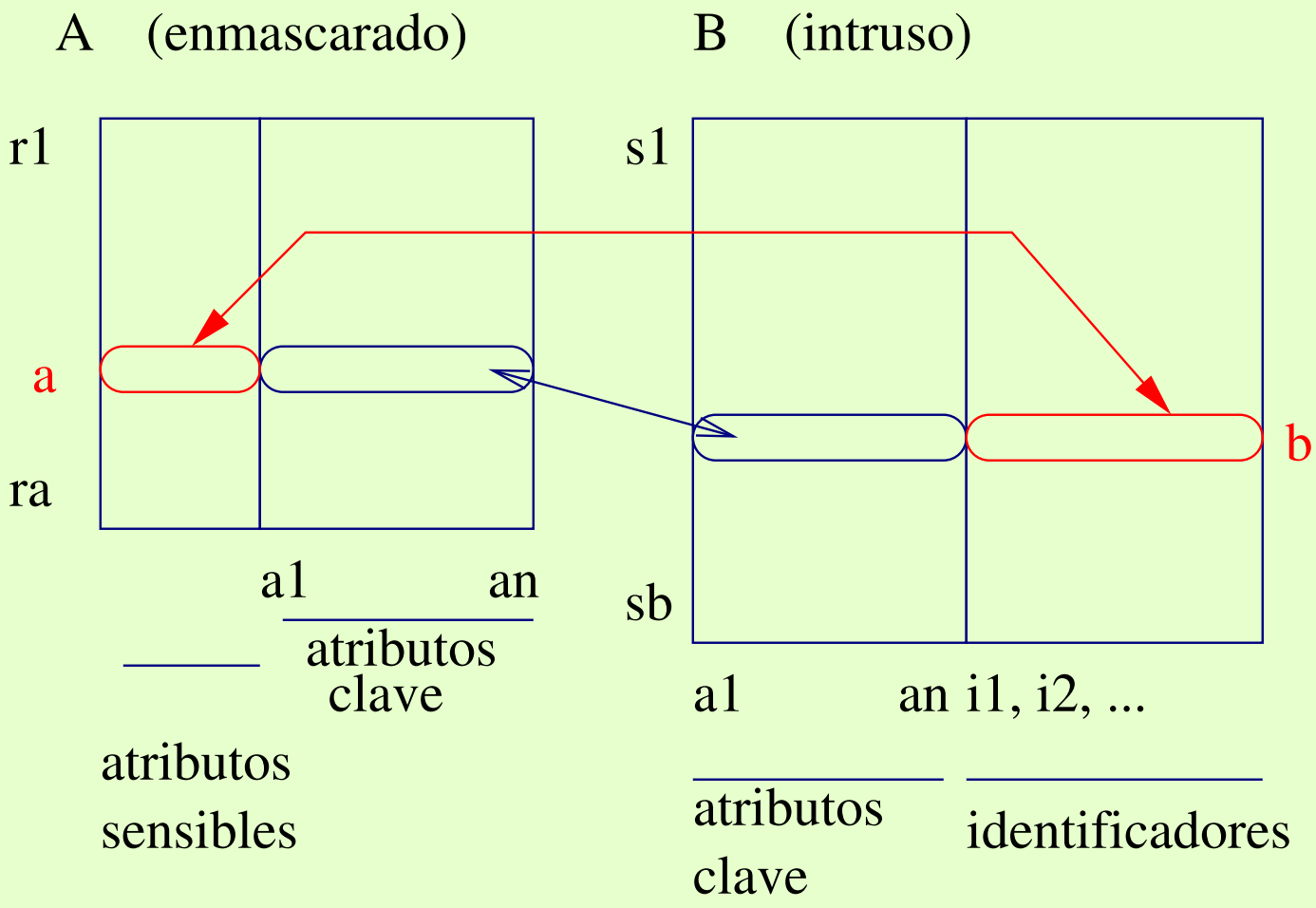
# Risc de revelació: escenari

## Mesura de risc de revelació: enllaç de registres



# Risc de revelació: escenari

Si $a$ és enllaçat amb $b$: deduïm informació confidencial sobre $b$.



# Risc de revelació: exemple

Fitxer original:

| Malaltia | ··· | Sexe | Estat civil | Ciutat | Edat |
|---|---|---|---|---|---|
| Cor | ··· | M | Casat | Barcelona | 33 |
| Embaras | ··· | F | Divorciat | Tarragona | 40 |
| Embaras | ··· | F | Casat | Barcelona | 36 |
| Apendicitis | ··· | M | Solter | Barcelona | 36 |
| SIDA | ··· | M | Solter | Gisclareny | 22 |
| Fractura | ··· | M | Vidu | Barcelona | 81 |

Fitxer públic: fitxer A

| Malaltia | ··· | Sexe | Estat civil | Ciutat | Edat |
|---|---|---|---|---|---|
| Cor | ··· | M | Casat | Barcelona | 33 |
| Embaras | ··· | F | Vidu-o-Divorciat | — | 40 |
| Embaras | ··· | F | Casat | — | 33 |
| Apendicitis | ··· | M | Solter | Barcelona | 40 |
| SIDA | ··· | M | Solter | Gisclareny | 25 |
| Fractura | ··· | M | Vidu-o-Divorciat | Barcelona | 81 |

# Risc de revelació: exemple

Fitxer original:

| Malaltia | ··· | Sexe | Estat Civil | Ciutat | Edat |
|---|---|---|---|---|---|
| Cor | ··· | M | Casat | Barcelona | 33 |
| Embaras | ··· | F | Divorciat | Tarragona | 40 |
| Embaras | ··· | F | Casat | Barcelona | 36 |
| Apendicitis | ··· | M | Solter | Barcelona | 36 |
| SIDA | ··· | M | Solter | Gisclareny | 22 |
| Fractura | ··· | M | Vidu | Barcelona | 81 |

Fitxer públic: fitxer A

| Malaltia | ··· | Sexe | Estat civil | Ciutat | Edat |
|---|---|---|---|---|---|
| Cor | ··· | M | Casat | Barcelona | 33 |
| Embaras | ··· | F | Vidu-o-Divorciat | — | 40 |
| Embaras | ··· | F | Casat | — | 33 |
| Apendicitis | ··· | M | Solter | Barcelona | 40 |
| SIDA | ··· | M | Solter | Gisclareny | 25 |
| Fractura | ··· | M | Vidu-o-Divorciat | Barcelona | 81 |

Fitxer de l'intrus: fitxer B

| Sexe | Estat Civil | Ciutat | Edat | | |
|---|---|---|---|---|---|
| M | Casat | Barcelona | 33 | ··· | Ricardo Gómez |
| F | Divorciat | Tarragona | 40 | ··· | María García |
| F | Casat | Barcelona | 36 | ··· | Juana Badía |
| M | Solter | Barcelona | 36 | ··· | Alberto Pérez |
| M | Solter | Gisclareny | 22 | ··· | Felipe Gil |
| M | Vidu | Barcelona | 81 | ··· | Antonio Fernández |

# Risc de revelació: exemple

Fitxer original:

| Malaltia | ··· | Sexe | Estat Civil | Ciutat | Edat |
|---|---|---|---|---|---|
| Cor | ··· | M | Casat | Barcelona | 33 |
| Embaras | ··· | F | Divorciat | Tarragona | 40 |
| Embaras | ··· | F | Casat | Barcelona | 36 |
| Apendicitis | ··· | M | Solter | Barcelona | 36 |
| SIDA | ··· | M | Solter | Gisclareny | 22 |
| Fractura | ··· | M | Vidu | Barcelona | 81 |

Fitxer públic: fitxer A

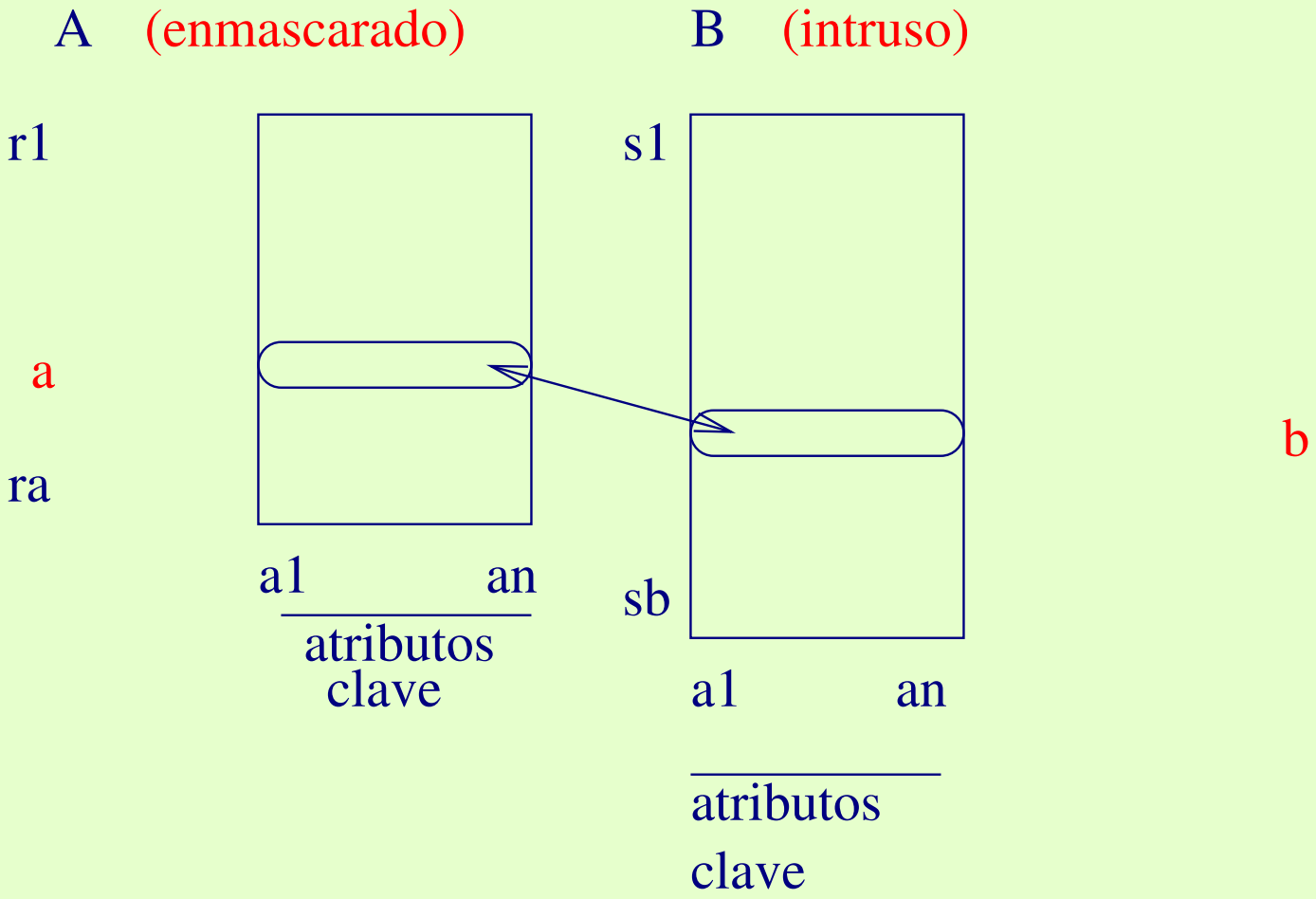
| Malaltia | ··· | Sexe | Estat civil | Ciutat | Edat |
|---|---|---|---|---|---|
| Cor | ··· | M | Casat | Barcelona | 33 |
| Embaras | ··· | F | Vidu-o-Divorciat | — | 40 |
| Embaras | ··· | F | Casat | — | 33 |
| Apendicitis | ··· | M | Solter | Barcelona | 40 |
| SIDA | ··· | M | Solter | Gisclareny | 25 |
| Fractura | ··· | M | Vidu-o-Divorciat | Barcelona | 81 |

Fitxer de l'intrus: fitxer B

| Sexe | Estat Civil | Ciutat | Edat | | |
|---|---|---|---|---|---|
| M | Casat | Barcelona | 33 | ··· | Ricardo Gómez |
| F | Divorciat | Tarragona | 40 | ··· | María García |
| F | Casat | Barcelona | 36 | ··· | Juana Badía |
| M | Solter | Barcelona | 36 | ··· | Alberto Pérez |
| M | Solter | Gisclareny | 22 | ··· | FELIPE GIL |
| M | Vidu | Barcelona | 81 | ··· | Antonio Fernández |

# Risc de revelació: avaluació

Risc de revelació: nombre de reidentificacions (original, protegit).



# Risc de revelació: reidentificació

Reidentificació per l'avaluació del risc

- Aplicables a tot tipus de dades:
  Atributs numèrics, categòrics
- Aplicables a diversos escenaris
  Intrusos amb accès a diverses bases de dades
- Aplicables als diferents mètodes d'emmascarament
  És fins i tot possible la reidentificació en el cas de dades sintètics

# Risc de revelació: reidentificació

Diferents algorismes de reidentificació
$\longrightarrow$ diferents estimacions del risc

Un intrús utilitzarà el millor mètode
$\longrightarrow$ estudiar els mètodes de re-identificació té interés

# Pèrdua d'informació

Construcció del fitxer públic (a partir del fitxer original):

- Modificació de l'original amb mètodes d'emmascarament
- Generació de dades sintètiques

Fitxer publicat que

- Mantingui la validesa:
  resultats semblants pels diversos anàlisis de l'usuari
  $\longrightarrow$ mesures de pèrdua d'informació
- Asseguri la privadesa:
  informació confidencial no deduïble
  $\longrightarrow$ Mesures de risc de revelació

# Pèrdua d'informació

Pèrdua mínima quan són semblants els ...

- resultats d'analitzar les dades originals
- resultats d'analitzar les dades protegides

Mesura de pèrdua d'informació:

- Promig de la variació entre

  valors, covariàncies, matrius de correlacions, mitjanes

# Compromis entre risc i pèrdua d'informació

Construcció del fitxer públic (a partir del fitxer original):

- Modificació de l'original amb mètodes d'emmascarament
- Generació de dades sintètiques

Fitxer publicat que

- Mantingui la validesa:
  resultats semblants pels diversos anàlisis de l'usuari
  $\longrightarrow$ Mesures de pèrdua d'informació
- Asseguri la privadesa:
  informació confidencial no deduïble
  $\longrightarrow$ Mesures de risc de revelació

$\longrightarrow$ Compromís entre la pèrdua d'informació i el risc

# Compromís entre risc i pèrdua d'informació

Comparació sistemàtica entre mètodes de protecció.

- Per exemple, mesurar el compromís amb un índex

$$score = 0.5 \cdot (IL + DR)$$

pèrdua d'informació (*IL: information loss*),
risc de revelació (DR)

Disminuir el risc de revelació $\rightarrow$ augmentar la pèrdua d'informació

# Compromís entre risc i pèrdua d'informació

Comparació sistemàtica entre mètodes de protecció.

- Per exemple, mesurar el compromís amb un índex

$$score = 0.5 \cdot (IL + DR)$$

pèrdua d'informació (*IL: information loss*),
risc de revelació (DR)

Podem calcular un índex per a cada parell *(mètode, paràmetre)*

- Millors mètodes: rank swapping i microagregació

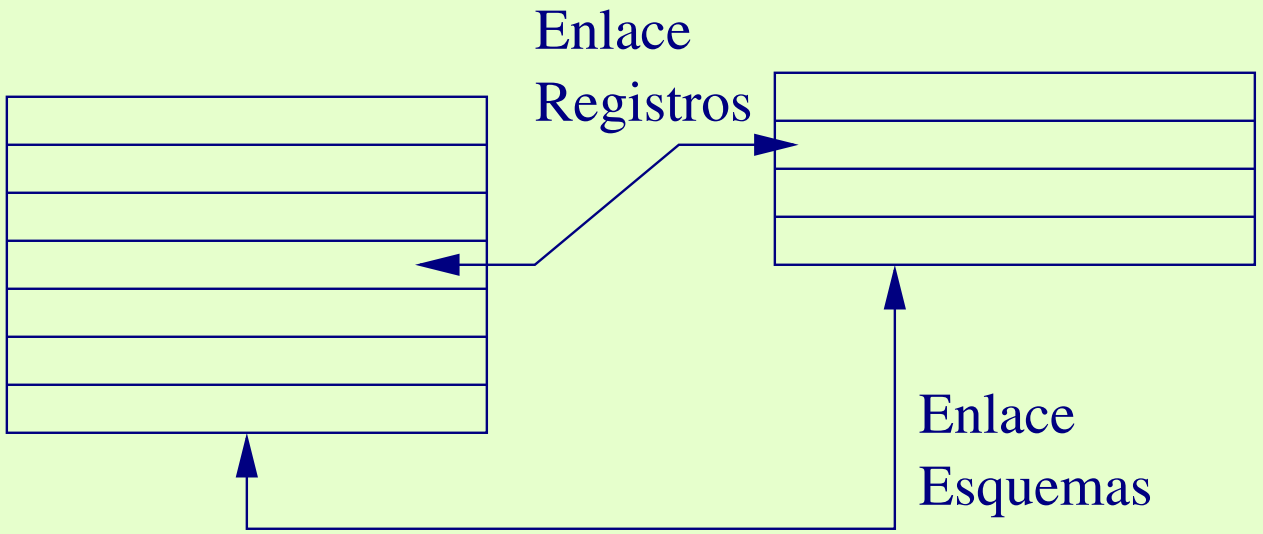
# Algorismes de reidentificació

# Algorismes de reidentificació
# Mètodes genèrics

# Reidentificació

Algorismes de reidentificació

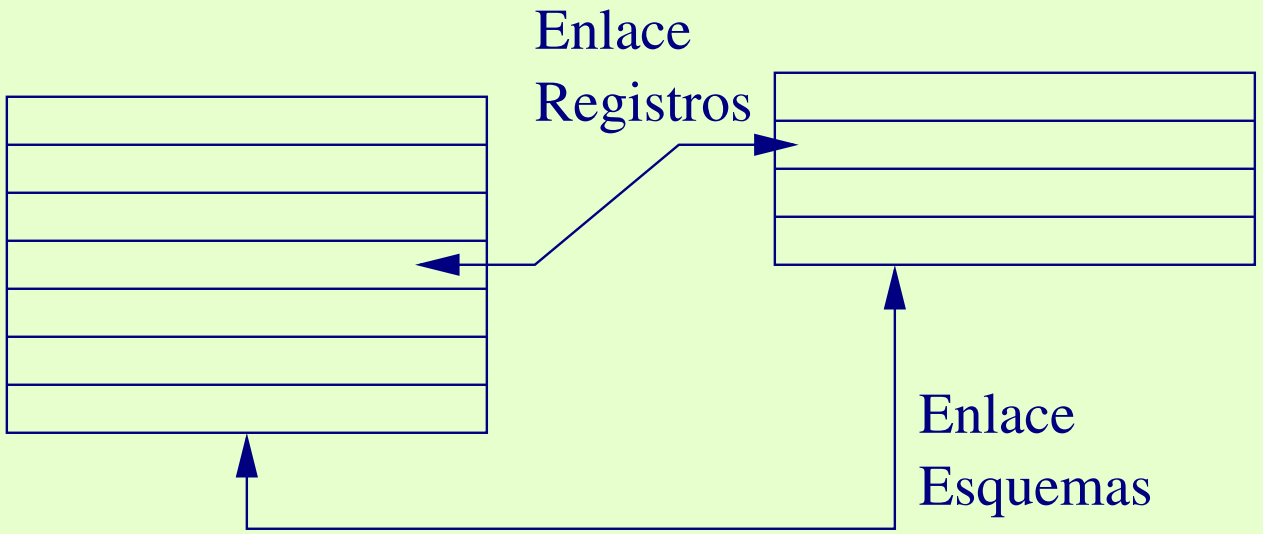
- Entre les dades subministrades per diverses fonts d'informació, establiment de relacions d'*identitat* entre objectes.
  - Consistència i qualitat de les dades
- En fitxers/bases de dades
  - Enllaç de registres
  - Enllaç d'esquemes



# Reidentificació

## Algorismes de reidentificació

- Entre les dades subministrades per diverses fonts d'informació, establiment de relacions d'*identitat* entre objectes.
  - Consistència i qualitat de les dades
- En fitxers/bases de dades
  - Enllaç de registres
  - Enllaç d'esquemes



# Enllaç d'esquemes (I)

Donats dos conjunts de dades $A$ i $B$,

- objectiu:
  trobar la correspondència adequada entre les variables d'$A$ i $B$

Algunes dificultats d'aquesta tasca són

- degudes al fet que la mateixa informació pot ser representada mitjançant variables amb noms diferents i, potser, semàntiques diferents $\rightarrow$ *adreça* (fitxer $A$) vs *carrer-número-ciutat-CP* (fitxer $B$).

# Enllaç d'esquemes (II)

Hi ha dues dimensions per classificar aquests mètodes:

**1a.** El tipus d'informació que utilitza el mètode:

Alguns mètodes només es basen en noms de variables i en els seus tipus de dades mentre que d'altres empren algun tipus d'informació estructural extreta de la base de dades.

Hem treballat amb enllaç de variables emprant com a informació estructural una aproximació suau de la funció de densitat de la variable (variables numèriques)

# Enllaç d'esquemes (III)

Hi ha dues dimensions per classificar aquests mètodes:

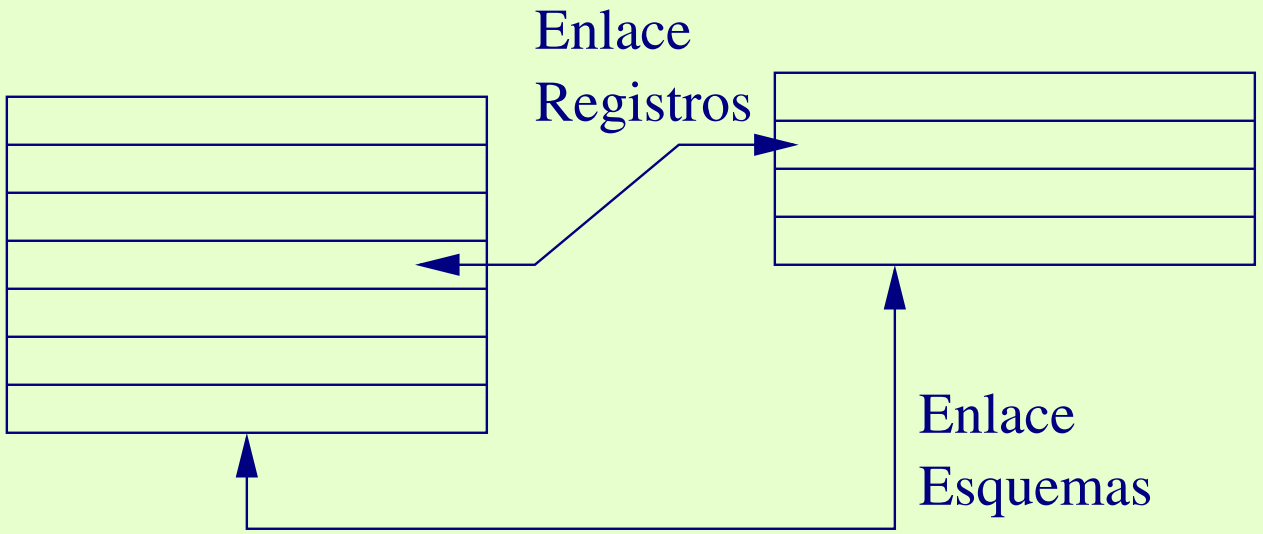
**2a.** Tipus de relacions permeses entre variables:

Alguns mètodes només permeten establir relacions attr-i amb attr-j (aquest cas es coneix per enllaç de variables), altres mètodes permeten relacions entre diverses variables (n:1, 1:m, n:m). Per exemple, pel cas 1:1 la variable *affiliation* (fitxer $A$) vs. la variable *organització* (fitxer $B$); pel cas 1:n l'associació entre la variable *adreça* (fitxer $A$) amb el conjunt de variables *carrer*, *nombre*, *ciutat* i *codi postal* (fitxer $B$).

# Reidentificació

## Algorismes de reidentificació

- Entre les dades subministrades per diverses fonts d'informació, establiment de relacions d'*identitat* entre objectes.
  - Consistència i qualitat de les dades
- En fitxers/bases de dades
  - Enllaç de registres
  - Enllaç d'esquemes



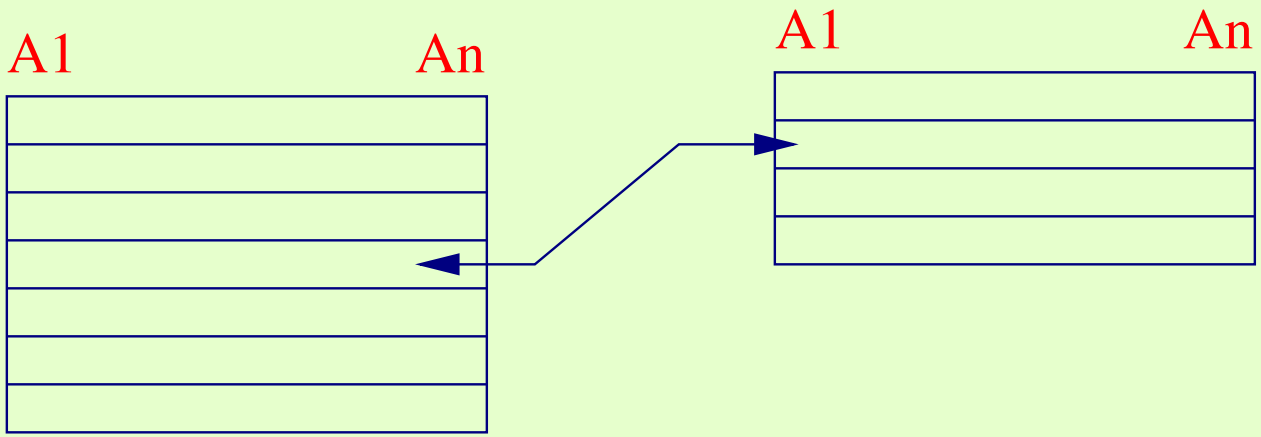
# Enllaç de registres

## Enllaç de registres

- Aplicació a un parell de fitxers $(A, B)$
- Atributs: numèrics, categòrics (escales ordinals, nominals)

## Enllaç de registres estàndard

- Els fitxers comparteixen atributs
- Dificultat: errors accidentals o intencionals en els fitxers



# Enllaç de registres estàndard

## Enllaç de registres estàndard

- Els fitxers comparteixen atributs
- Dificultat: errors accidentals o intencionals en els fitxers

# Enllaç de registres estàndard

Dades:

- Fitxers $A$, $B$

  Els individus d'$A$ i $B$ estan representats mitjançant els mateixos atributs

Mètodes:

- Classificació dels parells $(a, b)$ per $a \in A, b \in B$ en $LP$, $NP$:
  - Enllaç de registres probabilístic (PRL)
- Classificació d'un registre $b$ en una *classe* d'$A$:
  - Enllaç de registres basat en distàncies (DBRL)

# Enllaç de registres probabilístic (PRL)

Procediment:

- Per a cada parell de registres $(a, b)$ es calcula un índex $R(a, b)$ a partir de les probabilitats:
  - $P(coincidencia|\mathbf{M})$ de coincidència de valors condicionada a la correspondència de registres
  - $P(coincidencia|\mathbf{U})$ de coincidència de valors condicionada a la no correspondència de registres[1]
- Es classifiquen els parells a partir de l'índex d'acord amb dos llindars:
  1. Si $R(a, b) >$ *LlindarEnllaç*, el parell s'enllaça
  2. Si $R(a, b) <$ *LlindarNoEnllaç*, el parell no s'enllaça
  3. En qualsevol altre cas, el parell es classifica com *per revisar*

[1] $P(coincidencia|\mathbf{M})$, $P(coincidencia|\mathbf{U})$ s'estimen mitjançant l'algorisme EM

# Enllaç de registres probabilístic (PRL)

Paràmetres de l'algorisme:

- Llindars: a partir de les probabilitats d'error
  **(i)** probabilitat d'enllaçar un parell que no és una coinciencia real (falsos positius)
  $\rightarrow P(LP|\mathbf{U})$
  **(ii)** probabilitat de no enllaçar un parell que és una coincidència (falsos negatius)
  $\rightarrow P(NP|\mathbf{M})$

# Enllaç de registres basat en distàncies (DBRL)

Procediment:

- Cada registre $b$ de l'intrús s'enllaça amb el registre $a$ original més *proper*

Necessita:

- Definició d'una distància entre registres
  a partir de definicions de distància a nivell d'atributs

# Enllaç de registres basat en distàncies (DBRL)

Formalització:

basada en la regla de decisió de Bayes MAP (*màxim a posteriori*).

- Maximització de la probabilitat $P(\omega_i|x)$ calculada a partir de $P(x|\omega_i)$ i $P(\omega_i)$.
- Distribucions condicionades diferents (i $P(\omega_i)$ diferents) porten a distàncies diferents:
  - Distribució normal:
    - Atributs independents: distància euclidiana
    - Atributs no independents: distància de Mahalanobis

# Enllaç de registres basat en distàncies (DBRL)

- Altres distàncies: basades en les funcions de kernel
  - Transformació de les dades a un espai de grans dimensions:
$$x \longrightarrow \Phi(x) \in H$$
  - Aplicació de la reidentificació a aquest espai $H$

- Notació:
  - $K(x, y) = \Phi(x) \cdot \Phi(y)$, són les funcions de kernel

- Donats $a \in A$ i $b \in B$, la distància al quadrat en l'espai $H$ és:

$$||\Phi(a) - \Phi(b)||^2 = (\Phi(a) - \Phi(b))^2 =$$

$$= \Phi(a) \cdot \Phi(a) - 2\Phi(a) \cdot \Phi(b) + \Phi(b) \cdot \Phi(b) = K(a, a) - 2K(a, b) + K(b, b)$$

$\rightarrow$ No és necessari conèixer $\Phi(x)$, és suficient calcular $\Phi(x) \cdot \Phi(y)$

# Enllaç de registres basat en distàncies (DBRL)

- Procediment KDBRL:

  (enllaç de registres basats en funcions kernel)

  – Donats $A$ i $b$

  – Determinar el registre $a \in A$ més proper calculant

$$d_K(a,b)^2 = K(a,a) - 2K(a,b) + K(b,b)$$

# Enllaç de registres basat en distàncies (DBRL)

- Algunes funcions kernel:
  - $K_1(x,y) = e^{-\lambda||x-y||^2}$
  - $K_2(x,y) = e^{-(1/2)(x-y)^T\Sigma^{-1}(x-y)}$
  - $K_3(x,y) = (1 + x \cdot y)^d$
  - $K_4(x,y) = tanh((x \cdot y) + b)$

- KDBRL amb aquestes funcions:
  - KDBRL amb $d_{K_1}$ i $d_{K_2}$ equivalen a DBRL amb distància euclidiana i de Mahalanobis
  - KDBRL amb $d_{K_3}$ i $d = 1$ equival a DBRL amb distància euclidiana

# Enllaç de registres: comparació

## Basat en distàncies

- És més simple d'implementar
- És més fàcil incluir-hi informació subjectiva
- És difícil definir distàncies adequades per alguns atributs (per exemple, categòrics ordinals)
- És necessari establir els pesos dels diferents atributs

## Probabilístics

- És més complexe d'implementar
- Els paràmetres es determinen automàticament
- No és necessari establir els pesos
- Només fa falta definir les probabilitats d'error

# Enllaç de registres: aplicació, dades sintètiques

**Exemple:** reidentificació (1080 registres): dades "Census" vs. IPSO-A (generador de dades sintètiques)

| DBRL1 | DBRL2 | DBRLM-COV0 | DBRLM-COV | KDBRL | PRL |
|---|---|---|---|---|---|
| 145 | 133 | 135 | 123 | 146 | 133 |
| 91 | 75 | 126 | 60 | 89 | 82 |
| 95 | 87 | 137 | 66 | 94 | 103 |
| 98 | 87 | 129 | 62 | 97 | 86 |
| 23 | 40 | 123 | 67 | 24 | 97 |
| 104 | 92 | 93 | 84 | 100 | 92 |
| 59 | 65 | 63 | 57 | 61 | 65 |
| 94 | 85 | 89 | 68 | 91 | 86 |
| 109 | 104 | 106 | 44 | 106 | 103 |

DBRL1: attribute-standardizing implementation of distance-based record linkage (DBRL); DBRL2: distance-standardizing implementation of DBRL; DBRLM-COV and DBRLM-COV0: distance-based record linkage using Mahalanobis distance (covariances computed using the appropriate alignment or covariances set to zero); KDBRL: distance-based record linkage with kernel distance (polynomic kernel with d=2); PRL: probabilistic record linkage

# Algorismes de reidentificació
# Mètodes no estàndard

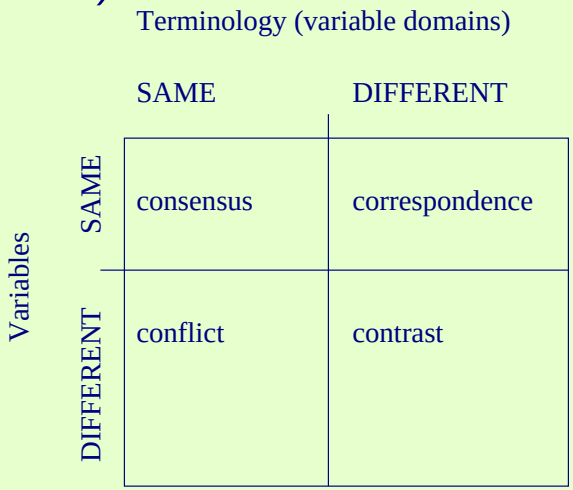
# Enllaç de registre no estàndard

Donats dos fitxers $A$ i $B$,

- objectiu:
  establir enllaços entre registres (de la mateixa entitat)

4 families diferents (d'acord amb la terminologia de l'extracció del coneixement a partir de grups, IA):

- d'acord amb la coincidència i no coincidència entre atributs i terminologia (la terminologia és el domini dels atributs/variables; això és, els termes emprats per avaluar els individus):



# Enllaç de registre no estàndard

Això és, hi ha 4 families diferents:

**Consensus:** El mateix atribut i la mateixa terminologia
**Correspondence:** Els mateixos atributs però terminologia diferent
**Contrast:** Variables diferents amb terminologia diferent
**Conflict:** Variables diferents amb la mateixa terminologia

Enllaç de registres estàndard: cau en el cas del consens o la correspondència (només es permeten petites variacions en quant a la terminologia: petites inconsistències entre noms, valors absents, etc. )

Altres mètodes d'enllaç de registres són possibles:

Per exemple, correspondència: quan el grau de no coincidència entre la terminologia no es limita a petites variacions dels noms (per exemple, termes completament diferents deguts a considerar granularitats diferents: població vs. comarca).

# Algorismes de reidentificació
# Mètodes adhoc

# Mètodes adhoc: introducció

- Cas:
  - protecció de dades
  - coneixem el mètode de protecció
  - coneixem els seus paràmetres

  $\longrightarrow$ intentem treure profit per augmentar la reidentificació

# Mètodes adhoc: introducció

- Cas:
  - protecció de dades
  - coneixem el mètode de protecció
  - coneixem els seus paràmetres

  $\longrightarrow$ intentem treure profit per augmentar la reidentificació

- Hem desenvolupat mètodes adhoc per a dos mètodes de protecció ...

# Mètodes adhoc: dos mètodes i una justificació

Comparació sistemàtica entre mètodes de protecció.

- Per exemple, mesurar el compromís amb un índex

$$score = 0.5 \cdot (IL + DR)$$

pèrdua d'informació (*IL: information loss*),
risc de revelació (DR)

Podem calcular un índex per a cada parell *(mètode, paràmetre)*

- Millors mètodes: rank swapping i microagregació

# Mètodes adhoc: Rank swapping, descripció

- Intercanvis d'ordre (*Rank swapping*) amb paràmetre $p$ per a un atribut
  - Ordenem els valors dels registres de forma creixent
    Aquí, els suposem ordenats $x_{ij} \leq x_{\ell j}$ per tot $1 \leq i < \ell \leq n$
  - Cada valor $x_{ij}$ s'intercanvia (aleatòriament, distribució uniforme) per un altre $x_{\ell j}$ en el rang $i < \ell \leq i + p$
  - Desfem l'ordenació inicial

- Aplicació:
  - Cada atribut és emmascarat de forma independent
  - $p$ controla el rank swapping: percentatge del nombre total de registres
    * Com més gran és $p$, més gran és la diferència entre els valors intercanviats
      $\longrightarrow$ disminueix el risc però incrementa la pèrdua d'informació
    * Com menor és $p$, més gran és el risc i menor és la pèrdua d'informació

# Mètodes adhoc: Rank swapping, descripció

- Exemple (aplicació del mètode per a cada atribut), $p = 2$.
  - 1,2,3,[4,5,6,7,8],9,10

| Fitxer original | | | | Fitxer emmascarat | | | |
|---|---|---|---|---|---|---|---|
| $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a'_1$ | $a'_2$ | $a'_3$ | $a'_4$ |
| 8 | 9 | 1 | 3 | 10 | 10 | 3 | 5 |
| 6 | 7 | 10 | 2 | 5 | 5 | 8 | 1 |
| 10 | 3 | 4 | 1 | 8 | 4 | 2 | 2 |
| 7 | 1 | 2 | 6 | 9 | 2 | 4 | 4 |
| 9 | 4 | 6 | 4 | 7 | 3 | 5 | 6 |
| 2 | 2 | 8 | 8 | 4 | 1 | 10 | 10 |
| 1 | 10 | 3 | 9 | 3 | 9 | 1 | 7 |
| 4 | 8 | 7 | 10 | 2 | 6 | 9 | 8 |
| 5 | 5 | 5 | 5 | 6 | 7 | 6 | 3 |
| 3 | 6 | 9 | 7 | 1 | 8 | 7 | 9 |

# Mètodes adhoc: Rank swapping, reidentificació

- Rank swapping record linkage (RS-RL)
  (mètode específic per a un escenari on s'utilitza rank swapping)
  - Si coneixem $p$, un registre de l'intrús (original) donat només pot generar, com a molt, $2p$ registres

# Mètodes adhoc: Rank swapping, reidentificació

- Rank swapping record linkage (RS-RL)
  (mètode específic per a un escenari on s'utilitza rank swapping)
  - Formalització: coneixem $p$ i un registre de l'intrús (original) donat
    * Per a cada valor $x_{ij}$ de l'intrús,
    * existeix un conjunt computable $B(x_{ij})$ de $2p$ registres emmascarats, que poden haver estat generats a partir del registre original $x_i$

# Mètodes adhoc: Rank swapping, reidentificació

- Registre (intrús) $x_2 = (6, 7, 10, 2)$, $p = 2$ i primer atribut $x_{21} = 6$
  - $B(x_{21} = 6) = \{(4, 1, 10, 10), (5, 5, 8, 1), (6, 7, 6, 3), (7, 3, 5, 6), (8, 4, 2, 2)\}$

| Fitxer original | | | | Fitxer emmascarat | | | | $B(x_{2j})$ |
|---|---|---|---|---|---|---|---|---|
| $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a'_1$ | $a'_2$ | $a'_3$ | $a'_4$ | $B(x_{21})$ |
| 8 | 9 | 1 | 3 | 10 | 10 | 3 | 5 | |
| 6 | 7 | 10 | 2 | 5 | 5 | 8 | 1 | X |
| 10 | 3 | 4 | 1 | 8 | 4 | 2 | 2 | X |
| 7 | 1 | 2 | 6 | 9 | 2 | 4 | 4 | |
| 9 | 4 | 6 | 4 | 7 | 3 | 5 | 6 | X |
| 2 | 2 | 8 | 8 | 4 | 1 | 10 | 10 | X |
| 1 | 10 | 3 | 9 | 3 | 9 | 1 | 7 | |
| 4 | 8 | 7 | 10 | 2 | 6 | 9 | 8 | |
| 5 | 5 | 5 | 5 | 6 | 7 | 6 | 3 | X |
| 3 | 6 | 9 | 7 | 1 | 8 | 7 | 9 | |

# Mètodes adhoc: Rank swapping, reidentificació

- Rank swapping record linkage (RS-RL)
  (mètode específic per a un escenari on s'utilitza rank swapping)
  - Formalització: coneixem $p$ i un registre de l'intrús (original) donat
    * Per a un valor $x_{ij}$ de l'intrús, $B(x_{ij})$
      ($2p$ registres generables a partir del registre original $x_i$)
    * Si coneixem diversos atributs $attr_1, \ldots, attr_c$,
      podem repetir el procés per a cada atribut $attr_j \rightarrow B(x_{ij})$
    * Naturalment, el registre emmascarat estarà en tots els $B(x_{ij})$

$$x'_\ell \in \cap_{1 \leq j \leq c} B(x_{ij}).$$

# Mètodes adhoc: Rank swapping, reidentificació

- Registre (intrús) $x_2 = (6, 7, 10, 2)$, $p = 2$ i segon atribut $x_{22} = 7$
  - $B(x_{22} = 7) = \{(5, 5, 8, 1), (2, 6, 9, 8), (6, 7, 6, 3), (1, 8, 7, 9), (3, 9, 1, 7)\}$

| Fitxer original | | | | Fitxer emmascarat | | | | $B(x_{2j})$ | |
|---|---|---|---|---|---|---|---|---|---|
| $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a'_1$ | $a'_2$ | $a'_3$ | $a'_4$ | $B(x_{21})$ | $B(x_{22})$ |
| 8 | 9 | 1 | 3 | 10 | 10 | 3 | 5 | | |
| 6 | 7 | 10 | 2 | 5 | 5 | 8 | 1 | X | X |
| 10 | 3 | 4 | 1 | 8 | 4 | 2 | 2 | X | |
| 7 | 1 | 2 | 6 | 9 | 2 | 4 | 4 | | |
| 9 | 4 | 6 | 4 | 7 | 3 | 5 | 6 | X | |
| 2 | 2 | 8 | 8 | 4 | 1 | 10 | 10 | X | |
| 1 | 10 | 3 | 9 | 3 | 9 | 1 | 7 | | X |
| 4 | 8 | 7 | 10 | 2 | 6 | 9 | 8 | | X |
| 5 | 5 | 5 | 5 | 6 | 7 | 6 | 3 | X | X |
| 3 | 6 | 9 | 7 | 1 | 8 | 7 | 9 | | X |

# Mètodes adhoc: Rank swapping, reidentificació

- De forma semblant:
  - $B(x_{21} = 6) = \{(4, 1, 10, 10), (5, 5, 8, 1), (6, 7, 6, 3), (7, 3, 5, 6), (8, 4, 2, 2)\}$
  - $B(x_{22} = 7) = \{(5, 5, 8, 1), (2, 6, 9, 8), (6, 7, 6, 3), (1, 8, 7, 9), (3, 9, 1, 7)\}$
  - $B(x_{23} = 10) = \{(5, 5, 8, 1), (2, 6, 9, 8), (4, 1, 10, 10)\}$
  - $B(x_{24} = 2) = \{(5, 5, 8, 1), (8, 4, 2, 2), (6, 7, 6, 3), (9, 2, 4, 4)\}$

- La intersecció d'aquests conjunts ...
  - és un únic registre $(5, 5, 8, 1)$.
    $\longrightarrow$ és l'enllaç correcte
    En cas d'haver-n'hi més d'un, aplicariem un enllaç de registres estàndard.

# Mètodes adhoc: Rank swapping, reidentificació



## RS-RL: Rank Swapping Record Linkage

**Data sets:**

- **Census (1080 records & 13 attributes)**
- **EIA (4092 records & 10 attributes)**

**Rank swapping configurations:**

- **p = 2 … 20**

**Score modifications:**

- **DR = 0.166 RSLD+ 0.166 DLD+ 0.166 PLD+ 0.5 ID**

# Mètodes adhoc: Rank swapping, reidentificació

Disclosure Risk Evaluation using Advanced Record Linkage

New record linkage methods

## RS-RL: Rank Swapping Record Linkage

| | Census | | | EIA | | |
|---|---|---|---|---|---|---|
| | RSLD | DLD | PLD | RSLD | DLD | PLD |
| rs 2 | 77.73 | 73.52 | 71.28 | 43.27 | 21.71 | 16.85 |
| rs 4 | 66.65 | 58.40 | 42.92 | 12.54 | 10.61 | 4.79 |
| rs 6 | 54.65 | 43.76 | 22.49 | 7.69 | 7.40 | 2.03 |
| rs 8 | 41.28 | 32.13 | 11.74 | 6.12 | 5.98 | 1.12 |
| rs 10 | 29.21 | 23.64 | 6.03 | 5.60 | 5.19 | 0.69 |
| rs 12 | 19.87 | 18.96 | 3.46 | 5.39 | 4.87 | 0.51 |
| rs 14 | 16.14 | 15.63 | 2.06 | 5.28 | 4.55 | 0.32 |
| rs 16 | 13.81 | 13.59 | 1.29 | 5.19 | 4.54 | 0.23 |
| rs 18 | 12.21 | 11.50 | 0.83 | 5.20 | 4.54 | 0.22 |
| rs 20 | 10.88 | 10.87 | 0.59 | 5.15 | 4.36 | 0.18 |

26

# Mètodes adhoc

- També s'han considerat mètodes adhoc
  - microagregació

# Conclusions

# Conclusions

- Diversitat de mètodes de reidentificació
- Com més gran el nombre de reidentificacions, major és el risc
- Es poden construir mètodes específics de reidentificació que incrementen el risc respecte els mètodes no específics.