Protecció de la privadesa de microdades mitjançant MDAV

Autor: Javier Vilalta Bautista
Tutor: Javier Parra Arnau
Professor: Helena Rifà Pous

Grau d'Enginyeria Informàtica
Tecnologies de la informació

2 de gener de 2020

Crèdits/Copyright

for the USGenWeb Census Project http://www.us-census.org/

1910 U.S. Military and Naval Federal Census
http://ftp.us-census.org/pub/usgenweb/census/_mil/1910/t624-1784/

This Census was transcribed by Linda Talbott and proofread by L. Talbott
for the USGenWeb Census Project®, http://www.us-census.org/

FITXA DEL TREBALL FINAL

Títol del treball:	*Protecció de la privadesa de microdades mitjançant MDAV*
Nom de l'autor:	*Javier Vilalta Bautista*
Nom del col·laborador/a docent:	*Javier Parra Arnau*
Nom del PRA:	*Helena Rifà Pous*
Data de lliurament (mm/aaaa):	*01/2020*
Titulació o programa:	*Grau d'Enginyeria Informàtica*
Àrea del Treball Final:	*Treball de final de grau*
Idioma del treball:	*Català*
Paraules clau	*MDAV, privacitat, microdades*

Resum del Treball (màxim 250 paraules): *Amb la finalitat, context d'aplicació, metodologia, resultats i conclusions del treball*

Aquest treball presenta una implementació de l'algorisme MDAV per tal de protegir un conjunt de dades de tal manera que no es perdin les seves característiques generals, i puguin protegir la privacitat dels individus i, alhora, donar prou informació per ser rellevants en estudis estadístics.

La metodologia utilitzada ha estat la implementació de l'algorisme i la seva avaluació amb diferents conjunts de dades.

Finalment, s'avaluen els inconvenients i limitacions del k-anonimat i es compara amb altres criteris que tenen millor comportament contra determinats tipus d'atacs.

Abstract (in English, 250 words or less):

This work presents an implementation of the MDAV algorithm in order to protect a dataset so that its general characteristics are not lost, and can protect the privacy of individuals, and at the same time provide enough information to be relevant in statistical studies.

The methodology used has been the implementation of the algorithm and its evaluation with different datasets.

Finally, the drawbacks and limitations of k-anonymity are evaluated and compared with other criteria that have better behavior against certain types of attacks.

Índex

Figures i taules

Índex de figures

Índex de taules

1.Introducció

1.1. Introducció/Prefaci

A mesura que les millores en tecnologia permeten mantenir més volum i més varietat de dades específiques sobre individus, augmenta la preocupació per la privacitat. La seva publicació, que pot ser molt necessària per a la realització d'estudis científics, s'ha de fer en condicions que permetin garantir la privacitat dels individus implicats, però sense perdre validesa pel seu estudi. En aquest conflicte, s'han desenvolupat una sèrie de tècniques per aconseguir aquest difícil equilibri.

A més d'aquestes tècniques, necessitem algun tipus de mesura per tal de quantificar-les de manera objectiva: un d'aquests indicadors és el de k-anonimat, que es defineix de la següent forma: un conjunt de dades proporciona k-anonimat si la informació de cada persona que hi conté no es pot distingir de, al menys, k-1 altres individus que també estiguin a les dades.

Amb aquests precedents, en aquest treball de final de grau treballarem el concepte de k-anonimat, des del punt de vista teòric, així com pràctic, mitjançant la implementació d'un algorisme que permet garantir el k-anonimat sobre un conjunt de dades.

Tot i que també es volia treballar el concepte de l-diversitat, per tal de veure les limitacions del concepte de k-anonimat i les diferències en implementació, al final no ha estat possible per restriccions de temps.

1.2. Descripció/Definició

En aquest treball ens preocupem de la necessitat de protegir la privadesa d'individus i institucions, en el context de la publicació d'informació que pot ser rellevant per a investigació. Podem imaginar el cas d'una institució, com ara un hospital, que té estadístiques d'una determinada malaltia i que pot pensar en publicar aquestes dades per tal d'ajudar a futurs investigadors a nous descobriments que ajudin en el tractament. Com ho fem, però, de manera que les dades siguin útils per la investigació, però a l'hora protegim la privacitat de cada malalt? Com mesurem aquesta privacitat?
D'entre els molts mecanismes que existeixen i que es comenten més endavant, un d'ells és l'algorisme MDAV, que garanteix el compliment de la característica de k-anonimat en un conjunt de dades.
El resultat d'aquest treball serà la implementació d'un algorisme de MDAV funcional, junt amb algunes proves amb alguns conjunts de dades per tal de validar el seu funcionament.

1.3. Objectius generals

Els objectius, en base a l'enunciat del projecte, seran els següents:

- Estudiar la privadesa de dades des d'un punt de vista algorísmic
  Aquest punt servirà de marc general de les explicacions de la resta de punts i ens basarem en tota la bibliografia del TFG.

- Analitzar avantatges i limitacions del requisit de k-anonimat
  En aquest punt, tractarem de veure els avantatges i limitacions del k-anonimat, tal com es descriu a [2]

- Implementar l'algorisme MDAV
  Es tracta d'implementar l'algorisme MDAV-genèric, tal com es defineix a [4]

- Estudiar l'equilibri entre privadesa i distorsió de dades per a diferents microdades
  Una vegada tinguem els algorismes implementats, haurem d'estudiar com es comporten, comparant la privades proporcionada amb la distorsió causada a les dades, per la qual cosa, haurem d'utilitzar els diferents criteris definits a la bibliografia.

- Millorar l'algorisme de MDAV per a que satisfaci el requisit de l-diversitat
  El requisit de l-diversitat se'ns explica a [5], que millora la seguretat del requisit de k-anonimat. En aquest mateix document, se'ns prova que, en general, és possible substituir un criteri per l'altra en algorismes de k-anonimat per la seva semblança i, per tant, podrem canviar el criteri a l'algorisme MDAV genèric anterior.
  Finalment, aquest últim objectiu no s'ha pogut assolir per falta de temps.

1.4. Metodologia i procés de treball

La metodologia utilitzada per tal de realitzar la implementació ha estat la de fer el desenvolupament de manera àgil, començant per objectius fàcilment assolibles, implementant parcialment l'algorisme pel diferents tipus de dades i, a partir d'aquí, anar progressant fins a tenir tota la casuística implementada.

1.5. Planificació

El projecte s'ha dividit en tasques, de manera que es puguin utilitzar els recursos i planificar de manera més eficient. A la taula següent es pot veure la descomposició realitzada:

		Temps		
		Durada	Inici	Final
Tasques	**PAC1**	**20 dies**	**18/09/19**	**07/10/19**
	Redacció	17 dies	18/09/19	06/10/19
	Lliurament PAC1		07/10/19	07/10/19
	PAC2	**28 dies**	**08/10/19**	**04/11/19**
	Integració feedback PAC1	1 dia	08/10/19	08/10/19
	Estudi de la bibliografia	7 dies	09/10/19	15/10/19
	Implementació algorisme MDAV	8 dies	16/10/19	23/10/19
	Redacció del document de PAC2	7 dies	24/10/19	30/10/19
	Implementació algorisme MDAV	5 dies	31/10/19	04/11/19
	Lliurament PAC2		*04/11/19*	*04/11/19*
	PAC3	**28 dies**	**05/11/19**	**02/12/19**
	Revisió situació projecte	1 dia	05/11/19	05/11/19
	Integració feedback PAC2	1 dia	06/11/19	06/11/19
	Estudi de la bibliografia	7 dies	07/11/19	13/11/19
	Implementació algorisme MDAV	7 dies	14/11/19	20/11/19
	Proves sobre dades	7 dies	21/11/19	27/11/19
	Redacció del document de PAC3	5 dies	28/11/19	02/12/19
	Lliurament PAC3		*02/12/19*	*02/12/19*
	Memòria i producte final	**31 dies**	**03/12/19**	**02/01/20**
	Revisió situació projecte	1 dia	03/12/19	03/12/19
	Integració feedback PAC3	1 dia	04/12/19	04/12/19
	Proves sobre dades	8 dies	05/12/19	12/12/19
	Ampliació amb l-diversitat	7 dies	13/12/19	19/12/19
	Preparar productes	2 dies	20/12/19	21/12/19
	Recopilar lliçons apreses	1 dia	22/12/19	22/12/19
	Redacció de la memòria	10 dies	23/12/19	01/01/20
	Elaboració de l'*Abstract* en català i anglès	1 dia	02/01/20	02/01/20
	Lliurament memòria i productes finals		*02/01/20*	*02/01/20*
	Lliurament de la presentació virtual	**7 dies**	**03/01/20**	**09/01/20**

	Preparació i gravació vídeo presentació	7 dies	03/01/20	09/01/20
	Lliurament de la presentació virtual		*09/01/20*	*09/01/20*
	Tribunal d'avaluació	**5 dies**	**13/01/20**	**17/01/20**
	Presentació a tribunal d'avaluació	5 dies	13/01/20	17/01/20
	Fi del projecte		*17/01/20*	*17/01/20*

Taula 1: Planificació

A continuació, podem veure aquesta planificació en forma de diagrama de Gantt:

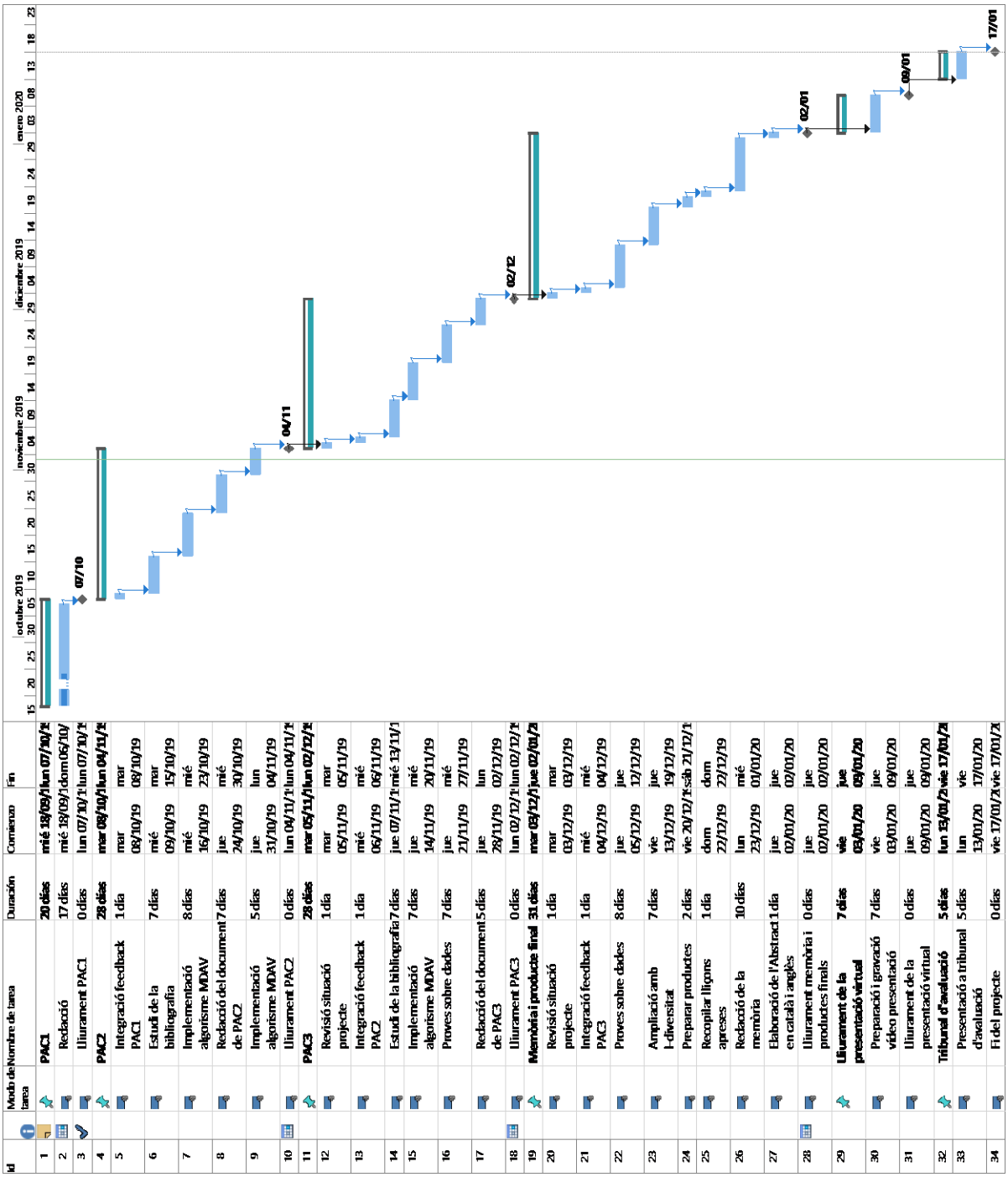

Figura 1: Diagrama de Gantt

Per restriccions de temps, no ha estat possible acabar la implementació de l'algorisme amb l'ampliació de l-diversitat. El problema principal ha estat que les correccions de l'algorisme genèric i les proves realitzades han ocupat més temps del previst.

1.6. Estructura de la resta del document

A la resta del document trobarem les següents seccions:

- Base teòrica del treball
  En aquest apartat farem una breu visió sobre el problema de la publicació de microdades i veurem alguns plantejaments formals que ens ajudaran a entendre el treball.

- Implementació realitzada de l'algorisme MDAV
  En aquesta part, farem una explicació de la implementació realitzada, així com mostrarem una sèrie de dades resultants.

- Estat de l'art
  En aquest apartat, parlarem de les limitacions del criteri de k-anonimat, així com de les vulnerabilitats i de quins plantejaments alternatius s'estan fent per evitar-los.

2. Plantejament teòric

2.1 Introducció

En aquest apartat revisarem els conceptes teòrics que hi ha al darrera del problema, així com un petit resum de la problemàtica de la revelació de microdades i el compromís entre la privacitat i la utilitat de les mateixes.

Aquest problema no és nou, històricament, els organismes oficials han hagut de fer pública informació estadística rellevant per a la investigació, però les solucions existents actualment no són adequades per l'entorn actual, amb conjunts més grans de dades i amb les noves necessitats d'anàlisi que alguns actors demanen: mineria de dades, anàlisi de costos, detecció del frau, etc., que requereixen dades més específiques a nivell de persona.

Algunes de les eines utilitzades fins ara són:

- Bases de dades estadístiques: aquestes bases de dades estan dissenyades per fer consultes a nivell agregat, protegint els elements individuals.

- Bases de dades multinivell: en aquest tipus, les dades estan assignades a nivells de seguretat, de manera que el sistema torna només informació dels nivells als que tenim permís, intentant sempre que no es puguin inferir dades d'un nivell a partir de les dades d'un altre. El problema d'aquest enfoc és que no es pot garantir al cent per cent que això sempre sigui així

També és important considerar que l'estudi de la privacitat no és el mateix que la seguretat, donat que aquesta última es preocupa de l'accés a les dades, mentre que la privacitat el que vol garantir és que, tot i tenir accés a les dades, no puguem arribar mai a conclusions respecte a persones individuals.

Una pràctica habitual és eliminar els identificadors explícits, com ara nom, telèfon o adreça. Però, en molts casos, la resta de les dades poden ser combinades per identificar, amb un grau alt de certesa, els individus, amb l'ajuda d'altres dades públicament disponibles. Per exemples, és possible utilitzar les dades gènere, data de naixement, ZIP i etnicitat d'una base de dades mèdica i combinar-la amb el cens, per obtenir amb un grau alt de certesa l'individu.

2.2 Marc formal

A continuació, indicarem algunes definicions importants a la resta del document:

Dades: informació específica d'una persona conceptualment organitzada en forma de taula de files (o registres o tuples) que no són necessàriament úniques i columnes (o camps o atributs) que sí són úniques (no està dues vegades el nom, per exemple)

Inferència: arribar a un nou fet en base a altra informació

Revelació (*disclosure*): el coneixement d'informació explícita o inferida no previst respecte a una persona

Control de revelacions (*disclosure control*): qualsevol intent de limitar o identificar revelacions respecte a un conjunt de dades

Quasi-identificador: conjunt d'atributs que, junts, poden identificar unívocament un individu amb l'ajuda d'una altra taula (com ara dades públiques) La identificació d'aquests quasi-identificadors no sempre és evident.

k-anonimat: donada una taula i un quasi-identificador associada amb ella, la taula es diu satisfà k-anonimat si i només si cada seqüència de valors del quasi-identificador apareix al menys k vegades.

2.3 Classificació de les microdades

Les microdades es poden classificar en els següents tipus:

- Continues
  Quan l'atribut és numèric i es poden aplicar operacions aritmètiques.

- Categòriques
  Una atribut és categòric si pren valor en un conjunt finit de valors possibles i no es poden aplicar operacions aritmètiques.
  A la seva vegada, es poden dividir en ordinals i nominals:
  - Ordinal
    Quan a l'atribut té sentit aplicar-li els operadors d'ordre, com ara <=, màxim, mínim, etc.
  - Nominal
    Quan no té sentit aplicar un ordre a les categories.

2.4 Definició de l'objectiu

Amb tot el definit anteriorment, el nostre objectiu consisteix en, donat un conjunt de microdades V, alliberar un conjunt de microdades alternatiu V' tal que:

- El risc de revelació d'informació és baix
  També anomenat d'identificació o reidentificació dels enquestats
- La pèrdua d'informació és baixa.

Més formalment, podem definir una puntuació:

$$Score(V, V') = \frac{IL(V, V') + DR(V, V')}{2}$$

On:

IL: Mesura de la pèrdua de la informació

DR: Mesura del risc de revelació

Per a la consecució d'aquest objectiu, els mètodes de protecció de microdades poden aplicar dues tècniques per generar un conjunt protegit:

- Emmascarar les dades
  D'aquests, n'hi ha dos tipus:
  - Amb pertorbació: amb aquests mètodes, poden aparèixer combinacions noves i desaparèixer d'altres però sempre preservant les propietats estadístiques. Alguns exemples són la microagregació i l'afegit de soroll.
  - Sense pertorbació: Aquest mètodes no distorsionen les dades originals. Alguns exemples són la generalització i la supressió.

- Generar dades sintètiques, però preservant algunes propietats estadístiques

Existeixen moltes formes possibles de concretar aquest procés. A la implementació, es veurà més en detall el cas concret de l'algorisme MDAV com a eina per aconseguir un conjunt de dades amb k-anonimat.

3.Implementació

3.1 Explicació de la implementació

L'algorisme MDAV s'ha implementat en llenguatge *Python*. L'estructura general del programa és la següent:

- Un mòdul `mdav`
- Programa principal que fa servir el mòdul

La forma de treballar amb el programa és mitjançant dos arxius:

- Arxiu en format CSV amb les dades
- Arxiu en format CSV amb les metadades
  Per convenció, s'entén que l'arxiu amb les metadades té el mateix nom que l'arxiu de dades.

Els diferents tipus de dades que s'esperen són:

- Contínues
- Categòriques (ordinals i nominals)

Per qüestions d'optimització es va afegir un tipus fictici per ignorar determinades columnes, però finalment s'ha decidit no fer-lo servir.

A la seva vegada, el mòdul `mdav` està composat dels següents

- `common.py`
  Aquest mòdul inclou la definició bàsica dels tipus de dades
- `helper.py`
  Aquest mòdul inclou funcions auxiliars per carregar arxius en format CSV i el seu arxiu de metadades vinculat.
- `mdav_generic.py`
  Aquest mòdul és el que fa el càlcul MDAV pròpiament i es veu una mica més endavant.
- `metric.py`
  Aquest mòdul calcula una sèrie de paràmetres, com ara mitjanes i variàncies per tal de comparar els diferents conjunts de dades generats.

L'algorisme principal es pot veure a continuació, amb la documentació amb l'algorisme explicat a la documentació:

```
def calc_mdav_generic(
        dataset: Dataset,
        k: int) -> list:
    """
    Function calc_mdav_generic
    """
    # region Auxiliary functions...

    # region Main functions to calculate averages, distances and clusters...

    logger.debug(f'Computing MDAC generic for {len(dataset.records)} rows')

    private_data = []
    while len(dataset.records) >= 3 * k:
        logger.debug(f'1. Iterating with {len(dataset.records)} records')

        columns, means, stdevs = compute_auxiliary_tables(dataset)

        # (1a) Compute the average record x̃ of all records in R. The average
        #      record is computed attribute-wise.
        x_avg = compute_average_record(dataset)
        logger.debug(f'1a. Average record is {x_avg}')

        # (1b) Consider the most distant record xr to the average record x̃
        #      using an appropriate distance
        xr = find_most_distant_record(dataset, x_avg)
        logger.debug(f'1b. Most distant record to average (xr) is {xr}')

        # (1c) Find the most distant record xs from the record xr considered in
        #      the previous step
        xs = find_most_distant_record(dataset, xr)
        logger.debug(f'1c. Most distant record to average record (xs) is {xs}')

        # (1d) Form two clusters around xr and xs , respectively. One cluster
        #      contains xr and the k − 1 records closest to xr. The other
        #      cluster contains xs and the k − 1 records closest to xs.
        # (1e) Take as a new dataset R the previous dataset R minus the
        #      clusters formed around xr and xs in the last instance of
        #      Step 1d.
        cluster_xr, dataset = extract_cluster(dataset, xr, k - 1)
        columns, means, stdevs = compute_auxiliary_tables(cluster_xr)
        cluster_xr_avg = compute_average_record(cluster_xr)
        logger.debug(f'1d. Cluster xr average record is {xr}')
        private_data.append(cluster_xr_avg)
```

```
        cluster_xs, dataset = extract_cluster(dataset, xs, k - 1)
        columns, means, stdevs = compute_auxiliary_tables(cluster_xs)
        cluster_xs_avg = compute_average_record(cluster_xs)
        logger.debug(f'1d. Cluster xs average record is {xs}')
        private_data.append(cluster_xs_avg)
        logger.debug(f'1e. New dataset size is {len(private_data)}')

    if len(dataset.records) >= 2 * k:

        logger.debug(
            f'2. Entering step 2 with {len(dataset.records)} records (>=2k)'
        )

        columns, means, stdevs = compute_auxiliary_tables(dataset)

        # (2a) Compute the average record x̃ of the remaining records in R
        x_avg = compute_average_record(dataset)
        logger.debug(f'2a. Average record is {x_avg}')
        # (2b) Find the most distant record xr from x̃
        xr = find_most_distant_record(dataset, x_avg)
        logger.debug(f'2b. Most distant record to average (xr) is {xr}')
        # (2c) Form a cluster containing xr and the k − 1 records closest to xr
        cluster_xr, dataset = extract_cluster(dataset, xr, k - 1)
        columns, means, stdevs = compute_auxiliary_tables(cluster_xr)
        cluster_xr_avg = compute_average_record(cluster_xr)
        logger.debug(f'2c. Cluster xr average record is {xr}')
        private_data.append(cluster_xr_avg)
        # (2d) form another cluster containing the rest of records

    else:

        logger.debug(
            f'2. Entering step 2 with {len(dataset.records)} records (<2k)'
        )
        # else (less than 2k records in R) form a new cluster with the
        # remaining records

    columns, means, stdevs = compute_auxiliary_tables(dataset)
    cluster_rest = compute_average_record(dataset)
    logger.debug(f'2/2d. Average record of rest of dataset is {cluster_rest}')
    private_data.append(cluster_rest)

    logger.debug(f'2. New dataset size is {len(private_data)}')

    return private_data
```

3.2 Dades de proves

Les dades utilitzades per fer les proves i els estudis són les següents:

- Sample 1

  Conjunt de dades fictícies, amb la intenció de recollir una casuística dels diferents tipus de dades. L'estructura és la següent:

Surname	name	year	balance

- Dades del cens dels EEUU

  Les dades que es faran servir seran històriques, del cens dels EEUU en concret

  Les dades s'han obtingut mitjançant aquestes adreces:

  https://www.census.gov/prod/2000pubs/cff-2.pdf

  https://1940census.archives.gov/

  https://www.census.gov/history/www/genealogy/decennial_census_records/census_records_2.html

  http://www.us-census.org

  Els exemples que s'estan utilitzant en concret són:

  - 1900 U.S. Military and Naval Federal Census

    http://us-census.org/pub/usgenweb/census/_mil/1900/t623-1842/ed009-pg019a.txt

    Census Year 1900
    Microfilm Roll #T623-1842
    Name of Military or Naval Station, or Vessel: U.S.S. Buffalo
    State --
    Country --
    Seaport Gibraltar
    Arm of Service U. S. Navy

    A l'arxiu auxiliar corresponent es pot veure l'estructura i els tipus utilitzats

  - 1910 U.S. Military and Naval Federal Census

    http://us-census.org/pub/usgenweb/census/_mil/1910/t624-1784/ed007-pg001a.txt

    Census Year 1910 CENSUS-DAY: April 15, 1910
    Microfilm Roll #T624-1784
    State --

County (Province) --
Township --
Incorporated-Place --
Institution Post of Camp Avery, Corregidor, Isle P.I.

A l'arxiu auxiliar corresponent es pot veure l'estructura i els tipus utilitzats

Tots dos arxius s'han convertir a CSV manualment, per donar una estructura més clara.

En tots els casos, s'ha generat tots els conjunts de corresponents als valors de k del tres al deu.

3.3 Resultats

A continuació, s'indiquen els resultats obtinguts per a diferents mesures obtingudes. En tots els casos, la columna *file* indica sobre quin arxiu s'ha fet el càlcul de la mesura i, dintre del nom, s'indica el valor de la *k* que s'ha fet servir per aplicar l'MDAV: per exemple, l'arxiu *sample1-mdav-generic-4-mean.csv* indica que s'ha aplicat *mdav* per 4-anonimat a l'arxiu *sample1.csv*. Quan el nom no inclou *mdav-generic*, com ara *sample1-mean.csv*, vol dir que el càlcul de mitjanes, en aquest cas, s'ha fet sobre l'arxiu original.

Algunes de les taules tenen molta informació i no es poden veure clarament. Adjunt al projecte s'inclouen els arxius Excels originals per tal de poder revisar-los millor.

El que s'ha de considerar en aquest cas és el següent:

- Per construcció, MDAV preserva mitjanes i variàncies
- Els valors individuals, els quantils, les covariàncies, les correlacions i les freqüències no es preserven.
- Quan més puja la k, més distorsió es produeix en les estadístiques no preservades.

És per aquest motiu que escollim aquestes mesures, amb l'objectiu de veure tant el comportament de mesures que es preserven com el comportament de mesures que no es preserven:

- Mitjana (*Mean*)
  En aquest cas, es calcula la mitjana per a cada valor. Aquest càlcul serveix per a tots els tipus de dades.

- Variància (*Variance*)
  Aquest valor només es calcular pels atributs continus.

- Quantil 50% (*Quantile 50*)

Aquest valor només es calcular pels atributs continus.

- Quantil 75% (*Quantile 75*)
  Aquest valor només es calcular pels atributs continus.

- Freqüències (*Frequencies*)
  Aquest valor és una mitjana de les freqüències i serveix per tenir una idea de com varien els diferents tipus de dades no continus (ordinals i nominals)

3.3.1 Dades fictícies

A continuació, els resultats per l'arxiu *sample1* amb dades fictícies:

file	surname	name	year	balance
sample1-mean.csv	Martinez	Pol	1988	19.923,77
sample1-mdav-generic-3-mean.csv	Rodriguez	Maria	1985	19.426,54
sample1-mdav-generic-4-mean.csv	Garcia	Jan	1989	19.292,61
sample1-mdav-generic-5-mean.csv	Martinez	Pol	1986	18.986,28
sample1-mdav-generic-6-mean.csv	Martinez	Jan	1989	18.871,07
sample1-mdav-generic-7-mean.csv	Martinez	Jan	1987	19.127,36
sample1-mdav-generic-8-mean.csv	Martinez	Emma	1990	18.350,92
sample1-mdav-generic-9-mean.csv	Martinez	Jan	1986	18.032,98
sample1-mdav-generic-10-mean.csv	Lopez	Pol	1988	18.353,48

Taula 2: Mitjanes de sample1

file	balance
sample1-variance.csv	302.901.118,81
sample1-mdav-generic-3-variance.csv	311.889.498,31
sample1-mdav-generic-4-variance.csv	310.630.808,72
sample1-mdav-generic-5-variance.csv	311.116.447,83
sample1-mdav-generic-6-variance.csv	290.081.412,93
sample1-mdav-generic-7-variance.csv	279.690.894,18
sample1-mdav-generic-8-variance.csv	304.684.810,09
sample1-mdav-generic-9-variance.csv	305.598.902,91
sample1-mdav-generic-10-variance.csv	305.482.447,63

Taula 3: Variàncies de sample1

file	balance
sample1-quantile50.csv	22.544,00
sample1-mdav-generic-3-quantile50.csv	18.991,80
sample1-mdav-generic-4-quantile50.csv	20.929,69

sample1-mdav-generic-5-quantile50.csv	18.880,17
sample1-mdav-generic-6-quantile50.csv	18.352,18
sample1-mdav-generic-7-quantile50.csv	19.071,37
sample1-mdav-generic-8-quantile50.csv	17.638,82
sample1-mdav-generic-9-quantile50.csv	16.525,30
sample1-mdav-generic-10-quantile50.csv	18.447,50

Taula 4: Quantil 50 de sample1

file	balance
sample1-quantile75.csv	34.168,26
sample1-mdav-generic-3-quantile75.csv	33.968,44
sample1-mdav-generic-4-quantile75.csv	33.913,40
sample1-mdav-generic-5-quantile75.csv	34.016,80
sample1-mdav-generic-6-quantile75.csv	32.732,82
sample1-mdav-generic-7-quantile75.csv	32.956,15
sample1-mdav-generic-8-quantile75.csv	32.749,71
sample1-mdav-generic-9-quantile75.csv	33.465,97
sample1-mdav-generic-10-quantile75.csv	34.819,76

Taula 5: Quantil 75 de sample1

file	surname	name	year
sample1-frequency.csv	13,3	14,778	2,714
sample1-mdav-generic-3-frequency.csv	4,3	4,778	1,536
sample1-mdav-generic-4-frequency.csv	3,2	3,556	1,28
sample1-mdav-generic-5-frequency.csv	2,5	2,778	1,389
sample1-mdav-generic-6-frequency.csv	2,333	2,333	1,167
sample1-mdav-generic-7-frequency.csv	2,25	2	1,2
sample1-mdav-generic-8-frequency.csv	1,5	1,875	1,071
sample1-mdav-generic-9-frequency.csv	1,625	1,625	1
sample1-mdav-generic-10-frequency.csv	1,714	1,5	1

Taula 6: Freqüències de sample1

Els resultats quadren amb el que estava previst, tot i que hauríem de fer una anàlisi estadística més exhaustiva per validar que les variacions són les esperades. En tot cas, sí que és cert que la poca quantitat de dades provoca algunes distorsions en alguns valors.

3.3.2 Cens del 1900

A continuació, els resultats per l'arxiu *ed009-pg019a.txt* amb dades del cens de 1900. Tot i que no és 100% correcte, s'ha considerat l'any com a continu, per aplicar tots els casos en aquest arxiu:

File	LAST NAME	FIRST NAME	OR CLASS	TOWN	STATE	RACE	SEX	MONTH	YEAR	AGE	W-D	PLACE	BIRTHPL	BIRTHPL2	READ	WRITE	ENG?
ed009-pg019a-mean.csv	Smith	John	Landsman	Philadelphia	New York	W	M	Nov	1.873,55	22	S	Pennsylvania	Ireland	Ireland	Yes	Yes	Yes
ed009-pg019a-mdav-generic-3-mean.csv	Smith	John	Landsman	Philadelphia	New York	W	M	Mar	807,552	21	S	Pennsylvania	Ireland	Ireland	Yes	Yes	Yes
ed009-pg019a-mdav-generic-4-mean.csv	Smith	John	Landsman	Philadelphia	New York	W	M	Aug	1.053,85	22	S	Pennsylvania	Ireland	Ireland	Yes	Yes	Yes
ed009-pg019a-mdav-generic-5-mean.csv	Smith	John	Landsman	Philadelphia	New York	W	M	Nov	1.131,99	21	S	Pennsylvania	Ireland	Ireland	Yes	Yes	Yes
ed009-pg019a-mdav-generic-6-mean.csv	Richards	John	Landsman	Philadelphia	New York	W	M	May	1.278,40	22	S	Pennsylvania	Ireland	Ireland	Yes	Yes	Yes
ed009-pg019a-mdav-generic-7-mean.csv	Sullivan	George	Landsman	Philadelphia	New York	W	M	June	1.393,10	22	S	Pennsylvania	Ireland	Ireland	Yes	Yes	Yes
ed009-pg019a-mdav-generic-8-mean.csv	Kane	George	Landsman	Philadelphia	New York	W	M	Aug	1.302,55	22	S	Pennsylvania	Ireland	Ireland	Yes	Yes	Yes
ed009-pg019a-mdav-generic-9-mean.csv	Kido	George	Landsman	Philadelphia	New York	W	M	Dec	1.426,28	22	S	Pennsylvania	Ireland	Ireland	Yes	Yes	Yes
ed009-pg019a-mdav-generic-10-mean.csv	King	John	Landsman	Philadelphia	New York	W	M	Nov	1.596,27	22	S	Pennsylvania	Ireland	Ireland	Yes	Yes	Yes

Taula 7: Mitjanes d'ed009-pg019a

file	YEAR
ed009-pg019a-variance.csv	5.140,61
ed009-pg019a-mdav-generic-3-variance.csv	864.875,60
ed009-pg019a-mdav-generic-4-variance.csv	869.681,20
ed009-pg019a-mdav-generic-5-variance.csv	847.304,62
ed009-pg019a-mdav-generic-6-variance.csv	768.671,85
ed009-pg019a-mdav-generic-7-variance.csv	679.070,64
ed009-pg019a-mdav-generic-8-variance.csv	754.412,80
ed009-pg019a-mdav-generic-9-variance.csv	649.390,50
ed009-pg019a-mdav-generic-10-variance.csv	452.763,55

Taula 8: Variàncies d'ed009-pg019a

file	YEAR
ed009-pg019a-quantile50.csv	1.878
ed009-pg019a-mdav-generic-3-quantile50.csv	1,881
ed009-pg019a-mdav-generic-4-quantile50.csv	1.868,33
ed009-pg019a-mdav-generic-5-quantile50.csv	1.872,38
ed009-pg019a-mdav-generic-6-quantile50.csv	1.872,80
ed009-pg019a-mdav-generic-7-quantile50.csv	1.876,17
ed009-pg019a-mdav-generic-8-quantile50.csv	1.874,14
ed009-pg019a-mdav-generic-9-quantile50.csv	1.875,50
ed009-pg019a-mdav-generic-10-quantile50.csv	1.876,56

Taula 9: Quantil 50 d'ed009-pg019a

file	YEAR
ed009-pg019a-quantile75.csv	1.880
ed009-pg019a-mdav-generic-3-quantile75.csv	1.876,50
ed009-pg019a-mdav-generic-4-quantile75.csv	1.877,50

ed009-pg019a-mdav-generic-5-quantile75.csv	1.878,25
ed009-pg019a-mdav-generic-6-quantile75.csv	1.877,80
ed009-pg019a-mdav-generic-7-quantile75.csv	1.878,83
ed009-pg019a-mdav-generic-8-quantile75.csv	1.878,86
ed009-pg019a-mdav-generic-9-quantile75.csv	1.879
ed009-pg019a-mdav-generic-10-quantile75.csv	1.879,83

Taula 10: Quantil 75 d'ed009-pg019a

file	LASTNAME	FIRSTNAME	ORCLASS	TOWN	STATE	STR-NUMBER	RACE	SEX	MONTH	AGE	W-D	MAR	PLACE	BIRTHPL	BIRTHPL2	IMMI	READ	WRITE	ENG?
ed009-pg019a-frequency.csv	1,188	1,664	8,612	3,57	17,667	1,154	172,25	344,5	32,81	20,879	114,833	38,278	10,132	10,766	10,6	21,531	229,667	229,667	229,667
ed009-pg019a-mdav-generic-3-frequency.csv	1,07	1,239	4,145	2,131	6,909	1,157	57	114	12	9,12	38	28,5	4,071	4,1145	4,071	8,444	76	76	76
ed009-pg019a-mdav-generic-4-frequency.csv	1,043	1,326	5,029	2,515	6,577	1,118	42,75	85,5	9	6,577	57	21,375	3,638	3,717	3,5562	8,143	85,5	85,5	85,5
ed009-pg019a-mdav-generic-5-frequency.csv	1,046	1,259	5,037	2,833	5,913	1,106	34	68	7,158	5,913	45,333	19,429	3,317	3,487	3,238	7,556	68	68	68
ed009-pg019a-mdav-generic-6-frequency.csv	1,018	1,215	5,65	3,139	6,278	1,153	28,25	56,5	5,947	5,65	37,667	22,6	3,229	2,974	3,229	7,533	56,5	56,5	56,5
ed009-pg019a-mdav-generic-7-frequency.csv	1,021	1,228	5,389	2,771	6,062	1,212	24,25	48,5	5,389	4,409	32,333	19,4	3,593	3,593	3,031	8,818	48,5	48,5	48,5
ed009-pg019a-mdav-generic-8-frequency.csv	1,012	1,181	5,667	2,931	5,312	1,214	21,25	42,5	4,722	4,048	42,5	17	2,833	3,036	3,4	10,625	42,5	42,5	42,5
ed009-pg019a-mdav-generic-9-frequency.csv	1	1,271	5	3,409	6,25	1,271	18,75	37,5	4,167	4,412	37,5	18,75	3,125	3,2261	2,8885	15	37,5	37,5	37,5
ed009-pg019a-mdav-generic-10-frequency.csv	1,047	1,175	4,786	3,35	5,583	1,34	16,75	33,5	4,786	3,941	33,5	16,75	2,792	2,68	2,913	9,571	67	67	67

Taula 11: Freqüències d'ed009-pg019a

3.3.3 Cens del 1910

A continuació, els resultats per l'arxiu *ed007-pg001a.txt* amb dades del cens del 1910. Igual que abans, tot i que no és 100% correcte, s'ha considerat l'edat com a continu, per aplicar tots els casos en aquest arxiu:

file	LASTNAME	FIRSTNAME	RELATION	SEX	RACE	AGELAST	#YRSMAR	#CHILD	#CHILDLIVING	BIRTHPLACE	FATHERBIRTHPLACE	MOTHERBIRTHPLACE	YEARIMMI	NATURALIZED	SPEAKENG?	TRADEORPROFESSION	GENERALINDUSTRY	EMPLACT	OUTOFWORK	#WKSUNEMPL	CANREAD	CANWRITE	ATTDSCH	OWNRENT	FREEMORT	FARMHOME	#FARMSCH	SURVWAR	BLIND	DEAFDUMB
ed007-pg001a-mean.csv	Anderson	Alvord VanP.	Head	M	W	38	2,194	0	0	New York	New York	Ohio	0	.	English	None	.	.	.	.	Yes	Yes	No	.	.	.	.	.	.	.
ed007-pg001a	Anderson	Alvord	Head	M	W	4	2,5	0	0	Iowa	Pennsylvania	Kansas	0	.	English	None	.	.	.	.	Yes	Yes	No	.	.	.	.	.	.	.

-md av-generic-3-mean.csv		an P., Jr																											
ed007-pg001a-md av-generic-4-mean.csv	Anderson	Alvord VanP., Jr	Wife	M	W	37	2,341	0	0	Montana	New York	Kansas	0	.	English	None	.	.	.	.	Yes	Yes	No	.	.	.	.	.	.	.
ed007-pg001a-md av-generic-5-mean.csv	Scott	Alvord VanP., Jr	Head	M	W	4	2,267	0	0	Montana	Pennsylvania	Cebu Isld P.I.	0	.	English	None	.	.	.	.	Yes	Yes	No	.	.	.	.	.	.	.
ed007-pg001a-md av-generic-6-mean.csv	Anderson	Alvord VanP., Jr	Wife	M	W	4	2,243	0	0	Montana	New York	Kansas	0	.	English	None	.	.	.	.	Yes	Yes	No	.	.	.	.	.	.	.
ed007-pg001a-md av-generic-7-mean.csv	Anderson	Alvord VanP., Jr	Son	M	W	4	1,989	0	0	Montana	Pennsylvania	Kansas	0	.	English	None	.	.	.	.	Yes	Yes	No	.	.	.	.	.	.	.
ed007-pg001a-md av-generic-8-mean.csv	Anderson	Alvord VanP., Jr	Son	M	W	4	1,605	0	0	Montana	New York	Kansas	0	.	.	None	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ed007-pg001a-md av-generic-9-mean.csv	Anderson	Alvord VanP., Jr	Son	F	W	4	1,818	0	0	Montana	New York	Kansas	0	.	.	None	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ed007-pg001a-md av-generic-10-mean.csv	Anderson	Alvord VanP., Jr	Son	F	W	4	2,03	0	0	Montana	New York	Kansas	0	.	English	None	.	.	.	.	Yes	Yes	.	.	.	.	.	.	.	.

Taula 12: Mitjanes d'ed007-pg001a

file	#YRS MAR
ed007-pg001a-variance.csv	11,428
ed007-pg001a-mdav-generic-3-variance.csv	9,188
ed007-pg001a-mdav-generic-4-variance.csv	14,992
ed007-pg001a-mdav-generic-5-variance.csv	5,199
ed007-pg001a-mdav-generic-6-variance.csv	3,569
ed007-pg001a-mdav-generic-7-variance.csv	2,65
ed007-pg001a-mdav-generic-8-variance.csv	4,275
ed007-pg001a-mdav-generic-9-variance.csv	4,914
ed007-pg001a-mdav-generic-10-variance.csv	5,76

Taula 13: Variàncies d'ed007-pg001a

file	#YRS MAR
ed007-pg001a-quantile50.csv	1
ed007-pg001a-mdav-generic-3-quantile50.csv	1,5
ed007-pg001a-mdav-generic-4-quantile50.csv	0,524
ed007-pg001a-mdav-generic-5-quantile50.csv	1,833
ed007-pg001a-mdav-generic-6-quantile50.csv	2,286
ed007-pg001a-mdav-generic-7-quantile50.csv	2,5
ed007-pg001a-mdav-generic-8-quantile50.csv	1,605
ed007-pg001a-mdav-generic-9-quantile50.csv	1,818
ed007-pg001a-mdav-generic-10-quantile50.csv	2,03

Taula 14: Quantil 50 d'ed007-pg001a

file	#YRS MAR
ed007-pg001a-quantile75.csv	3
ed007-pg001a-mdav-generic-3-quantile75.csv	3,5
ed007-pg001a-mdav-generic-4-quantile75.csv	2,179
ed007-pg001a-mdav-generic-5-quantile75.csv	4,25
ed007-pg001a-mdav-generic-6-quantile75.csv	3,35
ed007-pg001a-mdav-generic-7-quantile75.csv	2,9
ed007-pg001a-mdav-generic-8-quantile75.csv	2,336
ed007-pg001a-mdav-generic-9-quantile75.csv	2,601
ed007-pg001a-mdav-generic-10-quantile75.csv	2,878

Taula 15: Quantil 75 d'ed007-pg001a

file	LASTNAME	FIRSTNAME	RELATION	SEX	RACE	AGELAST	#CHILD	#CHILDLIVING	BIRTHPLACE	FATHERBIRTHPLACE	MOTHERBIRTHPLACE	YEARIMMI	NATURALIZED	SPEAKENG?	TRADEORPROFESSION	GENERALINDUSTRY	EMPLACT	OUTOFWORK	#WKSUNEMPL	CANREAD	CANWRITE	ATTDSCH	OWNRENT	FAREMORT	FARMHOME	#FARMSCH	SURVWAR	BLIND	DEAFDUMB
ed007-pg001a-	1,938	1	5,1	10,3	5,1	1,24	10,3	10,3	1,476	1,632	1,632	31	31	10,3	5,167	10,333	15,5	31	31	15,5	15,5	10,3	15,5	31	15,5	31	31	31	31

frequency.csv			67	33	67		33	33						33								33							
ed007-pg001a-mdav-generic-3-frequency.csv	1,286	1	3	3	2,25	1	4,5	4,5	1,125	1,286	1	9	9	3	3	4,5	4,5	9	9	4,5	4,5	4,5	4,5	9	4,5	9	9	9	9
ed007-pg001a-mdav-generic-4-frequency.csv	1,2	1	2	3	3	1,2	2	2	1	1,2	1	6	6	3	2	3	3	6	6	3	3	3	3	6	3	6	6	6	6
ed007-pg001a-mdav-generic-5-frequency.csv	1,25	1	1,667	2,5	2,5	1	2,5	2,5	1	1,25	1	5	5	2,5	1,667	2,5	2,5	5	5	2,5	2,5	2,5	2,5	5	2,5	5	5	5	5
ed007-pg001a-mdav-generic-6-frequency.csv	1,333	1	1,333	2	4	1	2	2	1	1	1	4	4	2	2	4	2	4	4	2	2	2	2	4	2	4	4	4	4
ed007-pg001a-mdav-generic-7-frequency.csv	3	1	1	1,5	3	1	1,5	1,5	1	1,5	1	3	3	1,5	1,5	3	1,5	3	3	1,5	1,5	1,5	1,5	3	1,5	3	3	3	3
ed007-pg001a-mdav-generic-8-frequency.csv	1	1	1	1	2	1	2	2	1	1	1	2	2	1	2	2	2	2	2	1	1	1	2	2	2	2	2	2	2
ed007-pg001a-mdav-generic-9-frequency.csv	2	1	1	2	2	1	2	2	1	1	1	2	2	1	2	2	2	2	2	1	1	1	2	2	2	2	2	2	2
ed007-pg001a-mdav-generic-10-frequency.csv	2	1	1	2	2	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	2

Taula 16: Freqüències d'ed007-pg001a

4. Estat de l'art

Dintre de les mesures utilitzades per quantificar la privacitat, la més comú és la de k-anonimat, tot i que aquesta és vulnerable a una sèrie d'atacs. Tota aquesta secció està basada en [7]

Els tipus d'atac al que es pot veure sotmès un conjunt de dades són els següents:

- Atac de vinculació (*linking attacks*)
  Contra aquest tipus d'atac és contra el que se suposa que protegeixen els algorismes que busquen la k-anonimat, perquè consisteix en vincular la informació publicada amb altres bases de dades per tal d'arribar a dades personals. La k-anonimat, en fer que els atributs clau estiguin compartits entre diversos registres, fa això més difícil, tot i que, això no vol dir que no sigui possible fer aquest tipus d'atacs en alguns casos, perquè mai podem estar segurs de quines bases de dades té accessibles un atacant.

- Atac d'homogeneïtat (*homogeneity attack*)
  Aquest atac aprofita el cas que tots els valors d'un atribut sensible d'un conjunt de registres k són iguals per tal de poder predir el valor sensible d'aquest conjunt de registres [10] Seria el cas de deduir que una persona està en una base de dades de pacients de SIDA, per exemple, en aquest cas, no ens fa falta saber quin registre de la base de dades és per tenir una informació privada.

- Atac de coneixement d'antecedents (*background-knowledge attack*)
  En aquest atac, aprofitem informació externa a les nostres dades per deduir informació personal, com per exemple, utilitzar l'origen d'un nom per determinar quines malalties són més probables d'unes dades processades per evitar filtracions.

- Atac d'obliqüitat (*skewness attack*)
  Un dels atacs més importants, que consisteix en aprofitar que la distribució de les dades originals i tractades pot no ser la mateixa. Això permet establir probabilitats de que un determinat individu compleixi una determinada característica sent aquesta una informació que hauria de ser privada.

Per tal d'evitar aquest tipus d'atacs, s'han plantejat alguns conceptes alternatius a la k-anonimat:

- l-diversitat (*l-diversity*)
  Dintre d'aquesta família de criteris, que inclou la *distinct l-diversity*, la *entropy l-diversity* i la *recursive (c-l)-diversity*, es defineix una ampliació de la k-anonimat on s'ha de garantir que el grup de registres *k* ha de contenir al menys *l* valors "ben representats" per cada atribut confidencial. De la definició d'aquest "ben representat" dependran les diferents variants.

Aquest criteri, però, continua sent vulnerable als atacs d'obliqüitat, tot i que no tan com la k-anonimat, així com als atacs de similitud, en el sentit de que, encara que es compleixi la l-diversitat, els diferents valors poden ser semànticament similars i permetre obtenir informació no desitjada.

- t-proximitat (*t-closeness*)
  El criteri de t-proximitat diu que, per cada grup que comparteixi un registre de comú d'atributs clau modificats, es compleix una distància màxima de t, per una determinada distància, entre la distribució després de l'aplicació de les pertorbacions i abans.
  El principal problema d'aquest criteri és que no es coneix cap procediment computacional per obtenir-lo

Altres criteris també proposats són la δ-divulgació (*δ-disclosure*) i la ε-privacitat diferencial (*ε-differential privacy*)

5. Bibliografia

[1] L. Willenborg and T. DeWaal, Elements of statistical disclosure control. New York: Springer-Verlag, 2001.

[2] L. Sweeney, "k-Anonymity: A model for protecting privacy," Int. J. Uncertain., Fuzz., Knowl.-Based Syst., vol. 10, no.5, pp. 557-570, 2002.

[3] J. Domingo-Ferrer and J. M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control," IEEE Trans. Knowl. Data Eng., vol. 14, no. 1, pp. 189-201, 2002.

[4] J. Domingo-Ferrer, V. Torra, "Ordinal, continuous and heterogenerous k-anonymity through microaggregation," Data Mining and Knowledge Discovery vol. 11, no. 2, pp. 195-212, 2005.

[5] A. Machanavajjhala, J. Gehrke, D. Kiefer, and M. Venkitasubramanian, "l-Diversity: Privacy beyond k-anonymity," in Proc. IEEE Int. Conf. Data Eng. (ICDE), Atlanta, GA, 2006.

[6] David Rebollo-Monedero, Jordi Forné, Esteve Pallarès, Javier Parra-Arnau, “A Modification of the Lloyd Algorithm for k-Anonymous Quantization,” Elsevier Inform. Sci., vol. 222, Feb. 2013, pp. 185-202. DOI: 10.1016/j.ins.20

Afegits per l’estat de l’art:

[7] Parra-Arnau, Javier & Rebollo-Monedero, David & Forné, Jordi. (2012). Privacy Protection of User Profiles in Personalized Information Systems.

[8] Rajendran, Keerthana & Jayabalan, Manoj & Rana, Muhammad Ehsan. (2017). A Study on k-anonymity, l-diversity, and t-closeness Techniques focusing Medical Data. 17. <https://www.researchgate.net/publication/322330948_A_Study_on_k-anonymity_l-diversity_and_t-closeness_Techniques_focusing_Medical_Data> [accedit 6 octubre 2019]

[9] Wikipedia contributors, 'T-closeness', *Wikipedia, The Free Encyclopedia,* 11 juliol 2019, 10:09 UTC, <https://en.wikipedia.org/w/index.php?title=T-closeness&oldid=905775506> [accedit 6 octubre 2019]

[10] Wikipedia contributors, 'L-diversity', *Wikipedia, The Free Encyclopedia,* 9 agost 2019, 10:16 UTC, <https://en.wikipedia.org/wiki/L-diversity> [accedit 29 desembre 2019]

Protecció de la privadesa de microdades mitjançant MDAV

Javier Vilalta Bautista (*Author*)
Treball de final de grau
UOC
Barcelona
xvilaltb@uoc.edu

***Abstract*— This work presents an implementation of the MDAV algorithm in order to protect a dataset so that its general characteristics are not lost, and can protect the privacy of individuals, and at the same time provide enough information to be relevant in statistical studies.**

The methodology used has been the implementation of the algorithm and its evaluation with different datasets. Finally, the drawbacks and limitations of k-anonymity are evaluated and compared with other criteria that have better behavior against certain types of attacks.

Keywords*— *MDAV, privacitat, microdades

I. INTRODUCCIÓ

A. Introducció/Prefaci

A mesura que les millores en tecnologia permeten mantenir més volum i més varietat de dades específiques sobre individus, augmenta la preocupació per la privacitat. La seva publicació, que pot ser molt necessària per a la realització d'estudis científics, s'ha de fer en condicions que permetin garantir la privacitat dels individus implicats, però sense perdre validesa pel seu estudi. En aquest conflicte, s'han desenvolupat una sèrie de tècniques per aconseguir aquest difícil equilibri.

A més d'aquestes tècniques, necessitem algun tipus de mesura per tal de quantificar-les de manera objectiva: un d'aquests indicadors és el de k-anonimat, que es defineix de la següent forma: un conjunt de dades proporciona k-anonimat si la informació de cada persona que hi conté no es pot distingir de, al menys, k-1 altres individus que també estiguin a les dades.

Amb aquests precedents, en aquest treball de final de grau treballarem el concepte de k-anonimat, des del punt de vista teòric, així com pràctic, mitjançant la implementació d'un algorisme que permet garantir el k-anonimat sobre un conjunt de dades.

Tot i que també es volia treballar el concepte de l-diversitat, per tal de veure les limitacions del concepte de k-anonimat i les diferències en implementació, al final no ha estat possible per restriccions de temps.

B. Descripció/Definició

En aquest treball ens preocupem de la necessitat de protegir la privadesa d'individus i institucions, en el context de la publicació d'informació que pot ser rellevant per a investigació. Podem imaginar el cas d'una institució, com ara un hospital, que té estadístiques d'una determinada malaltia i que pot pensar en publicar aquestes dades per tal d'ajudar a futurs investigadors a nous descobriments que ajudin en el tractament. Com ho fem, però, de manera que les dades siguin útils per la investigació, però a l'hora protegim la privacitat de cada malalt? Com mesurem aquesta privacitat?

D'entre els molts mecanismes que existeixen i que es comenten més endavant, un d'ells és l'algorisme MDAV, que garanteix el compliment de la característica de k-anonimat en un conjunt de dades.

El resultat d'aquest treball serà la implementació d'un algorisme de MDAV funcional, junt amb algunes proves amb alguns conjunts de dades per tal de validar el seu funcionament.

C. Objectius generals

Els objectius, en base a l'enunciat del projecte, seran els següents:

- Estudiar la privadesa de dades des d'un punt de vista algorísmic

  Aquest punt servirà de marc general de les explicacions de la resta de punts i ens basarem en tota la bibliografia del TFG.

- Analitzar avantatges i limitacions del requisit de k-anonimat

  En aquest punt, tractarem de veure els avantatges i limitacions del k-anonimat, tal com es descriu a [2]

- Implementar l'algorisme MDAV

  Es tracta d'implementar l'algorisme MDAV-genèric, tal com es defineix a [4]

- Estudiar l'equilibri entre privadesa i distorsió de dades per a diferents microdades

Una vegada tinguem els algorismes implementats, haurem d'estudiar com es comporten, comparant la privades proporcionada amb la distorsió causada a les dades, per la qual cosa, haurem d'utilitzar els diferents criteris definits a la bibliografia.

- Millorar l'algorisme de MDAV per a que satisfaci el requisit de l-diversitat

  El requisit de l-diversitat se'ns explica a [5], que millora la seguretat del requisit de k-anonimat. En aquest mateix document, se'ns prova que, en general, és possible substituir un criteri per l'altra en algorismes de k-anonimat per la seva semblança i, per tant, podrem canviar el criteri a l'algorisme MDAV genèric anterior.

  Finalment, aquest últim objectiu no s'ha pogut assolir per falta de temps.

II. Metodologia i procés de treball

La metodologia utilitzada per tal de realitzar la implementació ha estat la de fer el desenvolupament de manera àgil, començant per objectius fàcilment assolibles, implementant parcialment l'algorisme pel diferents tipus de dades i, a partir d'aquí, anar progressant fins a tenir tota la casuística implementada.

A. Planificació

El projecte s'ha dividit en tasques, de manera que es puguin utilitzar els recursos i planificar de manera més eficient. A la taula següent es pot veure la descomposició realitzada:

<Taules i Gantt eliminats>

Per restriccions de temps, no ha estat possible acabar la implementació de l'algorisme amb l'ampliació de l-diversitat. El problema principal ha estat que les correccions de l'algorisme genèric i les proves realitzades han ocupat més temps del previst.

III. Estructura de la resta del document

A la resta del document trobarem les següents seccions:

- Base teòrica del treball

  En aquest apartat farem una breu visió sobre el problema de la publicació de microdades i veurem alguns plantejaments formals que ens ajudaran a entendre el treball.

- Implementació realitzada de l'algorisme MDAV

  En aquesta part, farem una explicació de la implementació realitzada, així com mostrarem una sèrie de dades resultants.

- Estat de l'art

  En aquest apartat, parlarem de les limitacions del criteri de k-anonimat, així com de les vulnerabilitats i de quins plantejaments alternatius s'estan fent per evitar-los.

IV. Plantejament teòric

A. Introducció

En aquest apartat revisarem els conceptes teòrics que hi ha al darrera del problema, així com un petit resum de la problemàtica de la revelació de microdades i el compromís entre la privacitat i la utilitat de les mateixes.

Aquest problema no és nou, històricament, els organismes oficials han hagut de fer pública informació estadística rellevant per a la investigació, però les solucions existents actualment no són adequades per l'entorn actual, amb conjunts més grans de dades i amb les noves necessitats d'anàlisi que alguns actors demanen: mineria de dades, anàlisi de costos, detecció del frau, etc., que requereixen dades més específiques a nivell de persona.

Algunes de les eines utilitzades fins ara són:

- Bases de dades estadístiques: aquestes bases de dades estan dissenyades per fer consultes a nivell agregat, protegint els elements individuals.
- Bases de dades multinivell: en aquest tipus, les dades estan assignades a nivells de seguretat, de manera que el sistema torna només informació dels nivells als que tenim permís, intentant sempre que no es puguin inferir dades d'un nivell a partir de les dades d'un altre. El problema d'aquest enfoc és que no es pot garantir al cent per cent que això sempre sigui així

També és important considerar que l'estudi de la privacitat no és el mateix que la seguretat, donat que aquesta última es preocupa de l'accés a les dades, mentre que la privacitat el que vol garantir és que, tot i tenir accés a les dades, no puguem arribar mai a conclusions respecte a persones individuals.

Una pràctica habitual és eliminar els identificadors explícits, com ara nom, telèfon o adreça. Però, en molts casos, la resta de les dades poden ser combinades per identificar, amb un grau alt de certesa, els individus, amb l'ajuda d'altres dades públicament disponibles. Per exemples, és possible utilitzar les dades gènere, data de naixement, ZIP i etnicitat d'una base de dades mèdica i combinar-la amb el cens, per obtenir amb un grau alt de certesa l'individu.

B. Marc formal

A continuació, indicarem algunes definicions importants a la resta del document:

Dades: informació específica d'una persona conceptualment organitzada en forma de taula de files (o registres o tuples) que no són necessàriament úniques i columnes (o camps o atributs) que sí són úniques (no està dues vegades el nom, per exemple)

Inferència: arribar a un nou fet en base a altra informació

Revelació (disclosure): el coneixement d'informació explícita o inferida no previst respecte a una persona

Control de revelacions (disclosure control): qualsevol intent de limitar o identificar revelacions respecte a un conjunt de dades

Quasi-identificador: conjunt d'atributs que, junts, poden identificar unívocament un individu amb l'ajuda d'una altra taula (com ara dades públiques) La identificació d'aquests quasi-identificadors no sempre és evident.

k-anonimat: donada una taula i un quasi-identificador associada amb ella, la taula es diu satisfà k-anonimat si i només si cada seqüència de valors del quasi-identificador apareix al menys k vegades.

C. Classificació de les microdades

Les microdades es poden classificar en els següents tipus:

- Continues

  Quan l'atribut és numèric i es poden aplicar operacions aritmètiques.

- Categòriques

  Una atribut és categòric si pren valor en un conjunt finit de valors possibles i no es poden aplicar operacions aritmètiques.

  A la seva vegada, es poden dividir en ordinals i nominals:

- Ordinal

  Quan a l'atribut té sentit aplicar-li els operadors d'ordre, com ara <=, màxim, mínim, etc.

- Nominal

  Quan no té sentit aplicar un ordre a les categories.

D. Definició de l'objectiu

Amb tot el definit anteriorment, el nostre objectiu consisteix en, donat un conjunt de microdades V, alliberar un conjunt de microdades alternatiu V' tal que:

El risc de revelació d'informació és baix

També anomenat d'identificació o reidentificació dels enquestats

La pèrdua d'informació és baixa.

Més formalment, podem definir una puntuació:

$$\boldsymbol{Score}(\boldsymbol{V},\boldsymbol{V'}) = \frac{\boldsymbol{IL}(\boldsymbol{V},\boldsymbol{V'}) + \boldsymbol{DR}(\boldsymbol{V},\boldsymbol{V'})}{\mathbf{2}}$$

On:

IL: Mesura de la pèrdua de la informació

DR: Mesura del risc de revelació

Per a la consecució d'aquest objectiu, els mètodes de protecció de microdades poden aplicar dues tècniques per generar un conjunt protegit:

- Emmascarar les dades

  D'aquests, n'hi ha dos tipus:

- Amb pertorbació: amb aquests mètodes, poden aparèixer combinacions noves i desaparèixer d'altres però sempre preservant les propietats estadístiques. Alguns exemples són la microagregació i l'afegit de soroll.

- Sense pertorbació: Aquest mètodes no distorsionen les dades originals. Alguns exemples són la generalització i la supressió.

- Generar dades sintètiques, però preservant algunes propietats estadístiques

Existeixen moltes formes possibles de concretar aquest procés. A la implementació, es veurà més en detall el cas concret de l'algorisme MDAV com a eina per aconseguir un conjunt de dades amb k-anonimat.

V. IMPLEMENTACIÓ

A. Explicació de la implementació

L'algorisme MDAV s'ha implementat en llenguatge Python. L'estructura general del programa és la següent:

- Un mòdul mdav

  Programa principal que fa servir el mòdul

La forma de treballar amb el programa és mitjançant dos arxius:

- Arxiu en format CSV amb les dades
- Arxiu en format CSV amb les metadades

Per convenció, s'entén que l'arxiu amb les metadades té el mateix nom que l'arxiu de dades

Els diferents tipus de dades que s'esperen són:

- Contínues
- Categòriques (ordinals i nominals)

Per qüestions d'optimització es va afegir un tipus fictici per ignorar determinades columnes, però finalment s'ha decidit no fer-lo servir.

A la seva vegada, el mòdul mdav està composat dels següents

- common.py

  Aquest mòdul inclou la definició bàsica dels tipus de dades

- helper.py

  Aquest mòdul inclou funcions auxiliars per carregar arxius en format CSV i el seu arxiu de metadades vinculat.

- mdav_generic.py

  Aquest mòdul és el que fa el càlcul MDAV pròpiament i es veu una mica més endavant.

- metric.py

Aquest mòdul calcula una sèrie de paràmetres, com ara mitjanes i variàncies per tal de comparar els diferents conjunts de dades generats.

L'algorisme principal es pot veure a continuació, amb la documentació amb l'algorisme explicat a la documentació:

```
def calc_mdav_generic(
    dataset: Dataset,
    k: int) -> list:
    """
    Function calc_mdav_generic
    """
    private_data = []
    while len(dataset.records) >= 3 * k:
        logger.debug(f'1. Iterating with {len(dataset.records)} records')
        columns, means, stdevs = compute_auxiliary_tables(dataset)
        # (1a) Compute the average record x̃ of all records in R. The average
        #     record is computed attribute-wise.
        x_avg = compute_average_record(dataset)
        logger.debug(f'1a. Average record is {x_avg}')
        # (1b) Consider the most distant record xr to the average record x̃
        #     using an appropriate distance
        xr = find_most_distant_record(dataset, x_avg)
        logger.debug(f'1b. Most distant record to average (xr) is {xr}')
        # (1c) Find the most distant record xs from the record xr considered in
        #     the previous step
        xs = find_most_distant_record(dataset, xr)
        logger.debug(f'1c. Most distant record to average record (xs) is {xs}')
        # (1d) Form two clusters around xr and xs , respectively. One cluster
        #     contains xr and the k − 1 records closest to xr. The other
        #     cluster contains xs and the k − 1 records closest to xs.
        # (1e) Take as a new dataset R the previous dataset R minus the
        #     clusters formed around xr and xs in the last instance of
        #     Step 1d.
        cluster_xr, dataset = extract_cluster(dataset, xr, k - 1)
        columns, means, stdevs = compute_auxiliary_tables(cluster_xr)
        cluster_xr_avg = compute_average_record(cluster_xr)
        logger.debug(f'1d. Cluster xr average record is {xr}')
        private_data.append(cluster_xr_avg)
        cluster_xs, dataset = extract_cluster(dataset, xs, k - 1)
        columns, means, stdevs = compute_auxiliary_tables(cluster_xs)
        cluster_xs_avg = compute_average_record(cluster_xs)
        logger.debug(f'1d. Cluster xs average record is {xs}')
        private_data.append(cluster_xs_avg)
        logger.debug(f'1e. New dataset size is {len(private_data)}')
    if len(dataset.records) >= 2 * k:
        logger.debug(
            f'2. Entering step 2 with {len(dataset.records)} records (>=2k)'
        )
        columns, means, stdevs = compute_auxiliary_tables(dataset)
        # (2a) Compute the average record x̃ of the remaining records in R
        x_avg = compute_average_record(dataset)
        logger.debug(f'2a. Average record is {x_avg}')
        # (2b) Find the most distant record xr from x̃
        xr = find_most_distant_record(dataset, x_avg)
        logger.debug(f'2b. Most distant record to average (xr) is {xr}')
        # (2c) Form a cluster containing xr and the k − 1 records closest to xr
        cluster_xr, dataset = extract_cluster(dataset, xr, k - 1)
        columns, means, stdevs = compute_auxiliary_tables(cluster_xr)
        cluster_xr_avg = compute_average_record(cluster_xr)
        logger.debug(f'2c. Cluster xr average record is {xr}')
        private_data.append(cluster_xr_avg)
        # (2d) form another cluster containing the rest of records
```

```
        else:
            logger.debug(
                f'2. Entering step 2 with {len(dataset.records)} records (<2k)'
            )
            # else (less than 2k records in R) form a new cluster with the
            # remaining records
        columns, means, stdevs = compute_auxiliary_tables(dataset)
        cluster_rest = compute_average_record(dataset)
        logger.debug(f'2/2d. Average record of rest of dataset is {cluster_rest}')
        private_data.append(cluster_rest)
        logger.debug(f'2. New dataset size is {len(private_data)}')
        return private_data
```

B. Dades de proves

Les dades utilitzades per fer les proves i els estudis són les següents:

- Sample 1

Conjunt de dades fictícies, amb la intenció de recollir una casuística dels diferents tipus de dades. L'estructura és la següent:

Surname	name	year	balance

- Dades del cens dels EEUU

Les dades que es faran servir seran històriques, del cens dels EEUU en concret

Les dades s'han obtingut mitjançant aquestes adreces:

https://www.census.gov/prod/2000pubs/cff-2.pdf

https://1940census.archives.gov/

https://www.census.gov/history/www/genealogy/decennial_censuses_records/census_records_2.html

http://www.us-census.org

Els exemples que s'estan utilitzant en concret són:

1900 U.S. Military and Naval Federal Census

http://us-census.org/pub/usgenweb/census/_mil/1900/t623-1842/ed009-pg019a.txt

Census Year 1900

Microfilm Roll #T623-1842

Name of Military or Naval Station, or Vessel: U.S.S. Buffalo

State --

Country --

Seaport Gibraltar

Arm of Service U. S. Navy

A l'arxiu auxiliar corresponent es pot veure l'estructura i els tipus utilitzats

1910 U.S. Military and Naval Federal Census

http://us-census.org/pub/usgenweb/census/_mil/1910/t624-1784/ed007-pg001a.txt

Census Year 1910 CENSUS-DAY: April 15, 1910

Microfilm Roll #T624-1784

State --

County (Province) --

Township --

Incorporated-Place --

Institution Post of Camp Avery, Corregidor, Isle P.I.

A l'arxiu auxiliar corresponent es pot veure l'estructura i els tipus utilitzats

Tots dos arxius s'han convertir a CSV manualment, per donar una estructura més clara.

En tots els casos, s'ha generat tots els conjunts de corresponents als valors de k del tres al deu.

C. Resultats

A continuació, s'indiquen els resultats obtinguts per a diferents mesures obtingudes. En tots els casos, la columna file indica sobre quin arxiu s'ha fet el càlcul de la mesura i, dintre del nom, s'indica el valor de la k que s'ha fet servir per aplicar l'MDAV: per exemple, l'arxiu sample1-mdav-generic-4-mean.csv indica que s'ha aplicat mdav per 4-anonimat a l'arxiu sample1.csv. Quan el nom no inclou mdav-generic, com ara sample1-mean.csv, vol dir que el càlcul de mitjanes, en aquest cas, s'ha fet sobre l'arxiu original.

Algunes de les taules tenen molta informació i no es poden veure clarament. Adjunt al projecte s'inclouen els arxius Excels originals per tal de poder revisar-los millor.

El que s'ha de considerar en aquest cas és el següent:

- Per construcció, MDAV preserva mitjanes i variàncies
- Els valors individuals, els quantils, les covariàncies, les correlacions i les freqüències no es preserven.
- Quan més puja la k, més distorsió es produeix en les estadístiques no preservades.

És per aquest motiu que escollim aquestes mesures, amb l'objectiu de veure tant el comportament de mesures que es preserven com el comportament de mesures que no es preserven:

- Mitjana (Mean)

En aquest cas, es calcula la mitjana per a cada valor. Aquest càlcul serveix per a tots els tipus de dades.

- Variància (Variance)

Aquest valor només es calcular pels atributs continus.

- Quantil 50% (Quantile 50)

Aquest valor només es calcular pels atributs continus.

- Quantil 75% (Quantile 75)

Aquest valor només es calcular pels atributs continus.

- Freqüències (Frequencies)

Aquest valor és una mitjana de les freqüències i serveix per tenir una idea de com varien els diferents tipus de dades no continus (ordinals i nominals)

- Dades fictícies

A continuació, els resultats per l'arxiu sample1 amb dades fictícies.

<Taules eliminades>

Els resultats quadren amb el que estava previst, tot i que hauríem de fer una anàlisi estadística més exhaustiva per validar que les variacions són les esperades. En tot cas, sí que és cert que la poca quantitat de dades provoca algunes distorsions en alguns valors.

- Cens del 1900

A continuació, els resultats per l'arxiu ed009-pg019a.txt amb dades del cens de 1900. Tot i que no és 100% correcte, s'ha considerat l'any com a continu, per aplicar tots els casos en aquest arxiu:

<Taules eliminades>

- Cens del 1910

A continuació, els resultats per l'arxiu ed007-pg001a.txt amb dades del cens del 1910. Igual que abans, tot i que no és 100% correcte, s'ha considerat l'edat com a continu, per aplicar tots els casos en aquest arxiu:

<Taules eliminades>

VI. Estat de l'art

Dintre de les mesures utilitzades per quantificar la privacitat, la més comú és la de k-anonimat, tot i que aquesta és vulnerable a una sèrie d'atacs. Tota aquesta secció està basada en [7]

Els tipus d'atac al que es pot veure sotmès un conjunt de dades són els següents:

- Atac de vinculació (linking attacks)

Contra aquest tipus d'atac és contra el que se suposa que protegeixen els algorismes que busquen la k-anonimat, perquè consisteix en vincular la informació publicada amb altres bases de dades per tal d'arribar a dades personals. La k-anonimat, en fer que els atributs clau estiguin compartits entre diversos registres, fa això més difícil, tot i que, això no vol dir que no sigui possible fer aquest tipus d'atacs en alguns casos, perquè mai podem estar segurs de quines bases de dades té accessibles un atacant.

- Atac d'homogeneïtat (homogeneity attack)

Aquest atac aprofita el cas que tots els valors d'un atribut sensible d'un conjunt de registres k són iguals per tal de poder predir el valor sensible d'aquest conjunt de registres [10] Seria el cas de deduir que una persona està en una base de dades de pacients de SIDA, per exemple, en aquest cas, no ens fa falta saber quin registre de la base de dades és per tenir una informació privada.

- Atac de coneixement d'antecedents (background-knowledge attack)

En aquest atac, aprofitem informació externa a les nostres dades per deduir informació personal, com per exemple, utilitzar l'origen d'un nom per determinar quines malalties són més probables d'unes dades processades per evitar filtracions.

- Atac d'obliqüitat (skewness attack)

Un dels atacs més importants, que consisteix en aprofitar que la distribució de les dades originals i tractades pot no ser la mateixa. Això permet establir probabilitats de que un determinat individu compleixi una determinada característica sent aquesta una informació que hauria de ser privada.

Per tal d'evitar aquest tipus d'atacs, s'han plantejat alguns conceptes alternatius a la k-anonimat:

- l-diversitat (l-diversity)

Dintre d'aquesta família de criteris, que inclou la distinct l-diversity, la entropy l-diversity i la recursive (c-l)-diversity, es defineix una ampliació de la k-anonimat on s'ha de garantir que el grup de registres k ha de contenir al menys l valors "ben representats" per cada atribut confidencial. De la definició d'aquest "ben representat" dependran les diferents variants.

Aquest criteri, però, continua sent vulnerable als atacs d'obliqüitat, tot i que no tan com la k-anonimat, així com als atacs de similitud, en el sentit de que, encara que es compleixi la l-diversitat, els diferents valors poden ser semànticament similars i permetre obtenir informació no desitjada.

- t-proximitat (t-closeness)

El criteri de t-proximitat diu que, per cada grup que comparteixi un registre de comú d'atributs clau modificats, es compleix una distància màxima de t, per una determinada distància, entre la distribució després de l'aplicació de les pertorbacions i abans.

El principal problema d'aquest criteri és que no es coneix cap procediment computacional per obtenir-lo

Altres criteris també proposats són la δ-divulgació (δ-disclosure) i la ε-privacitat diferencial (ε-differential privacy)

VII. Bibliografia

[1] L. Willenborg and T. DeWaal, Elements of statistical disclosure control. New York: Springer-Verlag, 2001.

[2] L. Sweeney, "k-Anonymity: A model for protecting privacy," Int. J. Uncertain., Fuzz., Knowl.-Based Syst., vol. 10, no.5, pp. 557-570, 2002.

[3] J. Domingo-Ferrer and J. M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control," IEEE Trans. Knowl. Data Eng., vol. 14, no. 1, pp. 189-201, 2002.

[4] J. Domingo-Ferrer, V. Torra, "Ordinal, continuous and heterogenerous k-anonymity through microaggregation," Data Mining and Knowledge Discovery vol. 11, no. 2, pp. 195-212, 2005.

[5] A. Machanavajjhala, J. Gehrke, D. Kiefer, and M. Venkitasubramanian, "l-Diversity: Privacy beyond k-anonymity," in Proc. IEEE Int. Conf. Data Eng. (ICDE), Atlanta, GA, 2006.

[6] David Rebollo-Monedero, Jordi Forné, Esteve Pallarès, Javier Parra-Arnau, "A Modification of the Lloyd Algorithm for k-Anonymous Quantization," Elsevier Inform. Sci., vol. 222, Feb. 2013, pp. 185-202. DOI: 10.1016/j.ins.20

[7] Parra-Arnau, Javier & Rebollo-Monedero, David & Forné, Jordi. (2012). Privacy Protection of User Profiles in Personalized Information Systems.

[8] Rajendran, Keerthana & Jayabalan, Manoj & Rana, Muhammad Ehsan. (2017). A Study on k-anonymity, l-diversity, and t-closeness Techniques focusing Medical Data. 17. <https://www.researchgate.net/publication/322330948_A_Study_on_k-anonymity_l-diversity_and_t-closeness_Techniques_focusing_Medical_Data> [accedit 6 octubre 2019]

[9] Wikipedia contributors, 'T-closeness', Wikipedia, The Free Encyclopedia, 11 juliol 2019, 10:09 UTC, <https://en.wikipedia.org/w/index.php?title=T-closeness&oldid=905775506> [accedit 6 octubre 2019]

[10] Wikipedia contributors, 'L-diversity', Wikipedia, The Free Encyclopedia, 9 agost 2019, 10:16 UTC, <https://en.wikipedia.org/wiki/L-diversity> [accedit 29 desembre 2019]