Seguridad, privacidad y aspectos legales
Prácticas de Anonimización

Álvaro López García (aloga@ifca.unican.es)

Máster en Ciencia de Datos
Universidad de Cantabria – Universidad Internacional Menéndez Pelayo

Parte I
Ejercicios

1. Linking data, data aggregation

Una red social ha hecho público el siguiente dataset y considera que la anonimización realizada es suficiente.

Name	Gender	Age	City of birth	Favorite TV Series	Relationship Status
*	male	19-25	Bilbao	Game of Thrones	single
*	female	16-18	Santander	Game of Thrones	in relationship
*	male	12-15	Barcelona	Friends!	in relationship
*	female	19-25	Madrid	Big Bang Theory	in relationship
*	female	19-25	Oviedo	Big Bang Theory	single
*	female	19-25	Bilbao	Game of Thrones	single
*	male	16-18	Santander	Game of Thrones	single
*	female	12-15	Barcelona	Game of Thrones	in relationship
*	male	19-25	Madrid	Big Bang Theory	single

Cuadro 1: Dataset anonimizado de una red social.

Sin embargo, a través de dicha red social, se pueden ver las votaciones de algunos usuarios, resultando en la información mostrada a continuación en la Tabla XXX.

Name	Email	TV Show	Rating (1=bad, 5=great)
Alice	alice1995@email.com	Friends!	1
Bob	bobbybob@email.com	Friends!	4
Charlie	s9charchar@email.com	Friends!	2
Eve	evelyn@myhighscool.com	Friends!	1
Bob	bobbybob@email.com	Game of Thrones	1
Alice	alice1995@email.com	Game of Thrones	5
Charlie	s9charchar@email.com	Game of Thrones	5
Bob	bobbybob@email.com	Big Bang Theory	3
Charlie	s9charchar@email.com	Big Bang Theory	5
Alice	alice1995@email.com	Big Bang Theory	2
Eve	evelyn@myhighscool.com	Big Bang Theory	5

Cuadro 2: Información obtenida mediante scraping, de la misma red social que la Tabla 1.

Para este ejemplo podemos asumir que los usuarios de ambos dataset son comunes. Responde a las siguientes preguntas.

Question 1

¿Qué información podemos obtener de Alice?

Question 2

¿Qué información podemos obtener de Charlie?

Question 3

¿Qué información podemos obtener de Bob?

2. K-anonimato

Datos los siguientes conjuntos de datos:

ID	Age	Gender	Fav.Show
1	12-15	female	Friends!
2	19-25	male	Friends!
3	19-25	male	Friends!
4	12-15	female	Friends!
5	19-25	male	G.o.T.
6	19-25	male	G.o.T.
7	19-25	male	G.o.T.

ID	Age	Gender	Fav.Show
1	19-25	female	Grey's A.
2	19-25	female	Simpsons
3	19-25	female	Futurama
4	19-25	female	Friends!
5	19-25	male	G.o.T.
6	19-25	male	C.Minds
7	19-25	male	Br.Ba.

ID	Age	Gender	Fav.Show
1	19	male	Friends!
2	19	male	Friends!
3	19	male	Friends!
4	19	female	Friends!
5	20	male	G.o.T.
6	20	male	G.o.T.
7	20	male	G.o.T.

Question 4

¿Cual es el k-anonimato para cada uno de los anteriores conjuntos de datos?

Parte II
Prácticas

Prerequisitos

Para realizar las prácticas es necesario tener instalado ARX en nuestro ordenador, así como los ficheros contenidos en el archivo comprimido `anonimizacion.tar.gz` que se encuentran en Moodle.

Info: La versión del software que vamos a utilizar es la 3.9.

ARX es una herramienta open source para transformar datos estructurados (tabulares) que contengan datos personales o sensibles utilizando métodos de anonimización o de control de divulgación estadística (*Statistical disclosure control*). ARX soporta la transformación de conjuntos de datos según el modelo de privacidad definido por el usuario, asegurando que se cumplan los parámetros establecidos por el mismo. De esta forma se pueden realizar análisis de riesgos adecuados y mitigar ataques que puedan comprometer la privacidad de los datos.

3. Interfaz ARX

3.1. Carga de proyectos y conjuntos de datos

ARX requiere utilizar un proyecto nuevo para cada conjunto de datos que vayamos a anonimizar. Los ficheros de proyecto se guardan con extensión `.deid`. Esto nos permite trabajar en un proyecto a lo largo del tiempo.

Una vez creado el proyecto es necesario importar un conjunto de datos sobre el que trabajar. ARX permite importar desde diferentes formatos, especificando características de los datos que vamos a leer (las cuales se pueden cambiar más adelante).

3.2. Interfaz

El interfaz tiene tres pestañas principales (Figura 3.2):

- *Configure transformation*: Nos permite configurar las propiedades del conjunto de datos, así como las transformaciones que vamos a aplicar.
- Una vez cargado el conjunto de datos, este se muestra en la parte izquierda de la pantalla, en la pestaña
- La parte inferior nos permite especificar el subset de datos con el que vamos a trabajar
- La parte de la derecha está dividida en varias áreas.
    - El área superior nos permitirá especificar qué tipo de transformaciones vamos a llevar a cabo así como consultar sus atributos.
    - El área central nos permitirá definir que modelo de privacidad queremos aplicar.
    - La parte inferior nos permite definir parámetros generales del proceso de anonimización, las características poblacionales de nuestro conjunto de datos y factores para analizar el coste/beneficio de la publicación del conjunto de datos.
- *Explore results*: Nos permite visualizar el espacio de resultados del proceso de anonimización.
- *Analyze utility*: Nos permite analizar la utilidad del conjunto de datos anonimizado, así como visualizar las transformaciones realizadas.
- *Analyze risks*: Nos permite realizar un análisis de los riesgos de publicar el conjunto de datos.

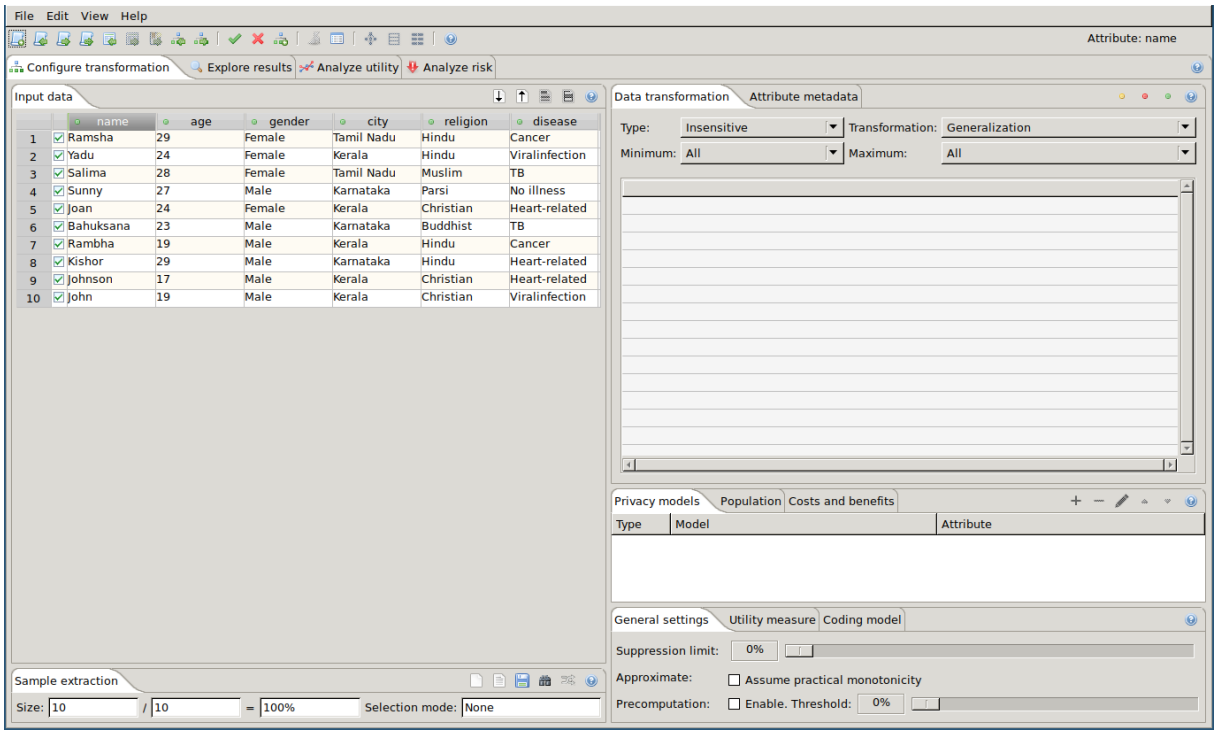

Figura 1: Intefaz de usuario de ARX

4. Conjunto de datos `hospital.csv`

4.1. Ejercicio 1

Cargar el conjunto de datos `hospital.csv` e identificar los tipos de atributos que debemos aplicar a las diferentes columnas. Familiarizate con la interfaz de usuario.

Para este ejercicio, solamente consideraremos atributos identificadores y cuasi-identificadores (es decir, sin datos sensibles). Pese a que es un atributo sensible, para este ejemplo no queremos realizar ninguna transformación ni aplicar ningún modelo a la columna *disease*, por lo que lo marcaremos como dato no sensible.

Una vez clasificados los atributos correctamente, realizar las transformaciones que creamos necesarias en los cuasi-identificadores. Para este ejemplo, deberemos realizar generalizaciones y/o supresiones en 3 columnas. Asegúrate que el *supression limit* de la pestaña *General settings* está a 100 %. Esto quiere decir que ARX va a suprimir registros si es necesario para anonimizar.

Aplicar un modelo de privacidad basado en k-anonimato con $k = 2$. Aplicar el modelo y evaluar los resultados.

Question 5

¿Se corresponde con el ejemplo realizado en clase?

Question 6

¿Cuantas clases de equivalencia se han generado?

Question 7

¿Se ha eliminado algún registro?

4.2. Ejercicio 2

Partiendo del resultado del Ejercicio 1, se quiere realizar la clasificación en grupos de edad a intervalos de 5 años. Para ello es necesario crear una jerarquía de generalización y reconfigurar el intervalo de edades que va a realizar la transformación para utilizar un intervalo de 5. Crearemos diferentes niveles en esta jerarquía de generalización, agrupando los niveles anteriores (con lo que tendremos intervalos de 5, de 10, de 20 y de 40 años, por ejemplo). Una vez hecho esto, anonimizar el conjunto de datos y ver los resultados obtenidos.

Como se puede ver, el nuevo conjunto de datos se corresponde al del ejercicio realizado en clase. Esto es debido a que ARX ha encontrado que esta es la mejor combinación de transformaciones posibles. Sin embargo, nosotros deseamos forzar el intervalo a 5 años, por lo que necesitaremos explorar los diferentes resultados a través de la pestaña *Explore results* (Figura 4.2). Analiza la visualización de los datos y familiarizate con ella.

La parte superior de la pantalla muestra una visualización del espacio de transformaciones en diferentes vistas. La parte inferior está dividida en tres partes.

- En la parte izquierda se pueden filtrar los resultados, ya sea por el número de transformaciones a realizar para cada columna o por la puntuación con respecto a la métrica de utilidad que ARX ha calculado por nosotros.
- La parte central contiene un listado con todos los resultados, tanto los que ARX ha generado como los que nosotros queramos guardar.
- La parte derecha nos muestra las propiedades de la transformación seleccionada. Un score bajo significa que es mejor con respecto a otro más alto.

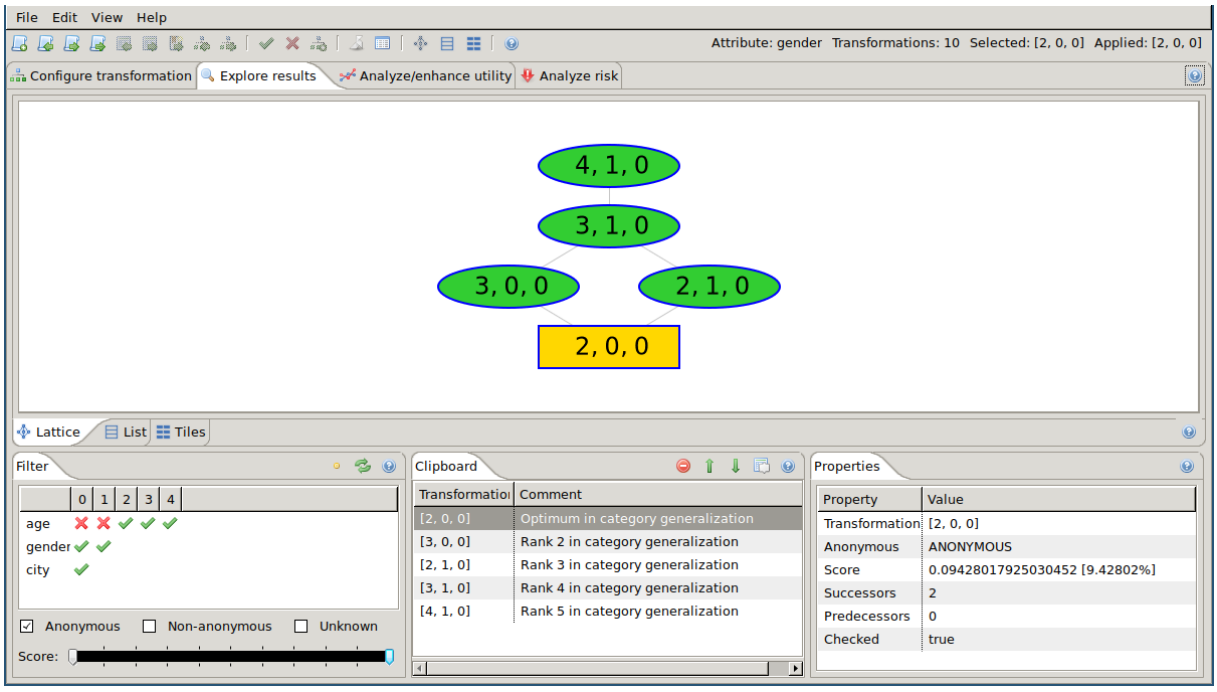

Figura 2: Intefaz de usuario de ARX: Explorando resultados

Puesto que queremos forzar a que sean intervalos de 5 años, filtramos la transformación que nos va a producir ese resultado. Aplicar las dos transformaciones resultantes y observar los resultados.

4.3. Ejercicio 3

A partir del ejercicio anterior, cambiar el *supression limit* de la pestaña *General settings* a $0\,\%$. Esto quiere decir que ARX no va a suprimir ningún registro.

Question 8

Realiza de nuevo la anonimización y busca una transformación que nos permita utilizar un intervalo de 5 años. ¿Qué sucede?

4.4. Ejercicio 4

Cargar el conjunto de datos `hospital_extended.csv` y realizar la configuración de los ejercicios anteriores, cambiando la columna *disease* a *sensitive*. Asegúrate que el limite de eliminación esta de nuevo al 100 % y aplica los siguientes modelos de privacidad:

- *k-anonimity* con $k = 2$
- *l-diversity* en la columna *disease* con $l = 2$

Aplica la anonimización y estudia los resultados.

Question 9

¿Qué ha sucedido? Cambia el ratio de supresión a 0 y estudia de nuevo los resultados.

5. Conjunto de datos `adult.csv`

5.1. Ejercicio

Cargar el conjunto de datos `adult.csv` con las siguientes columnas: *sex, age, race, marital-status, education, native-country, workclass, occupation, salary-class*.

Anonimizar el conjunto de datos del censo de los EEUU de 1996 (fichero `adult.csv`) teniendo en cuenta las siguientes acciones:

- Suprimir información de raza.
- Generalizar a intervalos la edad.
  - Rango: 0—100, agrupando $>= 80$ años.
  - Comenzando en intervalos de 5 años, realizar 6 niveles de transformación.
- Transformar todas las columnas, teniendo en cuenta lo siguiente:
  - 1 transformación: *sex, race, salary-class*.
  - 2 transformaciones: *marital-status, native-country, workclass, occupation*.
  - 3 transformaciones: *education*.

Para ello se pueden utilizar los CSV en el directorio `jerarquias`, aplicándolo a cada una de las diferentes columnas. Establecer el peso de la edad a 1 para indicar la importancia de este atributo con respecto al resto. Asegurarse de que el *supression limit* esté a $100\,\%$.

Establecer un k-anonimato de $k = 5$ y aplicar la anonimización. Explorar los resultados y exportar el conjunto de datos.

Question 10

¿Cuantos registros se han eliminado en este caso?

Una vez hecho eso, buscar una transformación que permita utilizar intervalos de 5 años y que siga siendo cercano al óptimo. Aplicar la transformación y exportar el conjunto de datos.

Question 11

¿Cuántos registros se han eliminado?

Question 12 **Entrega**

Adjuntar en Moodle ambos conjuntos de datos exportados (intervalos de 5 y de 10 años), así como el proyecto de ARX generado (fichero con extensión `.deid`).

Adjuntar en el texto en línea el número de registros eliminados.