Introduciendo la K-anonimización a los estudiantes del Grado en Ingeniería Informática: una práctica de la asignatura 'Análisis y Diseño de Algoritmos'

M. Mercedes Martínez González[1]
Departamento de Informática de la Universidad de Valladolid
mercedes@infor.uva.es

Este artículo es una versión previa y extendida de la publicación "Introducción práctica a la K-anonimización. LA LEY Privacidad (2), pp.1-8, 2019. Wolters Kluwer. ISSN 2659-8698". Por favor, use esta última referencia.

RESUMEN

La K-anonimización es una técnica que se utiliza para salvaguardar la identidad de los sujetos cuyos datos se comparten. Los responsables y encargados de los tratamientos de los datos, en aplicación del principio de responsabilidad proactiva que impone el Reglamento General de Protección de Datos (RGPD), deben tomar las medidas oportunas para evitar la reidentificación de los sujetos a partir de estos datos, siendo una de ellas la utilización de la K-anonimización. En esta línea, la Agencia Española de Protección de Datos (AEPD) ha publicado una nota técnica sobre K-anonimización en la cual introduce este concepto y proporciona un conjunto de pautas y herramientas para llevar a cabo esta tarea. Una tarea cuya responsabilidad habrá de recaer en muchos casos en profesionales de la informática encargados del almacenamiento y tratamiento de datos.

En este artículo se presenta una experiencia reciente en la asignatura *Análisis y Diseño de Algoritmos (ADA)*, del tercer curso del Grado en Ingeniería Informática de la Universidad de Valladolid. Durante el curso 2018/2019 se planteó una actividad práctica para introducir el concepto de K-anonimización a los alumnos de Informática y enfrentarles al reto de proponer e implementar un algoritmo para resolver este problema. De este modo, además de familiarizar a los alumnos con las nociones básicas sobre K-anonimización, se consiguió que relacionasen las técnicas utilizadas para este fin con los conocimientos que adquieren en sus estudios actuales.

Palabras clave: k-anonimización, algoritmos, anonimización, informática

ABSTRACT

K-anonymity is a technique used to protect identity when data are shared. The … introduced by the General Data Protection Regulation (GDPR) means that people in charge of data treatments should adopt the appropriate measures in order to prevent that persons whose data are shared can be reidentified. K-anonymity is a technique used for this aim. Related with it, the Spanish Data Protection Authority (Agencia Española de Protección de Datos, AEPD) published a technical note in which k-anonymity is

introduced together with a set of guidelines for using it. It seems reasonable that the anynymization task will in many cases fall under the responsability of computing professionals in charge of data management.

During the 2018/2019 academic year, an experience introducing K-anonymity to the Computer Engineering students following the third year course about Algorithms (Análisis y Diseño de Algoritmos, ADA). Besides introducing the basics of K-anonymity to the students, benefits of this experience was that students established a relationship between their knowledge about graphs and algorithms with the anonymity proble.

Keywords: k-anonymity, algorithms, anonymization, computing

1 INTRODUCCIÓN

Compartir datos o publicarlos para que estén disponibles para investigaciones, es cada vez más habitual en un contexto donde la disponibilidad de cantidades masivas de datos y de técnicas de análisis que permiten extraer nuevo conocimiento hace especialmente interesante explotar esta posibilidad. Los avances que se pueden conseguir gracias a este conocimiento, en nuevos tratamientos médicos, en gestión de riesgos y/o catástrofes, etc. son muy valorados. Sin embargo, es necesario combinar estas ventajas con la protección de la identidad de las personas cuyos datos nutren estos conjuntos. El campo de la salud, probablemente por la especial sensibilidad que todos tenemos hacia este tipo de información, es uno de los entornos más utilizados para ilustrar la importancia de esta clase de tratamientos.

En su considerando 26 el Reglamento General de Protección de Datos (RGPD) [Unión Europea, 2016] distingue entre datos seudonimizados y anonimizados, dejando los primeros bajo su ámbito de aplicación, y excluyendo los segundos: "*Los principios de la protección de datos deben aplicarse a toda la información relativa a una persona física identificada o identificable. Los datos personales seudonimizados, que cabría atribuir a una persona física mediante la utilización de información adicional, deben considerarse información sobre una persona física identificable. […] Por lo tanto los principios de protección de datos no deben aplicarse a la información anónima, es decir información que no guarda relación con una persona física identificada o identificable, ni a los datos convertidos en anónimos de forma que el interesado no sea identificable, o deje de serlo.*". La Opinión 05/2014 del grupo de trabajo del artículo 29 de la Directiva 45/96 discute varias técnicas de anonimización, estableciendo sus fortalezas y debilidades desde la perspectiva de su capacidad para garantizar las medidas jurídicas relativas a la protección de datos personales, en particular las posibilidades de reidentificación. Más recientemente, el Reglamento (UE) 2018/1807 relativo a la libre circulación de datos no personales en la Unión Europea [Unión Europea, 2018] se preocupa también por la posibilidad de conseguir reindentificar individuos a partir de datos no personales: "*Entre los ejemplos específicos de datos no personales se encuentran los conjuntos de datos agregados y anonimizados utilizados para análisis de datos a gran escala, los datos sobre agricultura de precisión que pueden ayudar a controlar y optimizar la utilización de plaguicidas y de agua, o los*

datos sobre las necesidades de mantenimiento de máquinas industriales. Si los avances tecnológicos hicieran posible transformar datos anónimos en datos personales, dichos datos se deben tratar como datos personales y, en consecuencia, se debe aplicar el Reglamento (UE) 2016/679". En concordancia con esta preocupación, la Agencia Española de Protección de Datos (AEPD) está adaptando sus guías al RGPD. Recientemente (julio de 2019) ha publicado una Nota Técnica dirigida a los responsables y encargados de los tratamientos de datos sobre el uso de la K-anonimidad (o k-anonimización) como medida de la privacidad [Agencia Española de Protección de Datos, 2019]. En ella, la AEPD introduce el concepto de K-anonimidad, proporciona un listado de herramientas que implementan la K-anonimización y da pautas sobre las bondades y limitaciones de esta técnica.

La K-anonimización es una técnica que se utiliza para prevenir ataques donde la agregación de datos permita la reidentificación de las personas a partir de los datos recolectados de diferentes fuentes. Responde a la necesidad de resolver el problema que supone la posibilidad de que, a través de técnicas de análisis e inferencia, se llegue a identificar personas a partir de datos de los cuales se han retirado los valores de atributos que habrían permitido su identificación directa (DNI, número de Seguridad Social, nombre y apellidos, etc.). Someramente, estos ataques se basan en agregar información de diversas fuentes para crear clases o categorías más específicas que las que aportan los conjuntos de datos utilizados como fuente de estos procesos, cuyo número de individuos sea cada vez más pequeño, hasta conseguir un número tan pequeño que sea posible identificar personas concretas.

Las implicaciones que el cumplimiento normativo en materia de privacidad tiene sobre el desempeño profesional de los profesionales del sector TIC son muchas. La privacidad desde el diseño, la implantación de medidas de seguridad que protejan los sistemas donde se alojan las bases de datos, son cuestiones que afectan directamente las tareas que estos profesionales desempeñan. Es lógico suponer que aquellos que se ocupan del almacenamiento, consulta y tratamiento de los datos, estarán también relacionados con los procesos de anonimización.

Instrucciones y guías como las ofrecidas por la AEPD son una valiosa ayuda para estos profesionales. Pero no es menos cierto que la formación de nuestros alumnos debe adaptarse aún a la eclosión de la preocupación por la protección de la privacidad, llegada de la mano de la entrada en vigor del RGPD en Europa en mayo de 2018. Los actuales planes de estudio se diseñaron con anterioridad, atendiendo al Libro Blanco del Título de Grado en Ingeniería Informática [ANECA, 2004], que se elaboró en 2004. La adaptación está en camino, la *Association for Computer Machinery, ACM*, incluyó en 2017 un sexto volumen sobre Ciberseguridad (*Cibersecurity*) en el *Computing Curricula* [ACM, IEEE-CS, 2017], una de las referencias más reconocidas en la elaboración de planes de estudio de Informática a nivel mundial, referencia también del Libro Blanco mencionado. Este volumen dedica una considerable atención a materias y aspectos relacionados con la privacidad. En los próximos años, es previsible que veamos aparecer planes de estudios vinculados a este volumen. Entretanto, es posible preguntarse qué instrumentos tienen nuestros actuales alumnos para abordar este tipo de responsabilidades y cómo hacerles conscientes de ellos.

En este artículo se presenta la experiencia realizada en la asignatura *Análisis y Diseño de Algoritmos (ADA)*, del tercer curso del *Grado en Ingeniería Informática* de la

Universidad de Valladolid. Durante el curso 2018/2019 una de sus prácticas sirvió para introducir a los alumnos de Informática en el concepto de K-anonimidad y enfrentarles al reto de proponer e implementar un algoritmo para resolver este problema. De este modo, además de familiarizar a los alumnos con las nociones básicas sobre K-anonimidad, se buscaba que relacionasen las técnicas utilizadas para la anonimización con los conocimientos que adquieren en sus estudios actuales, en concreto con sus conocimientos sobre algoritmos. Así, los alumnos toman contacto con los conceptos básicos sobre anonimización, con algunos de los problemas que se plantean y contemplan los retos a los que se enfrentan quienes abordan estas tareas. Un trabajo como el abordado en la práctica objeto de este artículo les ayuda a asociar este tipo de técnicas de anonimización con las técnicas de diseño de algoritmos que estudian, y a aplicar lo aprendido sobre análisis y comparación de algoritmos a los algoritmos que valoran para resolver su problema de K-anonimización.

A continuación se presentan los conceptos básicos sobre K-anonimización que se proporcionaron a los alumnos para esta práctica y el ejemplo con el que se les introdujo este problema. La sección 3 se dedica a describir la práctica: el contexto en el que se aborda, planteamiento y retos que los alumnos tienen que superar. Las Conclusiones reflexionan sobre el resultado de esta experiencia, e incluyen algunas consideraciones sobre el enfoque que se da al problema de K-anonimización en esta práctica y la orientación de la Nota sobre K-anonimidad de la AEPD.

2 CONCEPTOS BÁSICOS SOBRE K-ANONIMIZACIÓN

La necesidad de combinar adecuadamente las ventajas que supone compartir datos con la protección de la privacidad ha hecho de la anonimización un fin y un reto. La anonimización obvia y más sencilla, consiste en eliminar los atributos identificadores, aquellos que por sí mismos, sin necesidad de combinarlos con ningún otro, son capaces de identificar un individuo, sin posibilidad de confusión con ningún otro. Estos atributos suelen ser DNI, número de pasaporte, número de Seguridad Social, etc. Se eliminan cuando se comparten datos. Sin embargo, no es suficiente. Conseguir las técnicas de anonimización adecuadas, capaces de garantizar a la vez la protección de la identidad de las personas con la utilidad de los datos, es una tarea que ha ido volviéndose más compleja a medida que la disponibilidad de más datos, y de combinarlos con otros conjuntos de datos externos, ha hecho posible reidentificar personas allí donde en principio parecía imposible. Atributos como las fechas de nacimiento, sexo, o código postal, combinados de modo adecuados con otros datos, pueden llegar a ser la clave para reconocer la identidad de los individuos cuyos datos se exponen.

La K-anonimización es una técnica que se utiliza para prevenir este tipo de ataques, caracterizados por utilizar la agregración de datos de diversas fuentes para conseguir identificar los individuos cuya identidad se había intentado ocultar. Se basa en generalizar y/o suprimir porciones de datos, con el fin de que sea imposible distinguir un individuo entre un conjunto de k individuos que comparten características similares, de ahí el nombre de K-anonimización. Un ejemplo habitual para explicar esta idea suelen ser datos sanitarios, que se combinan con datos extraídos de algún censo. Así se consigue ver de modo sencillo los riesgos de la agregación y el objetivo de la K-anonimización. Uno de ellos se utilizó para esta práctica sobre K-anonimización. En las

tablas 1 y 2 se muestran dos pequeños conjuntos de datos, suficientes para que los alumnos entiendan el problema.

La tabla 1 contiene datos de pacientes, para cada uno de ellos: fecha de nacimiento (Fnac), sexo, código postal y dolencia que sufre. La tabla 2 son datos datos extraídos de algún censo: nombre, fecha de nacimiento, sexo y código postal. Aunque se ha eliminado de la tabla 'Datos de pacientes' (tabla 1) la información sobre DNI, e incluso nombre y apellidos (nótese que no hay columnas para estos atributos), si se cruzan sus datos con los de la tabla 'Datos del censo' (tabla 2) es fácil darse cuenta de que es posible saber que el paciente que padece "gripe" es Andrés; para llegar a esta conclusión es suficiente con fijarse en los valores de los atributos *sexo* y *código postal*, y contrastarlos con los datos de censo, ya que únicamente Andrés tiene valores "M" y "53715" en esos dos atributos. Éste es un ejemplo sencillo, con muy pocos datos, pero es fácil imaginar situaciones similares en la llamada 'España vacía' [Molino, 2016] si utilizamos los códigos postales combinados con algún otro atributo como la fecha de nacimiento. En nuestro ejemplo, los datos que se pretende preservar son los de pacientes, pero el cruce con los datos del censo (fuente externa) permiten identificar a los pacientes cuya identidad se quería proteger. Así pues, será necesario realizar una anonimización adecuada de los datos de pacientes, para resolver este problema y que deje de ser posible identificar a pacientes como Andrés. Lo vamos a hacer utilizando la K-anonimización.

Datos de pacientes

Fnac	**Sexo**	**Cod_postal**	**Dolencia**
21/1/86	M	53715	Gripe
13/4/96	F	53715	Neumonía
28/2/86	M	53703	Bronquitis
21/1/86	M	53703	Fractura brazo
13/4/96	F	53706	Apendicitis
28/2/86	F	53706	Fractura pierna

Tabla 1. Datos de pacientes utlizados como ejemplo en la práctica de la asignatura ADA.

Datos del censo

Nombre	**Fnac**	**Sexo**	**Cod_postal**
Andrés	21/1/86	M	53715
Beatriz	10/1/91	F	55410
Cecilia	1/10/54	F	90210
David	21/2/94	M	02174
Elisa	19/4/82	F	02237

Tabla 2. Datos del censo utlizados como ejemplo para explicar el problema de K-anonimización en la práctica de la asignatura ADA.

2.1 Definiciones útiles

Es necesario conocer algunas definiciones para trabajar con este problema:

1) **Conjunto de atributos identificadores**: conjunto de atributos que, de forma autónoma, permite identificar registros individuales. Suelen ser atributos como DNI, nombre, número de Seguridad Social… Se eliminan u ocultan como paso previo a la K-anonimización. En consecuencia, no se trabaja sobre ellos en la práctica.

2) **Conjunto de atributos sensibles**: conjunto de atributos que contienen información sensible, por ejemplo, sobre problemas de salud, financieros, etc. Interesan para entender el objetivo de la anonimización. Sin embargo, en la práctica se trabaja con el conjunto de atributos cuasi-identificadores.

2) **Conjunto de atributos cuasi-identificadores, *q***: conjunto de atributos que, unido a alguna información externa, permite identificar registros individuales. En el ejemplo de este enunciado el conjunto de atributos cuasi-identificadores seleccionado es *(sexo, cod_postal)*.

3) **Frecuencia de valores**: para cada combinación de valores de atributos cuasi-identificadores, *q*, número de registros que toman ese valor. Por ejemplo, en los datos de pacientes (tabla 1), frecuencia(sexo=M;cod_postal=53715)=1 y frecuencia(sexo=M; cod_postal=53703)=2.

4) **Propiedad de *k*-anonimidad**: un conjunto de datos cumple la propiedad de k-anonimidad respecto a *q* si la frecuencia de cualquier combinación de valores de *q* es mayor o igual que *k,* esto es, si hay **al menos** *k* individuos para los cuales los valores de q coinciden. Tiene que cumplirse para **todas** las combinaciones posibles de valores de q. Por ejemplo, si nos fijamos en las frecuencias de arriba, vemos que con el conjunto de atributos cuasi-identificadores (sexo, cod_postal), no se cumple la propiedad de 2-anonimidad (k=2), porque para los valores sexo=M y cod_postal=53715 hay un único individuo, frecuencia(sexo=M; cod_postal=53715)=1. Sin embargo, los datos de las tablas 3, 4 y 5 sí la cumplen.

2.2 Cómo conseguir la propiedad de k-anonimidad: generalización sobre los atributos cuasi-identificadores

Se trabaja sobre los valores de los atributos cuasi-identificadores, alterándolos hasta conseguir que se cumpla la propiedad de k-anonimidad. Las estrategias básicas son la **generalización** y la **eliminación** de registros. Ambas, así como las posibilidades de combinarlas, se abordan en la Nota Técnica de la AEPD sobre K-anonimización. Aquí nos centraremos en la generalización, que es la estrategia sobre la que se pidió trabajar a los alumnos. Ésta consiste en generalizar, en varios pasos si es necesario, los valores de los atributos cuasi-identificadores. Si pierden especificidad, será más fácil que varios individuos compartan los mismos valores. Para la práctica se utilizaron las jerarquías de generalización de las figuras 1 y 2, del modo que se explica en la sección 3.1.

2.3 Dificultades en la K-anonimización

Existen varias dificultades en la K-anonimización, todas ellas relacionadas con decisiones que deben tomarse antes de aplicar los algoritmos de K-anonimización. Una de ellas es la elección de k. Debe ser suficientemente alto para dificultar la reidentificación, pero no tanto que los datos pierdan su valor por exceso de generalidad. Sobre este asunto se pueden encontrar indicaciones en la Nota Técnica de la AEPD. Aquí no nos extenderemos más, puesto que la elección de k no fue una decisión que los alumnos debían abordar.

3 K-ANONIMIZACIÓN Y LA ASIGNATURA DE ANÁLISIS Y DISEÑO DE ALGORITMOS: UN TRABAJO PRÁCTICO

3.1 La K-anonimización y el tema de grafos de la asignatura Análisis y Diseño de Algoritmos

La asignatura *Análisis y Diseño de Algoritmos* es una asignatura del tercer curso del Grado en Ingeniería Informática, de carácter obligatorio. Tras varias asignaturas donde los alumnos han aprendido diversos paradigmas y técnicas de programación, ésta es la primera donde profundizan en la noción de algoritmo, distinguen claramente algoritmo y programa, aprenden a medir la calidad de un algoritmo en términos de coste y se introducen en las técnicas de diseño de algoritmos. La asignatura incluye dos trabajos prácticos, 'las prácticas', que los alumnos realizan en grupo. La segunda es una práctica donde se les plantea un problema y se les pide que propongan un algoritmo para resolverlo, para luego programarlo. No se les indica qué algoritmo deben utilizar, seleccionar el algoritmo adecuado es uno de sus retos. Han de mostrar cuál es la técnica de diseño de algoritmos, de entre las aprendidas de la asignatura, que utiliza su propuesta, y qué algoritmo o algoritmos de los vistos en la asignatura podrían servir para resolver el problema. Una vez seleccionado un algoritmo, proceden a su implementación y posterior prueba con diversos conjuntos de datos de entrada. El programa debe cumplir los requisitos indicados en el enunciado de la práctica, en particular en términos de entradas y salidas, así como parámetros que son susceptibles de variar en distintas ejecuciones.

Para la segunda práctica del curso 2018/2019 se buscaba un problema de grafos en el que pudieran aplicar alguno, o varios, de los algoritmos que los alumnos estudian en el tema sobre grafos. Para preparar el enunciado se utilizó como referencia el artículo de LeFevre, DeWitt y Ramakrishnan (2005). En este artículo se presenta *Incógnito*, uno de los algoritmos de K-anonimización más populares. Sin embargo, se decidió utilizar como referencia para preparar el enunciado de la práctica las secciones primera y segunda del artículo, que presentan el problema y el estado del arte. Entre las posibilidades presentadas se incluye la utilización del algoritmo de recorrido en anchura de un grafo, opción asequible para esta práctica, dado que los alumnos han visto los algoritmos de recorrido de grafos en la asignatura.

A los alumnos se les presentó el problema de la K-anonimización, que no conocían, con los conceptos necesarios para entender el problema que tienen que resolver en la práctica, que se han presentado en la sección 2. Se les planteó que debían recurrir a sus

conocimientos de grafos para resolver la práctica. Uno de los retos que puede llegar a presentar mayor dificultad cuando se aborda un problema de grafos es caracterizar correctamente el grafo sobre el que se debe trabajar, esto es, decidir cuáles son los nodos y las aristas del grafo. Una elección acertada del grafo sobre el que se trabaja puede facilitar mucho encontrar un algoritmo adecuado para el problema que se pretende resolver. En esta práctica se optó por guiar a los alumnos en la elección del grafo y cómo construirlo. Las secciones 3.1.1 y 3.1.2 aportan estas indicaciones.

3.1.1 Jerarquías de generalización de los atributos cuasi-identificadores

El conjunto de atributos cuasi-identificadores, *q*, elegido en el ejemplo es *q=(sexo, cod_postal)*. Los posibles valores del atributo *sexo* son los que se ven en la generalización de la figura 1. En la parte izquierda de la figura se muestran por niveles, los más específicos en el nivel 0 y los más generales en el nivel más alto. En la parte derecha de la figura se muestra la relación de generalización entre ellos. Para aplicar la generalización sobre los valores de ese atributo, se parte del nivel más bajo, que tiene los valores más específicos, y se recorre el árbol hacia arriba. El número de niveles que se sube es el coste. Por ejemplo, si nos fijamos en el atributo sexo, 'Persona' es más general que 'Masculino' y 'Femenino', y esta generalización se consigue con el mínimo coste posible, 1, porque ha sido necesario subir sólo un nivel. Los posibles valores del atributo cod_postal se muestran en la figura 2. Generalizar el valor 53706 del código postal hasta obtener el valor 537** tiene un coste=2.

Figura 1. Jerarquía de generalización para el atributo sexo.

Figura 2. Jerarquía de generalización para el atributo cod_postal.

3.1.2 Grafos para anonimización y estrategia básica para utilizarlos

Para anonimizar se construye un grafo para el conjunto de atributos cuasi-identificadores. Se trata de un *grafo de combinaciones de generalizaciones*. Cada combinación de generalizaciones representa las posibles combinaciones de valores de los atributos cuasi-identificadores que se pueden obtener. Si nos fijamos en las jerarquías de las figuras 1 y 2, comprobamos que <s1,p1> representa todas las combinaciones que se obtienen combinando el valor 'Persona' del atributo sexo con los valores 5370* ó 5371* del código postal, que son dos: (Persona, 5370*), (Persona, 5371*).

Recorriendo el grafo se pasa de una combinación de valores de atributos cuasi-identificadores, *q*, a otra con un nivel de generalización diferente. Se puede asociar a cada nodo de este grafo un conjunto de valores enteros, tantos como atributos tiene *q*, donde cada uno de ellos indica el número de pasos que es necesario dar para llegar desde el valor que dicho atributo tiene en los datos de entrada (sin generalizar) hasta el valor que se obtendría sustituyéndolo por un valor más general. Por ejemplo, la combinación <s1,p1> indica que se ha generalizado una vez sobre el atributo sexo y una sobre el código postal. Los costes asociados son (1,1). En el lado izquierdo de la figura 3 se muestra el grafo de generalizaciones que se obtiene para el ejemplo con el conjunto de atributos cuasi-identificadores q: (sexo,cod_postal), y en el lado derecho el grafo con los costes asociados a cada una de las combinaciones.

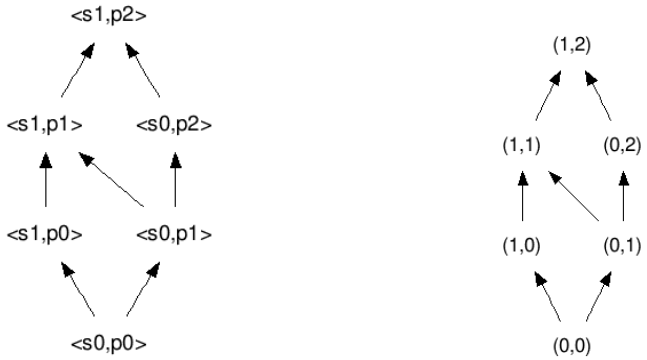

Figura 3. Grafo de combinaciones de generalizaciones para el conjunto de atributos *q=(sexo,cod_postal)*

3.2 Tareas de los alumnos

Los alumnos debían proporcionar un programa capaz de decidir qué combinación de generalización utilizar, para posteriormente aplicarla al conjunto de datos de entrada. Se considera que una combinación es **óptima** si se ha obtenido llevando a cabo las generalizaciones mínimas necesarias para garantizar la propiedad de k-anonimidad. En el ejemplo existen varias combinaciones que cumple la propiedad de k-anonimidad siendo k=2: *<s1,p0>, <s1,p1>, <s0,p2>*, y todas aquellas donde se generaliza aún más sobre estos dos atributos. *A*l anonimizar los datos de la tabla 1 usando la combinación óptima, <s1,p0>, se obtienen los datos anonimizados de la tabla 3. Los datos que se obtienen usando <s1,p1> son los de la tabla 4. Por último, la tabla 5 corresponde a la combinación <s0,p2>. De las combinaciones anteriores, sólo la combinación *<s1, p0>* es óptima (con subir un nivel de generalidad en el atributo sexo es suficiente, no es

necesario modificar los valores del código postal). Se busca aplicar combinaciones óptimas, porque se generaliza lo mínimo necesario para conseguir el fin buscado, la propiedad de k-anonimidad. La lista de tareas que los alumnos debían abordar en la práctica es la siguiente:
1) Construir el grafo de generalizaciones para el conjunto de atributos cuasi-identificadores, *q*.

2) Construir el grafo parejo que contiene los costes.

3) Buscar una combinación óptima. Es la principal tarea a la que se enfrentaban los alumnos.

4) Construir un subprograma ***k_anon(<valores_q>)***, que tomaba como entrada una combinación de generalizaciones de los atributos cuasi-identificadores, y debía devolver "cierto/falso" en función de si se cumplía la propiedad de k-anonimidad con esa combinación de q. Por ejemplo, para los datos de pacientes de la tabla 1, siendo *k=2*, *k_anon(s0,p0)=falso*, *k_anon(s0,p1)=falso, k_anon(s1,p0)=cierto, k_anon(s1,p1)=cierto* y *k_anon(s0,p2)=cierto.*

5) Generar una anonimización de los datos de entrada aplicando la combinación óptima.

Tal como se indicó en la sección 3.1, el programa de los alumnos debía trabajar con las entradas y salidas especificadas en el enunciado de la práctica. Las de esta práctica eran las que se muestran a continuación.

A) Entradas para el programa

El programa debía tomar como entrada:

1) Conjunto de datos de entrada. Se proporcionaron en un fichero. Podían ser distintos a los del ejemplo, con distintos atributos, procedentes de entornos diferentes.
2) Un conjunto de jerarquías de generalización de los valores de sus atributos. También disponibles en ficheros independientes, uno por cada atributo.
3) Un conjunto de atributos cuasi-identificadores, *q*, sobre los que se debía aplicar la anonimización en esa ejecución. Debía leerse por teclado.
4) Valor de *K*. También se leía por teclado en cada ejecución.

B) Salidas del programa

Las salidas que el programa debía producir en cada ejecución son:

1) Un fichero de datos anonimizado.
2) La combinación de generalizaciones que se utilizó para obtener los datos anonimizados.

Datos de pacientes

Fnac	**Sexo**	**Cod_postal**	**Dolencia**
21/1/86	persona	53715	Gripe
13/4/96	persona	53715	Neumonía

28/2/86	persona	53703	Bronquitis
21/1/86	persona	53703	Fractura brazo
13/4/96	persona	53706	Apendicitis
28/2/86	persona	53706	Fractura pierna

Tabla 3. Datos de la tabla 1 anonimizados usando la combinación de generalizaciones óptima para k=2, <s1,p0>.

Datos de pacientes

Fnac	**Sexo**	**Cod_postal**	**Dolencia**
21/1/86	persona	5371*	Gripe
13/4/96	persona	5371*	Neumonía
28/2/86	persona	5370*	Bronquitis
21/1/86	persona	5370*	Fractura brazo
13/4/96	persona	5370*	Apendicitis
28/2/86	persona	5370*	Fractura pierna

Tabla 4. Datos de la tabla 1 anonimizados usando la combinación <s1,p1>.

Datos de pacientes

Fnac	**Sexo**	**Cod_postal**	**Dolencia**
21/1/86	M	537**	Gripe
13/4/96	F	537**	Neumonía
28/2/86	M	537**	Bronquitis
21/1/86	M	537**	Fractura brazo
13/4/96	F	537**	Apendicitis
28/2/86	F	537**	Fractura pierna

Tabla 5. Datos de la tabla 1 anonimizados usando la combinación <s0,p2>.

3.3 Otras características de la práctica

Se tomaron algunas decisiones con el fin de ajustar la dificultad de la práctica al tiempo de trabajo que los alumnos deben dedicarle en función de su ubicación en el plan de

estudios y el número de créditos de la asignatura. Se decidió no incluir entre las tareas de los alumnos algunas que sí debe abordar el responsable de la aplicación de una K-anonimización. En concreto, los retos que no se incluyeron en la práctica son:

- Buscar el valor de K adecuado. Como se ha indicado en la sección 2, la elección de un valor de k adecuado no es evidente. En esta práctica el valor de K con el que se debía probar era un parámetro de entrada al programa.
- Elegir el conjunto de atributos cuasi-identificadores sobre el que trabajar. Como ya se ha dicho, el conjunto de atributos cuasi-identificadores, q, era otro de los parámetros de entrada al programa.
- Se restringió el método de trabajo sobre los atributos a la generalización y se les dieron las pautas para realizar la generalización sobre los atributos con los que trabajaban: o bien eran sustituciones/eliminaciones de caracteres sencillas como las mostradas para el atributo código postal, o bien se les daba directamente la generalización para cada atributo en un fichero que tomaban como entrada a su programa. Trabajar con jerarquías de generalización de los atributos cuasi-identificadores se presta a tomar la decisión sobre la técnica de diseño de algoritmos que se utilizará para ello, lo cual era un valor positivo en una práctica de esta asignatura.

4 CONCLUSIONES

La protección de la privacidad es un problema que concierne a prácticamente todos aquellos, empresas, instituciones o particulares, que manejan datos. La figura del Delegado de Protección de Datos (DPD) que introdujo el RGPD ha cobrado relevancia en una sociedad cada vez más digitalizada, que genera continuamente datos. Sin embargo, el procesamiento masivo de datos da lugar a un amplio abanico de tareas asociadas a la protección de la privacidad, desde la orientación sobre las posibilidades y efectos jurídicos y/o económicos de las decisiones que se tomen, hasta las técnicas sobre el diseño de sistemas, selección de algoritmos o evaluación de éstos, cuya cobertura requiere de equipos multidisciplinares. Equipos capaces de interaccionar con fluidez para dar respuesta a un asunto tan serio como es la privacidad, en los cuales profesionales de la informática con conocimientos sobre anonimización y evaluación de algoritmos habrán de participar. Es la reflexión sobre la formación de estos profesionales la que motiva la práctica descrita en este artículo.

La práctica sobre K-anonimización que se ha presentado fue la segunda práctica de la asignatura ‘Análisis y Diseño de Algoritmos’ del grado en Ingeniería Informática durante el curso 2018/2019. Una práctica no es suficiente para conocer el problema de la anonimización con la profundidad que sería deseable en los profesionales que asuman este tipo de tarea. No es posible abarcar aspectos como la evaluación de las vulnerabilidades de los datos anonimizados ni técnicas más elaboradas que la mostrada. Pero esta experiencia ha servido para explorar las posibilidades que hay para introducir en la formación de nuestros alumnos de Informática la conciencia sobre problemas como la protección de la privacidad a través de la anonimización.

Los resultados fueron buenos, al final de la práctica todos los grupos fueron capaces de aplicar y explicar la K-anonimización, también de manejar con adecuada soltura conceptos como ‘conjunto de atributos cuasi-identificadores’. Además establecieron una asociación entre sus conocimientos sobre coste mínimo de un algoritmo, que manejan

muy bien dado que están acostumbrados a enfrentarse con este tipo de retos a lo largo de sus estudios, con ideas como la de buscar el equilibrio entre el grado de generalización que se aplica en la anonimización y la utilidad de los datos.

Como se ha indicado, esta práctica se enmarca en la asignatura *Análisis y Diseño de Algoritmos*. En consecuencia, los alumnos debían trabajar sobre los algoritmos que utilizaban para k-anonimizar. En la Nota Técnica de la AEPD sobre K-anonimización se introducen los conceptos sobre K-anonimización, las limitaciones de esta técnica y se proporciona las referencias a un conjunto de herramientas para anonimizar. Lógicamente, no parece necesario ni oportuno programar los algoritmos de anonimización una vez que existe software que libera de esta tarea, y éste es el enfoque adoptado en esta Nota.

Lo cual lleva a otra reflexión, a preguntarse cuáles son los perfiles profesionales que debemos proponer desde nuestras universidades. Probablemente no todos ellos necesiten conocer los entresijos de las herramientas que utilizan, sean éstas de anonimización o cualquier otro fin. Sin embargo, sí será necesario que sigan existiendo perfiles capaces de entender los algoritmos subyacentes, y en consecuencia, sus limitaciones. A nadie se le escapa que los avances en Inteligencia Artificial, unidos a la disponibilidad cada vez mayor de datos y las posiblidades para su integración, pondrán a prueba la capacidad de los actuales algoritmos para resistir ataques de desanonimización. Anticiparse requiere entender el origen de los riesgos y ser capaz de proponer alternativas a las soluciones actuales.

REFERENCIAS

Agencia Española de Protección de Datos, AEPD. 2019. Unidad de Evaluación y Estudios Tecnológicos. 2019. *La K-anonimidad como medida de la privacidad.* 2019. Accesible en https://www.aepd.es/media/notas-tecnicas/nota-tecnica-kanonimidad.pdf (última visita: 22/07/2019)

Agencia Nacional de Evaluación de la Calidad y Acreditación, ANECA. 2004. *Libro Blanco del Título de Grado en Ingeniería Informática.* Marzo 2004. Accesible en www.aneca.es/Documentos-y-publicaciones/Libros-Blancos (última visita: 22/07/2019)

Association for Computing Machinery (ACM), IEEE Computer Society (IEEE-CS). 2017. Curriculum Guidelines for Post-Secondary Degree Programs in Cybersecurity, Dec. 2017. Accesible en https://www.acm.org/education/curricula-recommendations (última visita: 20/07/2019)

Article 29 Data Protection Working Party. 2014. *Opinion 05/2014 on Anonymisation Techniques. Adopted on 10 April 2014*. Accesible en https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (última visita: 22/07/2019)

LeFevre, K., DeWitt, D. J., and Ramakrishnan, R. 2005. *Incognito: Efficient full-domain k-anonymity*. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Baltimore, Maryland, USA, June 14-16, 2005, F. Özcan, Ed., ACM, pp. 49–60.

Molino, Sergio del. 2016. *La España vacía: Viaje por un país que nunca fue*. Turner ed.

Unión Europea. 2016. *Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento General de Protección de Datos)*. DOUE, 119, 4 de mayo. ELI: http://data.europa.eu/eli/reg/2016/679/oj

Unión Europea. 2018. *Reglamento (UE) 2018/1807 del Parlamento Europeo y del Consejo, de 14 de noviembre de 2018, relativo a un marco para la libre circulación de datos no personales en la Unión Europea*. DOUE, L 303, 28 de noviembre de 2018. ELI: http://data.europa.eu/eli/reg/2018/1807/oj

(1) M. Mercedes Martínez González es Profesora Contratada Doctora en el departamento de Informática de la Universidad de Valladolid. Imparte docencia de las asignaturas *Tecnología y Diseño de Bases de Datos* y *Análisis y Diseño de Algoritmos* en el *Grado de Ingeniería Informática*, y de la asignatura *Sistemas Avanzados para la Integración de Información* en el *doble Grado en Ingeniería Informática y Grado en Estadística (INdat)* de la Universidad de Valladolid. Es organizadora de la jornada *LexDatum: Derecho para profesionales TIC*, la cual se realiza con carácter anual desde el año 2015 en la Universidad de Valladolid. Forma parte del equipo de recensores de la revista *Computing Reviews* de la ACM.