LA K-ANONIMIDAD COMO MEDIDA DE LA PRIVACIDAD

I. RESUMEN

Esta nota técnica se dirige a responsables y encargados de tratamiento que aborden procesos de anonimización sobre conjuntos de datos. En una realidad en la que fuentes de datos independientes se interconectan y que, por diseño, pueden compartir atributos comunes, cabe la posibilidad de crear un rastro electrónico de los individuos, incluso cuando se hayan suprimido los datos que explícitamente les identifican, pudiendo llegar a establecerse vínculos entre dichas fuentes de información y constituir así una amenaza para la privacidad de los interesados cuyos datos están sujetos a tratamiento.

En aplicación del principio de Responsabilidad Proactiva establecido en el Reglamento (UE) 2016/679 General de Protección de Datos, el responsable debe abordar el estudio del riesgo inherente de reidentificación de los sujetos de los datos e implementar las medidas para gestionarlo. El objetivo de dicho análisis es alcanzar un balance correcto entre la necesidad de obtener unos resultados con una determinada fidelidad y el coste que el tratamiento puede tener para los derechos y libertades de los ciudadanos.

En la presente nota se desarrolla una de las posibles técnicas para gestionar el riesgo de reidentificación conocida como k-anonimización.

II. INTRODUCCIÓN

La Directiva 95/46 en su considerando 26 establecía que, para determinar si una persona era identificable, era necesario considerar el conjunto de los medios que pudieran ser razonablemente utilizados por el responsable del tratamiento, o por cualquier otro, para identificar a dicha persona. De esta forma, dejaban de ser aplicables los principios de protección de datos en aquellos casos en los que el conjunto de datos fuera hecho "anónimo" o disociados de manera tal que ya no fuera posible identificar al interesado.

En la misma línea, el considerando 26 del RGPD señala que datos personales "seudonimizados" constituyen información sobre una persona física a partir de la cual es posible realizar su identificación dentro de una probabilidad razonable teniendo en cuenta medios y factores objetivos, así como los costes, el tiempo y la tecnología necesarios para materializar su identificación.

La diferencia del término utilizado en ambas normas ha evolucionado desde una limitada "anonimización" a una materialización de ésta en el término "seudonimización" del RGPD donde viene a ponerse de manifiesto la dificultad de conseguir, en la actualidad, una anonimización perfecta o que garantice, en términos absolutos, el enmascaramiento de la identidad de las personas. No obstante, a lo largo del presente documento utilizaremos el término anonimización con independencia de que la identificación del sujeto de datos sea o no reversible en mayor o menor grado.

El tratamiento masivo de datos procedentes de los ciudadanos mediante el uso de técnicas basadas en Big Data, Inteligencia Artificial o Machine Learning obliga a la

implementación de garantías o mecanismos para preservar la privacidad y el derecho a la protección de datos de carácter personal, entre ellas las basadas en la anonimización de dichos datos.

Las fuentes de datos empleadas para dichos tratamientos contienen datos personales que se catalogan como *"identificadores"* ya que, por sí solos, están asociados de forma unívoca a un sujeto, como son el DNI, el nombre completo, el pasaporte o el número de la seguridad social. El proceso básico de anonimización consiste en disociar de los identificadores el resto de los datos más genéricos asociados a un sujeto como la fecha de nacimiento, el municipio de residencia, el género, etc. El conjunto de datos preservados serán aquellos necesarios para cumplir con el objetivo del tratamiento y, mediante la su conservación y enriquecimiento, explotarlo para extraer información adicional.

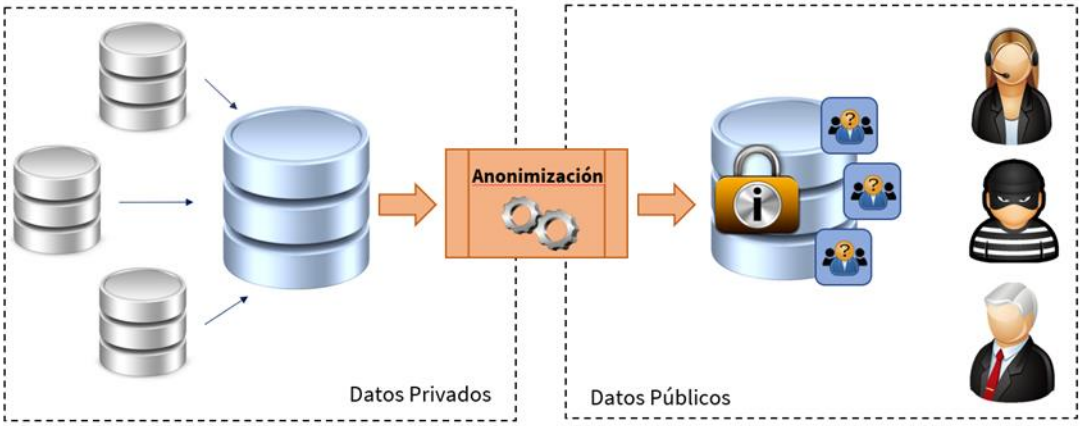

Figura 1: Anonimización

Sin embargo, aunque la realización de dicho proceso de anonimización aparentemente permite mantener el anonimato, dichos datos convenientemente agrupados y cruzados con otras fuentes de información, pueden llegar a identificar a un individuo e incluso relacionarlo con categorías especiales de datos. De ahí que al conjunto de datos que no son "identificadores" pero podrían llegar a señalar de forma univoca a un individuo se le denomina "*pseudo-identificadores*", *"cuasi-identificadores"* [1], o identificadores indirectos.

Existe un riesgo de que, una vez que se ha anonimizado un conjunto de datos, se pueda producir una desanonimización de éstos. Por lo tanto, es necesario tener una estimación objetiva de cuál es la probabilidad de reidentificación a partir del conjunto de cuasi-identificadores y, de esa forma, tener una medida de dicho riesgo.

Para gestionar este problema y evitar la desanonimización de un conjunto de datos se ha desarrollado una disciplina conocida como Control de Revelación Estadística o técnicas SDC (Statistical Disclosure Control)[2], cuyo objeto es estudiar la forma de realizar un tratamiento adicional sobre la información de los sujetos de datos de manera óptima, maximizando la privacidad al mismo tiempo que se mantiene los objetivos establecidos en la aplicación o servicio que explota dichos datos. Las técnicas utilizadas en SDC pueden ser claificadas genéricamente como perturbativas o no perturbativas, en función de si se introduce ruido en la fuente de datos original o no.

Una de estas técnicas es la K- anonimización, técnica que ya señalaba el grupo de trabajo del artículo 29 de la Directiva 45/96 en su Opinión 05/2014 [3].

III. ¿QUÉ ES LA *K*-ANONIMIDAD?

La ***K*-anonimidad** es una propiedad de los datos anonimizados que permite cuantificar hasta qué punto se preserva la anonimidad de los sujetos presentes en un conjunto de datos en el que se han eliminado los identificadores. Dicho de otro modo, es una medida del riesgo de que agentes externos puedan obtener información de carácter personal a partir de datos anonimizados.

Si clasificamos los atributos de los registros según su naturaleza o tipo de información que contienen distinguimos los siguientes tipos de datos [4]:

- **Atributos clave o identificadores**: son campos que identifican unívocamente a los sujetos de los datos (nombre, DNI, nº de pasaporte, teléfono, …). Este tipo de datos deben eliminarse de los registros anonimizados.
- **Cuasi-identificadores**: son campos que, si bien por si mismos y de forma aislada no identifican a un individuo, agrupados con otros atributos *cuasi-identificadores* pueden señalar de forma unívoca a un sujeto. Las técnicas de anonimización trabajan sobre estos datos, eliminando campos que no son necesarios para el tratamiento (en aplicación del principio de minimización), agregándolos o generalizándolos.
- **Atributos sensibles**: son los campos que contienen datos que podrían tener un mayor impacto en la privacidad de un individuo concreto, entre ellos las categorías especiales de datos, y que no deben ser vinculados con el sujeto de datos al que pertenecen (enfermedades, tratamientos médicos, nivel de renta, …). Esta información puede ser de gran interés en el objeto del tratamiento de datos, pero a menos que exista una legitimación para ello, debe mantenerse disociada de un sujeto concreto.

Se dice que un individuo es k-anónimo dentro del conjunto de datos en el que se encuentra incluido si, y sólo si, para cualquier combinación de los atributos cuasi-identificadores asociados, existen al menos otros $K-1$ individuos que comparten con él los mismos valores para esos mismos atributos [5].Hay que tener en cuenta que la K-anonimidad no se centra en los atributos sensibles de los registros [4], sino en los atributos cuasi-identificadores que pueden permitir la vinculación.

De este modo, la probabilidad de identificar a un individuo concreto en base a ese conjunto de cuasi-identificadores es como máximo $1/K$, por lo que para garantizar un bajo riesgo de reidentificación debe garantizarse un valor mínimo de K cuando se pretende llevar a cabo el diseño de un proceso de anonimización o disociación de datos.

Por ejemplo, imaginemos los siguientes conjuntos de datos en los que existen dos atributos de tipo cuasi-identificador como el “código postal” y la “edad” asociados a un atributo de tipo sensible que detalla datos de salud relativos a los sujetos de datos contenidos en el conjunto.

Código postal	Edad	Colesterol
37003	40	S
28108	44	S
24700	37	N
24700	37	N
37003	40	S

Código postal	Edad	Colesterol
37003	40	S
28108	44	S
24700	37	N
24700	37	N
37003	44	S

28108	44	S

Tabla 1: 2-anonimización

28108	40	S

Tabla 2: 1-anonimización

La tabla 1 esta 2-anonimizada, puesto que cada combinación de valores de los atributos cuasi-identificadores aparece al menos en dos filas, mientras que la tabla 2 no lo está pues no existe para cada registro al menos otro que contenga valores idénticos para dichos atributos.

Por lo tanto, se derivan dos conclusiones en relación con los valores de *K* en un conjunto de datos anonimizado:

1. Interesan valores altos de K para que, encontrado un sujeto incluido en varias fuentes de información y al que se le asocian determinados atributos, sea improbable saber a cuál de ellos se corresponde exactamente otro dato de interés asociado, por ejemplo, un tratamiento médico.
2. La 1-anonimidad equivale a decir que el individuo es perfectamente identificable dentro de su grupo [6]. Por lo tanto, ante las circunstancias oportunas y cruzando adecuadamente la información de otras fuentes en las que figuren datos de dicho individuo, podría ser posible desanonimizar la identidad de determinados sujetos de los incluidos en el universo de estudio.

A la hora de diseñar un tratamiento en el que se requiera hacer uso de datos anonimizados es importante responder a las siguientes cuestiones:

- ¿Qué valor de K es adecuado?

  Mayores valores de K se corresponden con requisitos de privacidad más exigentes dado que será necesaria la existencia de más sujetos dentro de un grupo que satisfagan idéntica combinación de rasgos identificativos. En la obtención de mayores valores de K se puede perder fidelidad en los datos de origen, por lo que hay que determinar si en esa pérdida de fidelidad hay o no pérdida de información que sea relevante para la finalidad del tratamiento. Si no hay pérdida de información relevante, hay que ejecutar ese proceso inicial. Si hay pérdida de información relevante, habrá que conseguir alcanzar el equilibrio entre los riesgos para los derechos y libertades de los sujetos y la potencial pérdida de fidelidad en el resultado del tratamiento.

- ¿Cómo conseguimos hacer un conjunto de datos K-anónimo?

  A responder esta cuestión está dedicado el apartado siguiente.

IV. MÉTODOS DE K-ANONIMIZACIÓN

Existen dos métodos ampliamente utilizados para implementar la K-anonimización y que no introducen perturbación en los datos: la generalización y la eliminación. Se dice que estos métodos son no perturbativos porque logran la protección mediante la sustitución de los valores originales de los atributos por otros valores, más generales, sin introducir información errónea en la fuente de datos original.

Generalización

La generalización consiste en hacer que el valor de los atributos cuasi-identificadores sea menos preciso, transformándolos o generalizándolos dentro de un conjunto o

intervalo que comparte los mismos valores, bien mediante la creación de rangos en el caso de atributos numéricos o el establecimiento de jerarquías para los atributos nominales. De este modo, el número de registros que poseen los mismos valores para un conjunto de atributos cuasi-identificadores se puede incrementar con el objeto de satisfacer los requisitos de privacidad a la vez que sigue siendo posible cumplir con la finalidad del tratamiento.

Partiendo de la tabla 2 mostrada anteriormente, es posible transformarla en un conjunto de datos 2-anónimo realizando una generalización del atributo 'Edad' dentro de un rango numérico y del atributo 'Código postal' clasificado en una jerarquía (figura 2). A su vez, la generalización puede ser global, si dado un mismo valor para un mismo tipo de atributo siempre se realiza la transformación de la misma manera (tabla 3), o local si se utilizan criterios de generalización diferentes para cada registro (tabla 4).

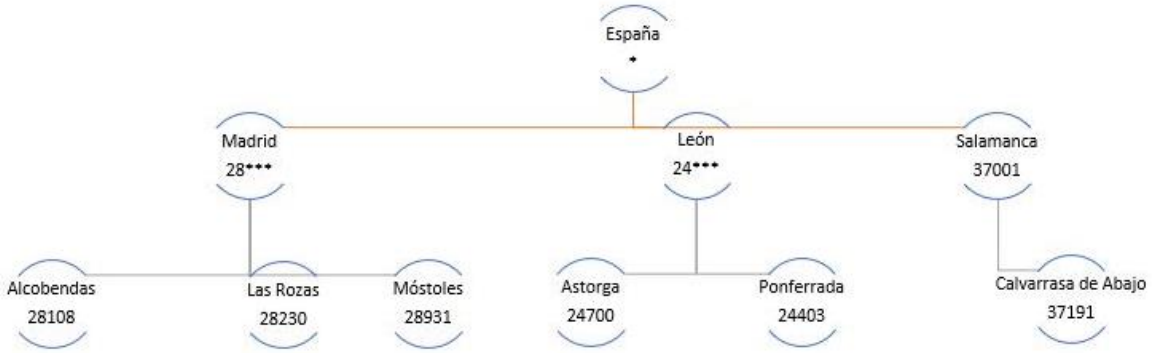

Figura 2: Jerarquía para el campo Código Postal

Código postal	Edad	Colesterol
37***	40 - 49	S
28***	40 - 49	S
24***	30 - 39	N
24***	30 - 39	N
37***	40 - 49	S
28***	40 - 49	S

Tabla 3 - Generalización global

Código postal	Edad	Colesterol
37***	40 - 49	S
28***	40 - 49	S
24700	30 - 39	N
24700	30 - 39	N
37***	40 - 49	S
28***	40 - 49	S

Tabla 4 - Generalización local

La ventaja de la generalización global es que simplifica el análisis de los datos mientras que la generalización local, si bien permite mantener valores más precisos, complica la representación de los resultados.

Eliminación

El otro método para implementar la K-anonimidad es la eliminación. En el ejemplo anterior los valores de los registros estaban bastante próximos entre sí, lo que permitía generalizar manteniendo una precisión razonable. Imaginemos que a la tabla 2 se le añaden los siguientes registros:

Código postal	Edad	Colesterol
37003	40	S
28108	44	S
24700	37	N
24700	37	N
37003	44	S
28108	40	S
37891	**33**	**N**
50011	**13**	**S**

Tabla 5: Tabla 2 expandida con valores fuera de rango

Para los seis primeros registros podemos hacer una generalización global o local tal y como se ha mostrado en las tablas 3 y 4, pero el último de los registros añadidos está fuera de rango. Intentar realizar una generalización definiendo un intervalo que lo contenga podría conllevar una pérdida de precisión tal que los datos dejarían de ser útiles para un análisis.

En estos casos, la solución pasa por suprimir o eliminar ese tipo de registros de modo que no "contaminen" el conjunto de datos y distorsionen los resultados. También los registros con valores muy poco usuales deben ser eliminados dado que aumentan significativamente la probabilidad de reidentificación.

Aplicando ambos métodos, generalización y eliminación, la tabla 5 del segundo ejemplo llegaría a ser 2-anónima tal y como se muestra en la tabla 6:

Código postal	Edad	Colesterol
37003	40	S
28108	44	S
24700	37	N
24700	37	N
37003	44	S
28108	40	S
37891	33	N
50011	13	S

Tabla 5: Original

Código postal	Edad	Colesterol
37***	40 - 49	S
28***	40 - 49	S
24700	30 - 39	N
24700	30 - 39	N
37***	40 - 49	S
28***	40 - 49	S
37***	30 - 39	N

Tabla 6: Generalización + Eliminación sobre la Tabla 5

Al intentar anonimizar utilizando de forma aislada el método de eliminación o combinado con el método de generalización se obtienen conjuntos de datos que contienen menos registros que en la fuente de datos original.

V. LIMITACIONES DE LA K-ANONIMIZACIÓN

Generalización y eliminación introducen distintos tipos y grados de distorsión en el proceso de anonimización. Anonimizar basándose en técnicas de eliminación puede suponer tener que eliminar un número considerable de registros del conjunto de datos tratados, introduciendo un sesgo en la distribución original de los valores que puede llegar a distorsionar el resultado de los análisis. Por su parte, la generalización hace que se desaproveche el potencial informativo de los datos atómicos haciendo que en el conjunto se pierda la capacidad de extraer conclusiones del valor de dichos atributos en su relación con otros campos de información. Si bien en el ejemplo mostrado el sesgo introducido es importante al tratarse de un número muy limitado de entradas, en el caso de fuentes de datos con un elevado número de registros, la pérdida de unos cuantos valores dispersos no distorsiona demasiado el resultado y evita introducir rangos de generalización amplios para contener esos extremos.

El problema matemático que hay detrás de transformar un conjunto de datos en otro conjunto de datos K-anónimo es un problema de complejidad NP-duro [7]. Existen diferentes algoritmos [8] , [9] para alcanzar una solución y sobre los que se construyen diferentes soluciones software, tanto abiertas como comerciales, que permiten K-anonimizar el conjunto de datos que se les introduce como entrada. Algunos ejemplos de este tipo de herramientas que permiten implementar las técnicas de K-anonimidad son [10]:

- **ARX Data Anonymization Tool**: ARX es una herramienta de código abierto que permite transformar conjuntos de datos personales estructurados utilizando diferentes métodos de anonimización y técnicas SDC. Permite eliminar los atributos identificadores directos (por ejemplo, nombres) de los conjuntos de datos y para aplicar reglas a los cuasi-identificadores para minimizar los ataques de vinculación. La herramienta soporta varias técnicas de privacidad, entre ellas la k-anonimidad, así como modelos de transformación de los datos como el muestreo aleatorio o la microagregación. ARX es capaz de manejar grandes conjuntos de datos y cuenta con una interfaz gráfica multiplataforma intuitiva además de una API de integración con Java para implementar capacidades de anonimización de datos desde software desarrollado bajo este lenguaje de programación.

  Enlace de descarga: https://arx.deidentifier.org/downloads/

- **Herramienta de anonimización UTD:** Es una herramienta de código abierto desarrollada en el UT Dallas Data Security y Privacy Lab, que implementa varios métodos de anonimización para uso público por parte de investigadores. Los algoritmos se pueden usar tanto directamente contra un dataset o conjunto de datos como a través de librerías de funciones implementadas dentro de otras aplicaciones. Utiliza métodos de anonimización diferentes, entre ellos la k-anonimidad.

  Enlace de descarga: http://cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php?go=download

- **Amnesia:** Amnesia es una herramienta de anonimización de datos, que permite eliminar la información no sólo asociada a los identificadores directos como nombres o números de documentos identificativos, sino que también transforma los atributos cuasi-identificadores como la fecha de nacimiento y el código postal para mitigar los riesgos de reidentificación de los sujetos que figuran en las fuentes de datos, utilizando para ello métodos de k-anonimato. Dispone de una versión cliente y de una versión online.

  Enlace de descarga: https://amnesia.openaire.eu/installation.html
  Enlace versión online: https://amnesia.openaire.eu/amnesia/

Sin embargo, aunque el K-anonimato impide desvelar la identidad de un sujeto de datos concreto dentro de un conjunto de individuos que compartan los mismos valores para los atributos cuasi-identificadores, aún puede fracasar en la protección de la revelación de información sensible asociada a este sujeto, pues en el caso de que los K elementos de una clase de equivalencia compartan un mismo valor para un atributo considerado confidencial, como ocurre en el ejemplo visto en esta nota, la simple determinación de la pertenencia de un individuo al grupo K-anonimizado hará que, sin saber su identidad exacta, se le asocie con total certeza el valor sensible protegido o con un porcentaje muy alto de acierto. En nuestro ejemplo, si conseguimos averiguar que un individuo residente en Madrid en el rango de 40 - 49 años pertenece a la muestra de estudio, sabremos que sufre de colesterol.

Este tipo de vulnerabilidades han motivado la aparición de técnicas de privacidad adicionales que están fuera del alcance de esta nota técnica, como la K-anonimidad p-sensible y la l-diversidad, que miden el grado de diversidad o variedad de los valores para los datos sensibles dentro de una clase de equivalencia, y la t-proximidad y la δ-revelación que miden la similitud entre la distribución de los valores de los atributos sensibles en cada clase de equivalencia y la distribución global de todos los registros. La herramienta ARX antes descrita implementa, además de la técnica de K-anonimidad, algunas de estas otras técnicas enumeradas dirigidas a mitigar los ataques de vinculación entre conjuntos de datos.

VI. CONCLUSIONES

El deber del responsable del tratamiento es velar por la privacidad de los sujetos de los que trata datos. Algunas entidades consideran que suprimir o enmascarar los atributos de carácter identificador resulta suficiente para garantizar la anonimidad de los sujetos objeto de estudio, sin embargo, es posible que campos comunes presentes en diferentes fuentes de datos, convenientemente agrupados y cruzados, se conviertan en un atributo seudoidentificador que llegue a comprometer la privacidad de las personas.

Por lo tanto, la anonimización no puede limitarse a la simple aplicación rutinaria y pasiva de determinadas reglas de uso común si no que, en aplicación del principio de *accountability*, el responsable del tratamiento debe analizar los riesgos de reidentificación en sus procesos de anonimización, escogiendo adecuadamente el tipo de atributos cuasi-identificadores utilizados con el objetivo de reducir la probabilidad de que el cruce de dichos campos con otros contenidos en fuentes de datos externas pueda representar un riesgo para los derechos y libertades de los individuos sujetos de su tratamiento.

Para ello, durante las fases de concepción y diseño de un tratamiento de datos de carácter personal, se ha de realizar un análisis del grado de fidelidad necesario en el resultado del tratamiento para determinar, de forma precisa, los márgenes adecuados de generalización y eliminación, dentro de límites razonables que impidan la distorsión de la realidad.

Igualmente, hay que hacer un análisis y correcto balance entre los riesgos para los derechos y libertades de los ciudadanos y los beneficios legítimos y para la sociedad que conlleva la realización de dicho tratamiento con un determinado grado de precisión.

Derivado de ambos análisis, es preciso alcanzar un equilibrio entre el beneficio que se obtendrá para la sociedad en la realización de un tratamiento con un grado de fidelidad determinado y el coste que dicho tratamiento implica para los derechos y libertades de los sujetos de los datos.

Existen diferentes técnicas orientadas a preservar la privacidad de los datos personales de individuos encaminadas a limitar las amenazas a la privacidad que pueden materializarse al desanonimizar información. La K-anonimidad es una técnica orientada a prevenir la reidentificación de un sujeto concreto dentro de un grupo, ya sea mediante generalización de los atributos cuasi-identificadores o la eliminación de registros fuera de rango, sin embargo, no ofrece garantías para evitar que sea posible, conocida la pertenencia de un sujeto a dicho grupo, inferir información de carácter sensible que tenga asociada.

VII. REFERENCIAS

[1] Jordi Casas Rom. *Data privacy*, Universitat Oberta de Catalunya (UOC) Data Day, 2017.

[2] Agustín Solanas, Antoni Martínez-Ballesté, Josep Domingo-Ferrer, Susana Bujalande, Josep M. Mateo-Sanz. *Métodos de microagregación para k-anonimato: privacidad en bases de datos*, Dept. Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili.

[3] Opinion 05/2014 on Anonymisation Techniques, adopted on 10 April 2014, 29WP.

[4] R. Somolinos Cristóbal, A. Muñoz Carrero, M.E. Hernando Pérez, M. Pascual Carrasco, R. Sánchez de Madariaga, O. Moreno Gil, J.A. Fragua Méndez, F. López Rodríguez, C. H. Salvador. *Pseudonimización de información clínica para uso secundario. Aplicación en un caso práctico ISO/EN 13606*, Unidad de Investigación en Telemedicina y e-Salud, Instituto Carlos III, Madrid, 2014.

[5] Latanya Sweeney. *K-anonymity: A model for protecting privacy*, School of Computer Science, Carnagie Mellon University, 2002.

[6] Carlos J. Gil Bellosta. *Microdatos y K-anonimidad: un enfoque cuantitativo en el contexto español*, Dananalytics, 2011.

[7] Adam Meyerson, Ryan Williams. *On the complexity of Optimal K-Anonymity*, Computer Science Departments of University of California and Carnagie Mellon University

[8] Aris Gkoulalas-Divanis, Grigorios Loukides, Jimeng Sun. *Publishing data from electronic health records while preserving privacy: A survey of algorithms*, Journal of Biomedical Informatics - Elsevier, 2014.

[9] Zakariae El Ouazzani, Hanan El Bakkali. *A new technique ensuring privacy in big data: K-anonimity without prior value of the threshold K*, ScienceDirect – Elsevier, 2018.

[10] Relación de software comercial y de código abierto relacionado con técnicas de anonimización (https://arx.deidentifier.org/overview/related-software/)