Privacidad en la publicación de datos

PID_00270338

Alexandre Viejo

Tiempo mínimo de dedicación recomendado: 5 horas

Universitat Oberta de Catalunya

Alexandre Viejo

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por la profesora: Mònica Vilasau Solana (2020)

Primera edición: febrero 2020
© Alexandre Viejo
Todos los derechos reservados
© de esta edición, FUOC, 2020
Av. Tibidabo, 39-43, 08035 Barcelona
Realización editorial: FUOC

Ninguna parte de esta publicación, incluido el diseño general y la cubierta, puede ser copiada, reproducida, almacenada o transmitida de ninguna forma, ni por ningún medio, sea este eléctrico, químico, mecánico, óptico, grabación, fotocopia, o cualquier otro, sin la previa autorización escrita de los titulares de los derechos.

Índice

Introducción..... 5

Objetivos..... 7

1. La publicación de datos personales..... 9
1.1. Tipos de datos 10
1.2. Registros de microdatos 11

2. Privacidad en la publicación de datos: anonimato y confidencialidad..... 14
2.1. Datos anonimizados y datos seudonimizados 15
2.2. El problema a resolver para anonimizar datos 17

3. Riesgos de revelación de información confidencial..... 19

4. Métodos de anonimización para microdatos..... 21
4.1. Métodos de enmascaramiento sin perturbación de datos 21
4.2. Métodos de enmascaramiento con perturbación de datos 24
4.3. Generación de datos sintéticos 26
4.4. Valoración general de los métodos de anonimización presentados 27

5. Preservación de la utilidad de los datos anonimizados..... 28

6. Modelos de privacidad para microdatos..... 31
6.1. k-Anonimidad 31
6.2. l-Diversidad 36
6.3. t-Proximidad 39

7. Anonimización de otros tipos de datos..... 40
7.1. Datos transaccionales 41
7.2. Datos de movimiento de objetos 43
7.3. Datos textuales 46

8. Integración de la privacidad en el desarrollo de productos tecnológicos..... 49
8.1. Principios relativos al tratamiento de los datos personales 49
8.2. Protección de datos desde el diseño y por defecto 50
8.3. Privacy enhancing technologies (PET) 51
8.4. Evaluación de impacto en la protección de datos (EIPD) 54

Resumen..... 58

Bibliografía..... 61

Introducción

En el contexto actual de la sociedad de la información es, generalmente, sencillo encontrar información sobre nuestra identidad, hábitos, intereses u opiniones; publicada en diversas y heterogéneas fuentes electrónicas como bases de datos gubernamentales, plataformas comerciales o redes sociales. Esta situación es especialmente sensible dado que la cantidad de información sobre nosotros que queda publicada al alcance de muchos tiende a aumentar con el tiempo y permanece disponible durante largos períodos de tiempo. Como resultado, los datos que se pueden recopilar sobre nosotros crecen y se vuelven más detallados con el paso del tiempo.

Lectura recomendada

Sánchez, D.; Viejo, A. (2017). «Personalized privacy in open data sharing scenarios». *Online Information Review* (vol. 41, núm. 3).

La recolección y explotación de datos personales se ha convertido en un negocio lucrativo que ha visto florecer la industria de los *data brokers*, entidades que compilan y analizan la información de los consumidores para revenderla o proporcionar servicios comerciales. Este tipo de negocio reportó ingresos anuales de más de mil millones de dólares en el 2014.

Lectura recomendada

US Federal Trade Commission (2014). Data brokers, a call for transparency and accountability.

En paralelo al imparable incremento de la publicación de datos personales en la red, la sociedad en conjunto ha experimentado un aumento significativo de la comprensión de las amenazas a la privacidad que la recolección y explotación incontroladas de estos datos personales pueden producir. Estas amenazas incluyen acciones discriminatorias, explotación no ética de datos, ataques informáticos de tipo *phishing* y, en definitiva, cualquier actividad lesiva para la dignidad y los derechos de los afectados. Uno de los frutos más significativos de esta preocupación global respecto a la privacidad de datos es el Reglamento general de protección de datos (RGDP): la regulación europea relativa a la protección de las personas físicas en lo que respecta al tratamiento de sus datos personales y a la libre circulación de estos.

Lectura recomendada

Parlamento Europeo (2016). *Reglamento (UE) 2016/679, general de protección de datos (RGPD)* [en línea]. <https://www.boe.es/doue/2016/119/L00001-00088.pdf>

Cabe destacar que el derecho a la privacidad no es un concepto nuevo, en 1890 el artículo «The Right To Privacy» de Warren and Brandeis presentaba la privacidad como una parte del derecho fundamental a la vida. No obstante, los avances tecnológicos actuales, incluyendo los ordenadores, la conectividad mediante internet, y el software capaz de recolectar y procesar grandes cantidades de datos, han puesto este derecho bajo el foco, convirtiéndolo en un derecho fundamental en sí mismo, presente en las constituciones de 150 países.

Lectura recomendada

Warren y Brandeis (1890). «The Right to Privacy». *Harvard Law Review* (vol. 193).

En lo que respecta a la protección de los datos de las personas, la legislación sobre privacidad se basa en varios principios, como limitar la recopilación de datos, especificar el propósito de dicha recolección, limitar el uso de los datos recolectados, proporcionar el control al individuo sobre la gestión y uso de sus datos, etc. No obstante, con el auge de la industria del *big data*, y su capacidad para extraer, analizar y monetizar toda pieza de información presente en la red, es claramente dudoso que esos principios en los cuales se basa la protección de datos se puedan aplicar de forma efectiva.

Lectura recomendada

Soria-Comas, J.; Domingo-Ferrer, J. (2015). «Big data privacy: challenges to privacy principles and models». *Data Science and Engineering* (pág. 1-8).

Entre todos los aspectos relacionados con la privacidad de la información, esta documentación se centra en la privacidad en la publicación o difusión de datos, un ámbito que ha recibido mucha atención por parte de la comunidad científica. En este sentido, la difusión de datos es la tarea principal de los institutos nacionales de estadística, los cuales pretenden ofrecer una imagen precisa de la sociedad; con ese fin, recopilan y publican datos estadísticos en una amplia gama de aspectos como economía, población, etc. En este ámbito, la legislación generalmente asocia la violación de la privacidad con la identificación de individuos en los datos difundidos.

Lectura recomendada

Domingo-Ferrer, J.; Sánchez, D.; Soria-Comas, J. (2016). *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*. Morgan & Claypool Publishers.

Objetivos

En los materiales didácticos asociados a este módulo el estudiante encontrará los contenidos necesarios para alcanzar los siguientes objetivos:

1. Conocer los tipos de datos habituales que se utilizan al publicar información en entornos abiertos como Internet.

2. Conocer los dos objetivos principales a la hora de proporcionar privacidad en la publicación de datos.

3. Entender la diferencia entre datos anonimizados y datos seudonimizados.

4. Conocer el problema que hay que resolver para anonimizar un conjunto de datos.

5. Entender qué es el riesgo de revelación de identidad y qué es el riego de revelación de atributo.

6. Conocer los métodos de anonimización de datos más habituales.

7. Entender la importancia de preservar la utilidad de los datos protegidos y conocer cómo los diferentes métodos de anonimización la reducen.

8. Conocer el modelo de privacidad *k*-Anonimidad y sus extensiones.

9. Conocer la anonimización de datos transaccionales, de movimiento de objetos y textuales, entendiendo sus diferencias respecto a los habituales microdatos.

10. Entender los conceptos de protección de datos desde la privacidad y protección de datos por defecto.

11. Conocer qué son las *privacy enhancing technologies* y su utilidad en la protección de datos personales.

12. Conocer qué son las evaluaciones de impacto en la protección de datos (EIPD) y su utilidad para cumplir con el principio de protección de datos desde el diseño.

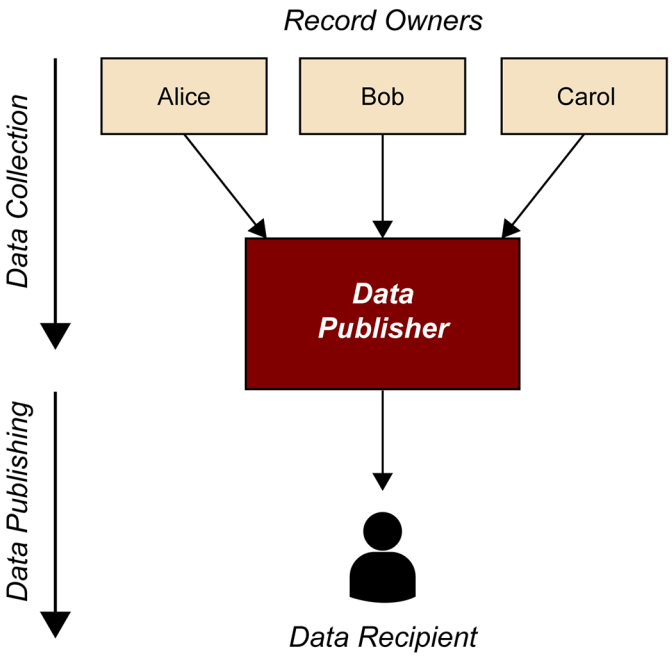
1. La publicación de datos personales

La figura 1 describe el esquema tipo de recolección y publicación de datos en un entorno abierto. En la **etapa de recolección** *(data collection)*, la entidad publicadora (*data publisher*) recolecta los datos de los individuos a los cuales pertenecen estos (*record owners*). En la **etapa de publicación** *(data publishing)*, la entidad publicadora difunde los datos recolectados a terceras personas (*data recipient*), esto es, entidades que aplicarán técnicas de minería de datos sobre dichos datos para generar nuevo conocimiento, o el público en general. Un ejemplo de este escenario sería un hospital que recoge datos de sus pacientes y publica/proporciona los registros de datos resultantes a un centro de investigación externo.

Record owners

El término *propietario de los datos* es un concepto controvertido y discutible. Esta terminología se utiliza habitualmente en la literatura sobre privacidad en el ámbito tecnológico. No obstante, en otros ámbitos, como el jurídico, este término es más cuestionable y se habla de *interesado* o *afectado* en su lugar.

Figura 1. Esquema tipo de recolección y publicación de datos



Fuente: elaboración propia (adaptado de Benjamin *et al.*, 2010)

Tal y como se indica en Benjamin, C. M. Fung; Ke Wang, RuiChen; Philip, S. Yu (2010), hay dos modelos de entidades publicadoras: 1) la entidad no confiable, la cual puede intentar identificar datos sensibles en los registros recolectados; y 2) la entidad confiable, la cual se comporta de forma honesta con los propietarios de los datos y se ciñe a la recolección y publicación de sus datos. En un escenario en el cual los datos recolectados deben ser protegidos antes de su publicación para impedir ataques a la privacidad de los individuos, la entidad publicadora de los datos está generalmente a cargo de aplicar las

Lectura recomendada

Benjamin, C. M. Fung; Ke Wang, Rui Chen; Philip, S. Yu (2010). «Privacy-Preserving Data Publishing: A Survey of Recent Developments». *ACM Computing Surveys* (vol. 42, núm. 4, art. 14).

técnicas necesarias para proporcionar dicha protección. Por esta razón, en la literatura habitualmente se asume que **la entidad publicadora es confiable**, lo cual también haremos en esta documentación.

Otra característica de la entidad publicadora es que se considera **no experta** en el uso posterior de los datos recolectados y publicados. Esto implica que los datos que difunde esta entidad no se procesan para un uso concreto por parte de los receptores finales. En lugar de eso, la entidad publicadora difunde un conjunto de registros de datos recolectados lo más general y útil posible, de forma que dicho conjunto tendrá múltiples aplicaciones diferentes, utilizables por diferentes tipos de receptores finales.

Respecto a la entidad receptora de los datos, esta se considera **no confiable**, esto es, asumimos que el receptor de los datos siempre intentará la identificación de individuos en los datos difundidos. De esta manera, la protección aplicada por la entidad publicadora a los datos recolectados tiene como función proteger a los propietarios de los datos de dichas entidades receptoras.

1.1. Tipos de datos

La tipología a la cual pertenecen los datos que se van a publicar determina las amenazas potenciales a la privacidad que estos representan, así como las técnicas de protección que habrá que aplicar sobre ellos.

Una base de datos representa una colección de datos bien organizada que permite su gestión de forma simple, añadiendo, eliminando, modificando u obteniendo la información almacenada de forma eficiente. Las **bases de datos estadísticas** son bases de datos utilizadas con fines de análisis estadístico y se consideran el vector de publicación de datos personales más habitual dado que son un elemento fundamental para la realización de estudios de todo tipo.

Lectura recomendada

Domingo-Ferrer, J.; Sánchez, D.; Soria-Comas, J. (2016). *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*. Morgan & Claypool Publishers.

Las bases de datos estadísticas se publican en dos formatos diferentes:

- Datos tabulares: Datos que han sido agregados bajo ciertas unidades colectivas y agrupaciones. Como resultado, las referencias a individuos han sido eliminadas de la base de datos. Estos tipos de datos son los habituales en las bases de datos estadísticas publicadas por institutos de estadística. Por ejemplo, una tabla en una base de datos de este tipo podría ser así:

Tabla 1.

Estado	Salario medio (dólares)
Alabama	39.832
Alaska	59.902

- Microdatos: El término *microdata* se refiere a cualquier registro que contiene información relativa a un individuo especifico (ciudadano, empresa,

etc.). Los microdatos se publican sin tratar y pueden contener información muy detallada sobre las personas o empresas (los *record owners*) a los cuales hacen referencia los datos almacenados. Siguiendo el ejemplo anterior, un conjunto de registros de microdatos podría ser así:

Tabla 2.

Nombre	Estado	Salario (dólares)
Alice	Alabama	25.000
Bob	Alaska	40.500
Carol	Alabama	51.000

De las dos tipologías indicadas, podemos concluir que los datos tabulares, por su naturaleza rígida, agregada y libre de referencias a individuos, no requieren la aplicación de técnicas para proteger la privacidad de la información contenida; por otro lado, los microdatos contienen detalles sensibles de los individuos y ofrecen el más alto grado de flexibilidad: el *data recipient* puede aplicar cualquier tipo de análisis sobre ellos dado que los datos se presentan sin ningún tipo de tratamiento ni precálculo (como la agregación que hemos visto en el caso de los datos tabulares). Es por ello por lo que la publicación de registros de microdatos es la más peligrosa para la privacidad de los individuos y es la que ha recibido más atención por parte de la comunidad científica.

1.2. Registros de microdatos

Como ya hemos visto anteriormente, un set de registros de microdatos generalmente se representa como una tabla donde cada fila corresponde a un individuo (*record owner*) y cada columna contiene información respecto a uno de los atributos de dicho individuo recolectados.

Los atributos de un registro de microdatos se pueden clasificar según las siguientes categorías:

- Identificadores: Un atributo es un identificador si proporciona una reidentificación inequívoca del individuo al que se refiere el registro. Ejemplos de atributos identificadores serían: número de DNI, número de pasaporte, o número de teléfono móvil. Si un registro contiene un identificador, cualquier información confidencial reflejada por los otros atributos puede vincularse inmediatamente a un individuo específico.
  Para evitar la reidentificación directa de un individuo por medio de un atributo identificador, estos se deben borrar de los registros antes de su publicación. Una alternativa al borrado sería cifrar estos atributos o sustituirlos por un valor que oculte el verdadero valor (un seudónimo). Cabe destacar que esta alternativa, aun con sus ventajas respecto al borrado, se

ha demostrado insegura para la privacidad de los *record owners*, y se discutirá más adelante.

- Cuasi-identificadores: Un atributo cuasi-identificador, por sí mismo y de forma aislada, no conduce a la reidentificación de un individuo en los registros. No obstante, la combinación de una serie de atributos cuasi-identificadores puede permitir la reidentificación inequívoca de algunas personas. En este sentido, el trabajo presentado en Latanya (2000) por Latanya Sweeney demostró que el 87 % de la población de los Estados Unidos puede identificarse sin ambigüedades combinando un código postal de 5 dígitos, la fecha de nacimiento y el sexo; todos ellos cuasi-identificadores aparentemente inocuos al considerarse por separado.
  A diferencia de los atributos identificadores, en este caso borrar (o cifrar) directamente todos los atributos cuasi-identificadores no es una opción viable. Esto es debido a que la mayoría de las veces se requieren cuasi-identificadores para realizar cualquier análisis útil de los datos. Más adelante explicaremos los métodos de anonimización que se pueden aplicar en este tipo de atributos para proteger la privacidad de los *record owners*.
  Cabe destacar que decidir si un cierto atributo debe ser considerado cuasi-identificador es un tema complejo: si consideramos *un atacante que no posee conocimiento adicional* sobre los individuos que hay que reidentificar, en este caso, solo aquellos atributos disponibles en un conjunto de datos externo no anonimizado (por ejemplo, una tabla con datos censales) deben clasificarse como cuasi-identificadores. Por el contrario, en presencia de *atacantes bien informados*, cualquier atributo puede ser potencialmente un cuasi-identificador; o puede no serlo, todo dependerá del grado de información adicional que posea dicho atacante.

Lectura recomendada

Latanya, S. (2000). «Uniqueness of Simple Demographics in the U.S. Population». *LIDAP-WP4*. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh PA.

- Confidenciales: Los atributos confidenciales contienen información confidencial sobre los individuos que participaron en el proceso de recopilación de datos (*record owners*). Ejemplos de información confidencial serían: salario, orientación sexual, estado de salud, etc. El objetivo principal de las técnicas de protección de la privacidad es evitar que los atacantes puedan aprender información confidencial sobre un individuo específico. Este objetivo implica no solo evitar que el intruso determine el valor exacto que toma un atributo confidencial para un cierto individuo, sino también evitar que el atacante pueda inferir de forma más o menos precisa dicho valor.

- No-confidenciales*:* Los atributos no-confidenciales son aquellos que no pertenecen a ninguna de las categorías anteriores. Este tipo de atributo no contiene información sensible sobre los individuos y no pueden utilizarse para reidentificar registros. Como resultado, estos atributos no afectan de ninguna manera al proceso de protección de la privacidad aplicado a la publicación de datos y no serán considerados en este documento.

A continuación, mostramos una tabla de ejemplo con registros de microdatos respecto a los pacientes de un hospital y las enfermedades que les han sido diagnosticadas. En esta tabla podemos ver los tres tipos de atributos considerados: identificadores, cuasi-identificadores y confidenciales.

Tabla 3.

DNI	Código postal	Edad	Sexo	Enfermedad
37713522Z	08017	37	Hombre	Hepatitis
41343621B	08242	25	Hombre	Hepatitis
36689764P	43005	45	Mujer	HIV
74452667A	08040	49	Hombre	Gripe

El campo DNI es el atributo identificador que debería ser borrado como primer paso en un proceso de anonimización de los datos. Los campos código postal, edad y sexo son atributos cuasi-identificadores, los cuales, combinados, pueden llegar a reidentificar a un paciente. Finalmente, el campo enfermedad es el atributo confidencial, el cual, en ningún caso, un atacante debería poder vincular al individuo correspondiente.

2. Privacidad en la publicación de datos: anonimato y confidencialidad

Hasta ahora hemos hablado de registros de microdatos, y de proteger la privacidad de los *record owners* que han generado dichos microdatos. Cuando hablamos de proteger la privacidad, tenemos que tener en cuenta que, en la literatura, esta protección se realiza partiendo de dos objetivos de privacidad diferentes:

- *Obtener confidencialidad* (en inglés, *confidentiality*): acceder al conjunto de datos publicado no debería revelar información confidencial vinculada a ningún individuo específico.

- *Obtener anonimato* (en inglés, *anonymity*): no debería ser posible reidentificar a ningún individuo en el conjunto de datos publicado.

Proteger datos para obtener confidencialidad busca limitar la cantidad de información que un conjunto de registros de microdatos proporciona al *data recipient*. Una forma habitual de conseguir esto es enmascarando los valores confidenciales añadiendo ruido. Por ejemplo, si sumamos/restamos números aleatorios dentro de un cierto rango, por ejemplo, entre 5 y10, (esto sería el «ruido») al campo edad de una tabla de microdatos, el *data recipient* obtendrá las edades de los individuos con una distorsión de entre 5 y 10 puntos, lo cual no le permitirá conocer las edades exactas de los *record owners*. La cantidad de ruido añadida determina directamente el nivel de confidencialidad obtenido. De esta manera, siguiendo el ejemplo, si añadimos muy poco ruido (valores entre 1 y 3), el *data recipent*, aun no obteniendo el valor exacto, sí obtendría un valor muy parecido al real; por otro lado, si añadimos mucho ruido (valores entre 10 y 20), el *data recipient* obtendría un valor tan alejado del valor real que habría que valorar si este ha dejado de ser útil. En este sentido, más adelante hablaremos sobre el equilibrio entre la protección de la privacidad y la utilidad de los datos protegidos.

Por otro lado, proteger datos para obtener anonimato busca ocultar cada individuo (cada *record owner*) en un grupo de individuos «similares», de forma que no sea posible identificar a un individuo concreto dentro de ese grupo (una idea relacionada con el «anonimato de la multitud»). Este objetivo es típicamente aplicado por la comunidad científica para desarrollar sus modelos de privacidad, que a su vez son esenciales para ofrecer garantías de privacidad al proteger conjuntos de datos. Un ejemplo de modelo de privacidad basado en este objetivo es la célebre *k-Anonimidad* (en inglés, *k-Anonymity*), modelo del cual hablaremos con detalle más adelante.

Lectura recomendada

Samarati, P.; Latanya, S. (1998). «Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression». *Technical report* . SRI International.

2.1. Datos anonimizados y datos seudonimizados

Como ya hemos dicho, un proceso basado en *anonimizar datos* busca evitar la reidentificación de individuos en los datos anonimizados. De esta manera, el primer paso en cualquier proceso de anonimización es eliminar los atributos identificadores de los registros que hay que proteger; los siguientes pasos se basarán en alterar, siguiendo ciertas técnicas, el resto de los atributos presentes en los datos. Como resultado a este proceso, obtendremos unos datos anonimizados, en los cuales **se ha roto todo vínculo entre los individuos que generaron los datos (*record owners*) y los datos resultantes protegidos**. Cabe destacar que eliminar el vínculo entre individuos y sus datos implica la destrucción de información inherente a dichos datos, lo cual, a su vez, repercute en la utilidad que se puede extraer de los datos protegidos (unos datos que proporcionan menos información forzosamente serán menos útiles para un *data recipient*).

Al explicar los atributos identificadores, indicamos que una alternativa a eliminarlos era cifrarlos o sustituirlos por otros valores que ocultasen el identificador real. En ambos casos lo que hacemos es sustituir el atributo identificador de un individuo por un seudónimo. Esta es la base de los *datos seudonimizados*. Las dos principales ventajas respecto a los datos anonimizados son: 1) **los seudónimos ocultan la identidad, pero mantienen el vínculo entre los datos** (los datos protegidos retienen más información así que son más útiles); y 2) la entidad que ha generado los seudónimos (sea mediante cifrado o por simple sustitución de valores) es capaz de revertir el proceso de seudonimización y recuperar los datos originales en caso de ser necesario.

Lectura recomendada

Barbaro M.; Zeller T. (2006). «A face is exposed for AOL searcher no. 4417749». *New York Times* (agosto).

Para poner un ejemplo representativo de las dos formas de protección, podemos asumir la base de datos de un motor de búsqueda web (como Google), donde esta entidad almacena todas las consultas que los usuarios envían (estas bases de datos se conocen como *query logs*). La primera tabla corresponde a cinco consultas reales extraídas de los *query logs* publicados por la empresa AOL en el año 2006. Estos *query logs* estaban formados por veinte millones de consultas realizadas por 657426 usuarios. Para protegerlos, AOL sustituyó los atributos identificadores de los usuarios por seudónimos en forma de número:

Tabla 4.

Seudónimo	Consulta	Fecha	URL seleccionada
2178	*Dog cornea treatment*	2006-03-27	<http://www.2ndchance.info>
2178	*Pergola house entrance*	2006-04-08	<http://www.gardenstructure.com>
2178	*Pennsylvania college savings*	2006-03-16	<http://www.patreasury.org>
2178	*Inducing dog vomiting*	2006-05-26	<http://dogs.about.com>
1326	*Holiday mansion house boat*	2006-04-06	<http://www.iboats.com>

Un observador externo que tenga acceso a esta tabla, *a priori* no puede conocer la identidad del individuo detrás del seudónimo «2178», pero sí puede hacer un perfil más o menos preciso de los intereses, problemas, lugar de residencia, etc. del individuo que se oculta detrás de dicho seudónimo. Esto puede permitir al observador saber qué páginas web acostumbra a consultar una persona con estos intereses, y quizá pueda servir para realizar algún estudio de márquetin o similar. Esta tabla basada en datos seudonimizados es capaz de proporcionar este conocimiento porque mantiene el vínculo entre los datos y el individuo que los ha generado, aun ocultando su identidad.

En contraste, eliminar directamente los atributos identificadores nos proporcionaría una tabla anonimizada de esta forma:

Tabla 5.

Consulta	Fecha	URL seleccionada
Dog cornea treatment	2006-03-27	<http://www.2ndchance.info>
Pergola house entrance	2006-04-08	<http://www.gardenstructure.com>
Pennsylvania college savings	2006-03-16	<http://www.patreasury.org>
Inducing dog vomiting	2006-05-26	<http://dogs.about.com>
Holiday mansion house boat	2006-04-06	<http://www.iboats.com>

En la cual, un observador externo no puede vincular de ninguna manera las consultas que ahí aparecen. Desde su punto de vista, cada consulta es independiente de las anteriores y no puede inferir un perfil de intereses de un individuo en concreto.

Desde el punto de vista de preservar la utilidad de los datos, seudonimizar en lugar de anonimizar representa una clara ventaja. No obstante, se ha demostrado que una protección basada en seudonimizar es débil y, de hecho, no es un método de protección de datos aceptado por la comunidad científica. En particular, siguiendo con la publicación masiva de datos que realizó AOL, en ese caso los periodistas del *New York Times*, Michael Barbaro y Tom Zeller, en su artículo explican con qué facilidad fueron capaces de rastrear un seudónimo hasta su identidad real asociada. En particular, los dos periodistas fueron capaces de asociar el número «4417749» con la persona llamada Thelma Arnold, la cual había realizado consultas como: «hombres solteros de 60», «perro que se mea en todo», «paisajes en Lilburn» y múltiples consultas relacionadas con el apellido «Arnold». Tras el escándalo, AOL eliminó el acceso al *query log* publicado; no obstante, hoy en día aún circulan copias de dichos registros.

Lectura recomendada

Barbaro M.; Zeller T. (2006). «A face is exposed for AOL searcher no. 4417749». *New York Times* (agosto).

Debido a la debilidad del uso de seudónimos, la literatura sobre privacidad generalmente se centra en la anonimización de datos, ya que su objetivo es la total disociación de los datos y es, por tanto, la metodología más segura desde el punto de vista de la privacidad. En consecuencia, en esta documentación

también nos centraremos en la anonimización de datos, y los métodos existentes para conseguirla. No obstante, cabe destacar que, a la práctica, conseguir una anonimización perfecta, entendiendo como tal una disociación irreversible y total de los datos, es una tarea de una dificultad muy elevada. Además, no es posible proporcionar garantías de que un conjunto de datos anonimizados permanecerá protegido indefinidamente, dado que la reidentificación por parte de un atacante depende en gran medida de la información externa que este posea, y la posibilidad de que un atacante obtenga nueva información adicional en el futuro que rompa la anonimización realizada a fecha de hoy está fuera del control de la entidad encargada de proteger los datos.

La imposibilidad a nivel práctico de realizar una anonimización perfecta (más allá de simplemente distorsionar/eliminar los datos protegidos hasta llegar al punto de que la utilidad de los datos resultantes sea cero) tiene ciertas implicaciones a la hora de proteger datos aplicando la legislación.

En concreto, según se indica en *La k-Anonimidad como medida de la privacidad* (2019), 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 se hubiese anonimizado de manera tal que ya no fuese posible identificar al interesado. En la misma línea, el considerando 26 del RGPD señala que los datos 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.

Lectura recomendada

Agencia Española de Protección de Datos (AEPD) (2019). *La k-Anonimidad como medida de la privacidad* [en línea]. <https://www.aepd.es/media/notas-tecnicas/nota-tecnica-kanonimidad.pdf>

La primera implicación de esta situación es que no es raro encontrar documentación en la cual se igualan los conceptos seudonimización y anonimización, dado que el hecho de que la anonimización no sea perfecta abre la puerta a que sea un proceso reversible, a semejanza de la seudonimización. La segunda implicación es que los datos anonimizados (como los seudonimizados), dada la probabilidad razonable de que su protección sea reversible y la reidentificación de individuos posible, quedarían amparados por el ámbito de aplicación de las normas sobre protección de datos (siempre y cuando exista esa probabilidad de reidentificación).

2.2. El problema a resolver para anonimizar datos

Anteriormente hemos argumentado la aplicación de la anonimización de datos (con sus imperfecciones) por parte de los expertos en el área a la hora de proteger información que va a ser publicada en entornos abiertos. En este

apartado formalizaremos el problema que esos expertos deben resolver cada vez que deben anonimizar un cierto conjunto de datos. Esto es lo que se conoce como el problema de la anonimización (en inglés, *anonymization problem*).

El **problema de la anonimización** es producir una tabla anonimizada, a partir de una tabla original, que satisfaga unos ciertos requerimientos de privacidad, determinados por el modelo de privacidad seleccionado, y que retenga la mayor cantidad de información (utilidad) posible.

Más adelante detallaremos las operaciones que se utilizan para anonimizar un conjunto de datos, introduciremos los modelos de privacidad que se suelen aplicar, y explicaremos el equilibrio requerido entre la necesidad de obtener información de la tabla protegida con una determinada fidelidad y el coste que el proceso de anonimización puede tener para la privacidad de los individuos.

3. Riesgos de revelación de información confidencial

Cuando se publican registros de microdatos, la entidad publicadora (*data publisher*) debe evitar que se divulgue información sensible o confidencial de los individuos (*record owners*) de los cuales se ha obtenido los datos. Generalmente, se consideran dos tipos de riesgos de revelación:

- Revelación de la identidad (en inglés, *identity disclosure*): Este riesgo aplica directamente a la privacidad vista como obtención del anonimato y ocurre cuando el atacante reidentifica a un individuo en el conjunto de datos publicado y protegido, esto es, el atacante es capaz de vincular un registro de dicho conjunto de datos con el individuo que lo originó. Una vez el atacante ha re-identificado a un individuo, puede asociar los valores de los atributos confidenciales a esa persona.
  Para medir el riesgo de revelación de la identidad, generalmente el *data publisher* aplica un método de vinculación entre registros (en inglés, *record linkage*) sobre los datos protegidos antes de publicarlos. Existen varios algoritmos de *record linkage* en la literatura, pero la idea básica es buscar la similaridad entre los registros protegidos y los registros originales para calcular el porcentaje de «parejas» detectadas. Detectar una pareja de registros implica que el registro original y el registro anonimizado son demasiado parecidos y, por lo tanto, un atacante podría conseguir una reidentificación. Si el porcentaje de parejas detectado es elevado, el *data publisher* debería volver a anonimizar los datos hasta conseguir un nivel de riesgo de revelación de la identidad razonable.
  Como se ha dicho, existen diferentes algoritmos de *record linkage*. El más sencillo se basaría en buscar valores de atributos coincidentes entre la tabla original y la tabla protegida, por ejemplo, podríamos buscar la combinación de los tres atributos cuasi-identificadores más célebres: código postal, fecha de nacimiento y sexo en ambas tablas, y detectar las coincidencias:

Lectura recomendada

Sobre los dos tipos de riesgos de revelación:

Hundepool, A.; Domingo-Ferrer, J.; Franconi, L.; Giessing, S.; Schulte-Nordholt, E.; Spicer, K.; De Wolf, P. P. (2012). *Statistical Disclosure Control*. Wiley.

Tabla 6. Tabla original

DNI	C. postal	Edad	Sexo	Enfermedad
37713522Z	08017	37	Hombre	Hepatitis
36689764P	43005	45	Mujer	HIV

Tabla 7. Tabla protegida

DNI	C. postal	Edad	Sexo	Enfermedad
*	08017	37	Hombre	Hepatitis
*	*	*	Mujer	HIV

En este caso, el algoritmo detectaría en la tabla protegida el registro correspondiente al usuario con DNI 37713522Z dado que los tres cuasi-identificadores coinciden. Por otro lado, el usuario con DNI 36689764P no sería detectado en la tabla protegida.

- Revelación del atributo (en inglés, *attribute disclosure*): Este riesgo aplica directamente a la privacidad vista como obtención de la confidencialidad y ocurre cuando el atacante es capaz de determinar, con suficiente precisión, el valor de un atributo confidencial a partir del conjunto de datos publicado y protegido.

Tal y como se indica en Domingo-Ferrer *et al.* (2016), los dos tipos de riesgo de revelación considerados son independientes. Esto implica que, si un atacante es capaz de reidentificar a un individuo en los datos publicados, esto no garantizaría que el atacante acabase conociendo el atributo confidencial (o atributos confidenciales) asociados a dicho registro, dado que estos se podrían haber enmascarado añadiendo ruido, por lo tanto, en este caso concreto, habría revelación de la identidad, pero no habría revelación del atributo

Lectura recomendada

Domingo-Ferrer, J.; Sánchez, D.; Soria-Comas, J. (2016). *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*. Morgan & Claypool Publishers.

Por otro lado, la revelación de atributo puede ocurrir incluso sin que exista revelación de la identidad. Por ejemplo, suponiendo una tabla que contiene los trabajadores de una empresa en la cual el salario es un atributo confidencial, y el puesto de trabajo un cuasi-identificador, si un atacante quiere conocer el salario de un cierto trabajador y sabe que su puesto de trabajo es «administrativo», aun no pudiendo reidentificar al trabajador con su registro en la tabla solamente sabiendo que es un «administrativo» (dado que habrá muchos «administrativos» en dicha tabla), sí podrá conocer el rango de salario en el que dicho trabajador se encuentra, dado que podrá conocer todos los salarios de todos los trabajadores; si el rango de posibles valores salariales para «administrativos» es suficientemente estrecho, ocurrirá la revelación del atributo.

4. Métodos de anonimización para microdatos

Para evitar los riesgos de revelación de identidad o atributo, los *data publishers* no publican los datos recolectados originales, sino que publican una versión anonimizada (o protegida) de los mismos. Para esta tarea, los *data publishers* utilizan una serie de métodos de anonimización basados tanto en enmascarar los datos originales como sustituirlos por datos sintéticos. En los siguientes subapartados clasificaremos y explicaremos los métodos de anonimización más habituales. Cabe destacar que estos métodos, además de ser herramientas de anonimización de datos, también son métodos básicos para satisfacer los requerimientos de privacidad determinados por los diversos modelos de privacidad que veremos más adelante.

4.1. Métodos de enmascaramiento sin perturbación de datos

Los métodos de enmascaramiento no perturbativos no alteran los datos originales, en lugar de eso, los suprimen o reducen su nivel de detalle. Bajo esta categoría, podemos destacar cuatro métodos:

- Generalización (en inglés, *generalization*): Este método, también conocido en la literatura como *global recoding*, sustituye los valores originales de los atributos por valores más generales, de esta manera se reduce el nivel de detalle de los datos originales y, por lo tanto, la cantidad de información que estos proporcionan. Para enmascarar un atributo siguiendo esta técnica, es necesario representar todos los valores que el atributo puede tomar siguiendo una jerarquía de generalización. En dicha jerarquía el valor más general estará situado en la cima, mientras que los valores más específicos estarán situados en la base. El proceso de generalización trabaja sustituyendo los valores específicos de los datos originales por valores más generales situados en cotas más altas de la jerarquía y que serán los que aparecerán en el conjunto de datos protegido.

  Por ejemplo, en el caso de un atributo categórico (los valores de una variable categórica son categorías o grupos mutuamente excluyentes y no admiten operaciones aritméticas) que indica el puesto de trabajo de un individuo, los valores «profesor ayudante doctor» y «profesor contratado doctor» se podrían sustituir por el valor «profesor», más general que los anteriores y que, por tanto, los engloba. En el caso de un atributo numérico (valores que admiten operaciones aritméticas) que indica el salario de un individuo, los valores originales se podrían sustituir por intervalos numéricos.

Lectura recomendada

Sobre el *global recoding*:

Samarati, P.; Latanya, S. (1998). «Protecting privacy when disclosing information: *k*-anonymityand its enforcement through generalization and suppression». *Technical report*. SRI International.

Tabla 8. Tabla original

DNI	Puesto de trabajo	Salario
37713522Z	Profesor contratado doctor	38.000
36689764P	Profesor ayudante doctor	27.000

Tabla 9. Tabla protegida

DNI	Puesto de trabajo	Salario
*	Profesor	35.000-40.000
*	Profesor	25.000-30.000

- *Top and bottom coding:* Este método es un caso especial de generalización en el cual los valores extremos, aquellos que se sitúan por encima de un determinado umbral superior y por debajo de un determinado umbral inferior, son sustituidos por un valor único que representa el valor máximo posible (*top-code*) o el valor mínimo posible (*bottom-code*) respectivamente. La idea que late detrás de este método es que los valores extremos son raros y, por lo tanto, pueden facilitar la reidentificación de los individuos a los cuales están vinculados. Siguiendo con el ejemplo anterior con la tabla de trabajadores, es raro encontrar trabajadores con edades muy elevadas en una empresa. Si la empresa tiene un único trabajador de, por ejemplo, 72 años, de nada servirá enmascarar dicha edad en una franja tipo 70-75 años mediante el método de generalización, dado que solamente ese trabajador estará en dicha franja y será fácilmente reidentificable en la tabla protegida. En lugar de eso, se aplicaría *top coding* y se sustituiría la edad por el valor «>60», bajo el cual se englobarán muchos más trabajadores de la empresa. Por definición, este método solo se puede aplicar en atributos que pueden ser ordenados de mayor a menor, como la edad o el salario.
  En las siguientes tres tablas consideramos el atributo salario como confidencial y, además, no le aplicamos ningún enmascaramiento (podríamos tener riesgo de revelación de atributo aun evitando la revelación de identidad). El atacante quiere saber el sueldo de un individuo, y sabe que este tiene más de 70 años. La tabla mal protegida no protege al individuo 76533777B: solo él está en la franja de edad 70-75 y, por lo tanto, el atacante reidentifica el registro con dicho individuo (tenemos revelación de identidad), y descubre su salario (tenemos revelación de atributo). Por el contrario, en la tabla protegida el atacante se encuentra tres registros que podrían corresponder con el individuo, de esta manera no puede reidentificarlo, y solo puede saber que su sueldo está entre 38.000 y 58.000. Si hubiésemos enmascarado el sueldo, el atacante aún hubiese conseguido menos información.

Tabla 10. Tabla original

DNI	Edad	Salario
37713522Z	26	35.000
36689764P	29	41.000
56564576X	62	44.000
15434434V	61	38.000
76533777B	72	58.000

Tabla 11. Tabla mal protegida

DNI	Edad	Salario
*	25-30	35.000
*	25-30	41.000
*	60-65	44.000
*	60-65	38.000
*	**70-75**	**58.000**

Tabla 12. Tabla protegida

DNI	Edad	Salario
*	25-30	35.000
*	25-30	41.000
*	>60	44.000
*	>60	38.000
*	>60	58.000

- Eliminación (en inglés, *local supression*): Este método se basa simplemente en eliminar ciertos valores de los atributos que se cree que pueden facilitar la reidentificación de los *record owners*. Es la técnica que aplicamos para eliminar los valores de los atributos identificadores (el primer paso en cualquier proceso de anonimización). Adicionalmente, esta técnica también se utiliza con valores de atributos cuasi-identificadores cuya existencia resulta en una combinación única en la tabla y que, por lo tanto, llevará a una reidentificación. Por ejemplo, en el caso anterior podríamos haber eliminado la edad de la persona 76533777B en lugar de sustituirlo por el *top-code* (cabe destacar que, en ese caso, tener un único registro con la edad borrada podría representar un problema de privacidad también). Finalmente, esta técnica también se puede utilizar para borrar registros enteros directamente (lo cual solucionaría el problema anterior respecto a la persona 76533777B). No obstante, es muy importante tener en cuenta que aplicar esta técnica implica borrar datos, y cada vez que se borran datos,

la tabla protegida resultante está perdiendo información y, por lo tanto, utilidad. La pérdida de utilidad afecta directamente al problema de la anonimización, que nos lleva a que el conjunto de datos protegido retenga la mayor cantidad de información (utilidad) posible.

- *Sampling:* esta técnica tiene alguna relación con la anterior en el sentido de que se basa en elegir un set de registros del conjunto de datos original y publicarlos directamente sin ningún enmascaramiento adicional (más allá de eliminar los atributos identificadores). Por ejemplo, podríamos publicar solamente los registros impares del conjunto de datos original, eliminando los pares. La protección que ofrece se sustenta en la incertidumbre respecto a si el registro de un cierto individuo se encuentra entre los datos publicados o no. De forma similar al método anterior, el hecho de eliminar registros afecta directamente a la utilidad de los datos protegidos.

Lectura recomendada

Sobre el *sampling*:

Willenborg, L.; De Waal T . (2001). *Elements of Statistical Disclosure Control*. Nueva York: Springer-Verlag.

4.2. Métodos de enmascaramiento con perturbación de datos

Este tipo de métodos se basa en distorsionar los datos originales para generar el conjunto de datos protegidos. Como hemos indicado anteriormente, los métodos no-perturbativos reducen el detalle de los datos originales o los eliminan totalmente, pero tenemos la garantía de que todos los datos que aparecen en el conjunto de datos protegidos son legítimos. En el caso de los métodos basados en perturbación, la distorsión añadida genera unos datos nuevos vinculados a cada individuo que son diferentes a los originales. Bajo esta categoría, podemos destacar tres métodos:

- Añadir ruido (en inglés, *noise addition*): Este método se usa habitualmente para enmascarar atributos numéricos considerados confidenciales (por ejemplo, el salario de una persona). La idea general es sustituir el valor original v por un valor nuevo $v+r$, donde r es un valor aleatorio (el ruido que queremos añadir) obtenido de alguna distribución de probabilidad. Según qué distribución de probabilidad elijamos para generar el ruido, obtendremos valores aleatorios más o menos altos, y con más o menos frecuencia, obteniendo así unos valores protegidos más o menos distorsionados respecto a los originales. Como resultado de añadir ruido a los datos originales, los datos protegidos resultantes serán diferentes de los legítimos, lo cual puede tener efectos negativos respecto a la utilidad de los datos; no obstante, dependiendo del tipo de ruido añadido, los datos protegidos, aun siendo diferentes de los originales, serán capaces de preservar información estadística como la media o la varianza de los datos legítimos.

Lectura recomendada

Sobre *noise addition*: **Hundepool, A.; Domingo-Ferrer, J.; Franconi, L.; Giessing, S.; Schulte-Nordholt, E.; Spicer, K.; De Wolf, P. P.** (2012). *Statistical Disclosure Control*. Wiley.

- Intercambio de datos (en inglés, *data swapping*): La idea general detrás de esta técnica es la de anonimizar una tabla de registros mediante el intercambio de los valores del atributo seleccionado entre los registros individuales. De esta forma, los valores originales se mantienen, aunque en el conjunto de datos protegido estos aparecen vinculados a otros individuos. Este método mantiene la información estadística de la tabla protegida, dado que los datos son los mismos que en la tabla original (sería exactamente lo mismo calcular la media de salarios en la tabla original que en la tabla protegida); no obstante, el intercambio de vinculación entre individuos y valores representa una pérdida de información relevante y, el hecho de que sea un intercambio aleatorio puede dar lugar a vinculaciones poco consistentes (por ejemplo, salarios que no concuerdan con puestos de trabajo). Para mitigar este problema de consistencia entre valores, existe una variación de este método llamado *rank swapping,* en el cual, primero se ordenan los valores que van a ser intercambiados, y los intercambios se producen entre valores cercanos dentro de dicho orden; de esta manera, siguiendo con el ejemplo anterior, el salario de una persona en un cierto puesto de trabajo se intercambiaría por el salario de otra persona dentro de un mismo rango de salarios, por lo tanto, el sueldo protegido asignado al individuo, aun no siendo igual al original, sí sería parecido.

Lectura recomendada

Sobre el *data swapping*:

Dalenius, T.; Reiss S. P. (1978). *Data-swapping: a technique for disclosure control, Proceedings of the ASA Section on Survey Research Methods* (págs. 191-194).

Sobre el *rank swapping*:

Greenberg, B. (1996). *Rank swapping for masking ordinal microdata*. U.S. Bureau of the Census, Washington DC.

- Microagregación (en inglés, *microaggregation*): Este método se basa en sustituir los valores originales de un atributo por un valor agregado que oculte los valores originales, pero que sea suficientemente parecido a ellos como para preservar su utilidad. El proceso de enmascaramiento tiene dos etapas: 1) el conjunto de datos original se parte en grupos de *K* registros, intentando que los valores del atributo que hay que enmascarar de dichos *K* registros sean lo más similares posibles; y 2) para cada grupo de registros, los valores del atributo a enmascarar se sustituyen por un valor representativo del grupo. Este valor representativo típicamente será la media de todos los valores originales; de esta manera, cuanto más homogéneo sea el grupo de *k* registros construido en el primer paso, más fidedigno será el valor representativo obtenido en este segundo paso.

Lectura recomendada

Sobre el *microaggregation*:

Defays, D.; Anwar, M. N. (1998). «Masking microdata using micro-aggregation». *Journal of Official Statistics* (vol. 14, n.º 4, págs. 449-461).

Tabla 13. Tabla original

DNI	Edad	Salario
37713522Z	26	35.000
36689764P	29	41.000
67457676C	65	61.000
56564576X	62	44.000
15434434V	61	38.000
76533777B	72	68.000

Tabla 14. Tabla protegida (K=2)

DNI	Edad	Salario	
*	26	36.500	(35.000+38.000)/2 = 36.500
*	61	36.500	(35.000+38.000)/2 = 36.500
*	29	42.500	(41.000+44.000)/2 = 42.500
*	62	42.500	(41.000+44.000)/2 = 42.500
*	65	64.500	(61.000+68.000)/2 = 64.500
*	72	64.500	(61.000+68.000)/2 = 64.500

Cuando se quieren enmascarar múltiples atributos, esta técnica se puede aplicar de forma iterativa para cada atributo, simplemente tomando como entrada la tabla protegida resultante de la iteración anterior; esta manera de trabajar se denomina *univariate* y produce baja pérdida de utilidad en los datos protegidos a cambio de un riesgo significativo de revelación. Otra opción es proteger todos los atributos a la vez, bajo la metodología llamada *multivariate*; en este caso, la pérdida de utilidad será muy elevada, pero el riesgo de revelación será bajo.

4.3. Generación de datos sintéticos

Este tipo de métodos se basa en eliminar todos los datos originales y sustituirlos completamente por datos nuevos generados de forma aleatoria. De esta manera se espera proteger la privacidad de los *record owners* dado que ningún dato real suyo realmente aparece en los datos protegidos. No obstante, el problema que se presenta en estos métodos es conseguir unos datos aleatorios que, dentro de su aleatoriedad, retengan suficiente información real como para que los datos protegidos resultantes tengan algún tipo de utilidad.

Lectura recomendada

Rubin, D. B. (1993). «Discussion: statistical disclosure limitation». *Journal of Official Statistics* (vol. 9, págs. 462-468).

Para conseguir retener cierta información en los datos protegidos, estos métodos lo que buscan es, primero, crear un modelo de datos (esto sería, una secuencia finita de valores posibles) del cual se pueda obtener de forma aleatoria valores similares a los valores originales que queremos «simular». Una vez tengamos ese modelo de datos, simplemente sustituiremos los valores originales por los valores que obtengamos de forma aleatoria de dicho modelo. Como todos los datos posibles de dicho modelo se supone que son «parecidos» a los originales, los datos protegidos resultantes se espera que tendrán suficiente similaridad con los originales como para retener cierta información estadística; por ejemplo, podríamos calcular la media de los valores y se espera que el resultado no difiera demasiado de la media que calcularíamos con los valores originales.

La literatura sobre privacidad considera que estos métodos solo tienen valor a nivel teórico, dado que a nivel práctico es complicado encontrar modelos de datos que sirvan para simular de forma fidedigna los valores de cualquier tipo de atributo que quiera ser protegido. Además, dado que los datos protegidos son sintéticos (esto es, falsos), el objetivo principal de estos métodos es publicar un conjunto de datos protegido que retenga cierta información estadística (media, mediana, varianza, etc.). Siendo este el objetivo principal, nos podríamos preguntar si no sería más seguro y eficiente simplemente publicar la información estadística que queremos retener directamente calculada de los datos originales y evitar todo el proceso de generación de datos falsos aleatorios que intentan retener dicha información estadística.

Lectura recomendada

Domingo-Ferrer, J.; Sánchez, D.; Soria-Comas, J. (2016). *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*. Morgan & Claypool Publishers.

4.4. Valoración general de los métodos de anonimización presentados

En este apartado se han explicado una serie de métodos de anonimización de datos y, en general, no es posible presentar uno de ellos como el más útil o el único necesario. Cada método podrá ser útil en un cierto ámbito de actuación. No obstante, de entre todos ellos sí podemos destacar los métodos de generalización y eliminación pertenecientes a la categoría de enmascaramiento no-perturbativo. Sin duda, estos dos métodos son los más utilizados para la tarea de proteger los datos tanto a nivel académico como en la industria. Es por esto por lo que dichos métodos son los que utilizaremos preferentemente en los siguientes apartados para explicar cómo se aplican los diversos modelos de privacidad para poder satisfacer los requerimientos de privacidad y resolver el problema de la anonimización.

Las razones que encontramos detrás del éxito de estos dos métodos son las siguientes: 1) son métodos muy intuitivos, con lo cual es sencillo entender cómo funcionan y qué resultado vamos a obtener al aplicarlos; 2) ofrecen facilidad en su utilización e integración en cualquier sistema de protección de datos; y 3) generan datos protegidos totalmente legítimos en contraste con otras soluciones que generan datos falsos, lo cual ayuda a retener la utilidad de los datos anonimizados.

5. Preservación de la utilidad de los datos anonimizados

Tal y como se indica en Rodríguez (2017), limitar los riesgos de revelación en los datos que hay que publicar implica modificar de alguna manera los datos originales. Estas modificaciones causan pérdida de información en los datos protegidos y estos, a su vez, pierden utilidad para los *data recipients* que quieren analizarlos y extraer conocimiento de ellos. En cualquier caso, la medida real de utilidad que proporcionan unos datos protegidos depende directamente del uso posterior que se le vaya a hacer: un conjunto de datos protegidos puede ser útil para ciertos tipos de análisis, pero totalmente inútil para otros. La gran variedad de usos potenciales de los datos hace generalmente imposible para el *data publisher* saber, a la hora de anonimizar unos datos, para qué se van a usar posteriormente dichos datos protegidos, lo cual representa un importante problema que impide ajustar correctamente la balanza entre utilidad de los datos y protección de la privacidad. Dado que los datos generalmente se protegen sin tener en cuenta el uso posterior que se les va a dar, en la literatura se considera más apropiado hablar de **pérdida de información** en lugar de utilidad.

Lectura recomendada

Sobre la pérdida de información:

Domingo-Ferrer, J.; Sánchez, D.; Soria-Comas, J. (2016). *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*. Morgan & Claypool Publishers.

Lectura recomendada

Rodríguez, M. (2017). *Semantic Perturbative Privacy-preserving Methods for Nominal Data*. Tesis doctoral. Universitat Rovira i Virgili.

Para minimizar la pérdida de información de un conjunto de datos protegidos, la estrategia más habitual es maximizar la preservación de la estructura analítica de los datos originales. Ejemplos de los elementos que sería deseable preservar serían: la media, la varianza, la generalidad/especificidad de los atributos, los valores extremos, etc. La pérdida de información de unos datos protegidos se puede medir mediante la observación de las diferencias entre los datos originales y los anonimizados, y mediante la premisa de que, si hay una pérdida limitada de la estructura analítica de los datos, esto implica que los datos protegidos son similares a los datos originales.

Lectura recomendada

Hundepool, A.; Domingo-Ferrer, J.; Franconi, L.; Giessing, S.; Schulte-Nordholt, E.; Spicer, K.; De Wolf, P. P. (2012). *Statistical Disclosure Control*. Wiley.

El análisis que se debe realizar para evaluar la pérdida de información sufrida por un conjunto de datos protegidos variará según el conjunto de datos, los métodos de anonimización aplicados, y cómo hayan sido aplicados estos. Dada la imposibilidad de dar un procedimiento concreto y definitivo para calcular la pérdida de información, a continuación, describiremos las fortalezas y debilidades respecto a la preservación de la utilidad de los datos y de los métodos típicos de anonimización comentados anteriormente:

- Enmascaramiento no-perturbativo*:* Estos métodos se caracterizan por preservar la legitimidad de los datos originales en el conjunto de datos protegidos, lo cual resulta en un conjunto de registros enmascarados consistentes con los contenidos de los registros originales, pero con menos detalle. A pesar de mantener la legitimidad de los datos originales, estos métodos incurren generalmente en grandes pérdidas de información debido a que

se basan en eliminar datos, o generalizarlos, destruyendo su nivel de detalle. Por ejemplo, los valores extremos en un atributo son uno de los objetivos preferidos de estos métodos que actúan eliminándolos o generalizándolos de forma que dejen de ser extremos. Tratar los valores extremos es un comportamiento normal dado que este tipo de valores son «raros» y, por lo tanto, facilitan la reidentificación de individuos; no obstante, los valores raros también son particularmente útiles para los investigadores ya que identifican nuevas áreas de trabajo, como por ejemplo enfermedades «raras». Cabe destacar que generalizar siempre retendrá mayor cantidad de información que eliminar; no obstante, también hay que tener en cuenta que la pérdida de información sufrida a la hora de generalizar dependerá de la jerarquía de generalización/especialización aplicada: si hay pocas opciones posibles para generalizar un valor a tratar, o dichas generalizaciones son demasiado generales (podemos generalizar el valor «Audi A3» a «Coche», pero también podríamos generalizarlo a «Objeto») la pérdida de información puede ser similar a la de una eliminación.

- Enmascaramiento perturbativo: Los métodos que perturban los datos los distorsionan: sumando valores aleatorios (añadiendo ruido), sustituyéndolos por valores agregados suficientemente «representativos», o sustituyéndolos por valores pertenecientes a otros registros. En cualquier caso, estos métodos acaban generando unos datos protegidos que no son legítimos, en otras palabras, son diferentes a los originales y pueden ser inconsistentes respecto a ellos, en el sentido de que se pueden dar combinaciones ilógicas de valores de atributos. Por ejemplo, suponiendo que los atributos del conjunto de datos sean [sexo, enfermedad], o [puesto de trabajo, salario], este tipo de métodos podrían generar unos datos protegidos como: [mujer, cáncer de próstata] o [auxiliar administrativo, 90.000 €] respectivamente. Estos resultados ilógicos pueden llevar a los análisis posteriores de los datos protegidos a conclusiones falsas que pueden ser más dañinas que simplemente la pérdida de utilidad de los datos generada por los métodos no-perturbativos. Es por ello por lo que, al aplicar este tipo de métodos, es importante controlar el nivel de distorsión introducido en los datos protegidos que serán finalmente publicados.

Como ya se ha indicado, la pérdida de legitimidad de los datos que estos métodos aplican no es un tema menor; no obstante, una ventaja importante de estos métodos es que, dependiendo de la técnica en concreto aplicada, son bastante efectivos a la hora de preservar ciertos análisis estadísticos: por ejemplo, el *intercambio de datos* preserva perfectamente estadísticos como la media, la varianza, la distribución de frecuencia o los valores extremos; dado que la *microagregación* sustituye los valores originales por agregados como la media, en este caso concreto la media de los valores protegidos quedaría perfectamente preservada; finalmente, la técnica de añadir ruido es capaz de preservar la media de los datos originales y, además, ofrece una varianza proporcional a la varianza propia del ruido añadido.

- Generación de datos sintéticos: Estos métodos generan valores nuevos totalmente aleatorios, en comparación con un método basado en añadir ruido, en el cual el dato nuevo tiene una parte original y una parte aleatoria (esto es, valor original + ruido). Los datos sintéticos tienen el mismo problema que el enmascaramiento perturbativo en el sentido de que generan datos no legítimos que pueden llevar a datos protegidos ilógicos y, a su vez, a conclusiones falsas y dañinas. Respecto a su capacidad de preservar estadísticos como media, varianza, etc., esto viene directamente determinado de cómo de similares sean los datos generados falsos respecto a los originales. Dado que la estrategia habitual se basa en generar unos datos falsos que preserven un determinado estadístico de los datos originales (por ejemplo, podríamos generar unos valores falsos nuevos para el atributo «salario», cuya media sea exactamente la misma que la media de los datos originales). Una alternativa a esta estrategia sería publicar directamente la información estadística que queremos preservar. en este caso estaríamos publicando datos agregados tabulares que no presentan problemas para la privacidad de las personas y estaríamos evitando un proceso costoso como sería la generación sintética de un conjunto de datos con escaso valor analítico (dado que los datos son falsos), más allá del estadístico concreto seleccionado.

6. Modelos de privacidad para microdatos

Anteriormente indicamos que el problema de la anonimización es producir una tabla anonimizada, a partir de una tabla original, que satisfaga unos ciertos requerimientos de privacidad, determinados por el modelo de privacidad seleccionado. Los modelos de privacidad establecen, por adelantado, unas condiciones que los datos protegidos deben cumplir para garantizar un nivel mínimo de anonimidad a los *record owners*. En otras palabras, si anonimizamos un conjunto de datos según un cierto modelo de privacidad, estaremos garantizando que los datos protegidos resultantes ofrecerán exactamente el nivel de anonimidad prometido por dicho modelo de privacidad. El rol de los métodos de anonimización que hemos visto en el apartado anterior es aplicar el modelo de privacidad que deseemos sobre los datos que hay que proteger.

A continuación, explicaremos el modelo de privacidad más célebre, la *k*-Anonimidad, un modelo que ha sido recientemente sugerido en una nota pública por la Agencia Española de Protección de Datos (AEPD) como herramienta básica para gestionar el riesgo de reidentificación. La *k*-Anonimidad es un modelo de privacidad con relevantes fortalezas, pero también con importantes debilidades; por eso, también hablaremos de dos variaciones de este modelo que buscan mejorar su nivel de protección: la *l*-Diversidad y la *t*-Proximidad. Cabe destacar que hay muchos otros modelos de privacidad en la literatura, no obstante, por diversas razones, estos no han alcanzado el nivel de éxito y aplicabilidad de la *k*-Anonimidad; por eso, en esta documentación nos centraremos en este último modelo, y en sus variaciones más relevantes, el lector interesado puede consultar la literatura de este campo de investigación (por ejemplo, Benjamin *et al.*, 2010) para conocer otros modelos de privacidad.

Lectura recomendada

Benjamin, C. M. Fung; Ke Wang, Rui Chen; Philip, S. Yu (2010). «Privacy-Preserving Data Publishing: A Survey of Recent Developments». *ACM Computing Surveys* (vol. 42, núm. 4, art. 14).

Agencia Española de Protección de Datos (AEPD) (2019). *La k-Anonimidad como medida de la privacidad* [en línea]. <https://www.aepd.es/media/notas-tecnicas/nota-tecnica-kanonimidad.pdf>

6.1. k-Anonimidad

Este modelo de privacidad (Samarati; Latanya, 1998) asume que el conjunto de atributos que un atacante puede usar para reidentificar un individuo son conocidos (el atacante los puede obtener de alguna fuente de información externa); a partir de aquí, la idea básica es ocultar cada registro de datos del conjunto de datos original en un grupo de *k* registros indistinguibles entre sí.

Lectura recomendada

Samarati, P.; Latanya, S. (1998). «Protecting privacy when disclosing information: *k*-anonymityand its enforcement through generalization and suppression». *Technical report*. SRI International.

Entrando un poco más en detalles técnicos, hemos explicado anteriormente que las combinaciones de atributos cuasi-identificadores son el factor principal de riesgo de revelación de la identidad; de esta manera, para evitar la reidentificación de registros en los datos protegidos a partir de los cuasi-identificadores, **la *k*-Anonimidad requiere como condición que cada combinación posible de valores de los cuasi-identificadores sea compartida por *k* o más**

registros. De esta manera, **anonimizar bajo *k*-Anonimidad garantiza a los *record owners* que la probabilidad de reidentificarlos en los datos protegidos será de 1/k.**

La condición que acabamos de indicar es la que el *data publisher* debe hacer cumplir en los datos protegidos, aplicando los métodos de enmascaramiento (perturbativos o no-perturbativos) que él considere oportunos para poder generar un conjunto de datos protegidos bajo las garantías de la *k*-Anonimidad.

A continuación, ilustraremos la generación de un conjunto de datos protegidos bajo este modelo de privacidad utilizando los métodos de anonimización perturbativos generalización y eliminación, los cuales han sido explicados anteriormente y que son la elección habitual a la hora de aplicar este modelo de privacidad (y muchos otros). Primero, tenemos la tabla con los datos originales que queremos proteger. Estos datos corresponden a una tabla de pacientes con sus enfermedades diagnosticadas. En particular, tenemos el atributo identificador DNI, los atributos cuasi-identificadores Edad y Código postal, y el atributo confidencial Diagnóstico.

Tabla 15. Tabla original

DNI	Edad	Código postal	Diagnóstico
37713522Z	21	23058	Cáncer pulmón
36689764P	24	23059	Cáncer páncreas
67457676C	26	23060	Cáncer colon
56564576X	27	23061	Cáncer colon
15434434V	43	23058	Gripe H1N1
76533777B	43	23059	Cáncer colon
45646767V	47	23060	Gripe H1N1
89476473L	49	23061	Gripe H1N1
64577365S	32	23058	Cáncer colon
34567646E	34	23059	Gripe H1N1
78456337R	35	23060	VIH
58776867D	38	23061	Cáncer pulmón

A continuación, mostramos la tabla protegida mediante *4*-Anonimidad, esto es, *k*-Anonimidad cuando *k=4*. De esta manera, ocultaremos los registros originales en conjuntos indistinguibles de mínimo 4 registros.

Tabla 16. Tabla protegida bajo *4*-Anonimidad

DNI	Edad	Código postal	Diagnóstico
*	20-30	230**	Cáncer pulmón
*	20-30	230**	Cáncer páncreas
*	20-30	230**	Cáncer colon
*	20-30	230**	Cáncer pulmón
*	40-50	230**	Gripe H1N1
*	40-50	230**	Cáncer colon
*	40-50	230**	Gripe H1N1
*	40-50	230**	Gripe H1N1
*	30-40	230**	Cáncer colon
*	30-40	230**	Gripe H1N1
*	30-40	230**	VIH
*	30-40	230**	Cáncer pulmón

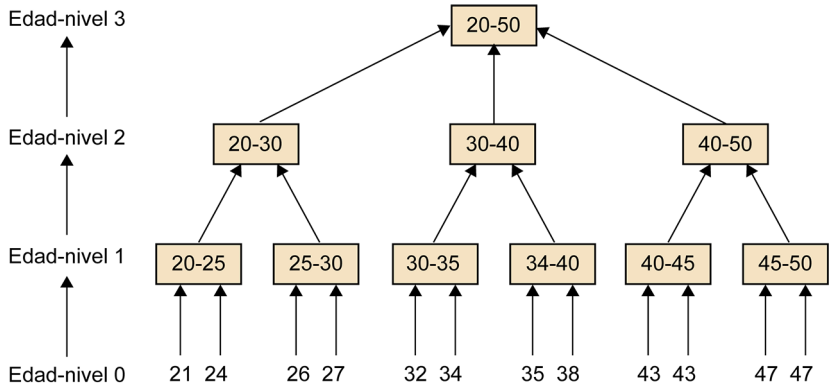
La tabla protegida es solo una de las posibles versiones que obtendríamos aplicando *4*-Anonimidad mediante generalización y eliminación. Diferentes formas de aplicar estas dos técnicas generaran diferentes versiones de esta tabla. Cabe destacar que en la literatura también se utiliza el método no-perturbativo microagregación para generar tablas protegidas bajo *k*-anonimidad (Domingo-Ferrer, 2016); aplicando este método de anonimización también se generarían diferentes versiones de esta tabla. En todo caso, lo que hemos hecho en esta tabla es, primero, eliminar el atributo identificador (DNI), dado que permitía la reidentificación directa de cada registro. A continuación, hemos generalizado los atributos cuasi-identificadores (Edad y Código postal) reduciendo la cantidad de información que proporcionan hasta conseguir que cada combinación de valores sea compartida por al menos cuatro registros (la condición que nos exigía la *4*-Anonimidad).

Lectura recomendada

Domingo-Ferrer, J.; Sánchez, D.; Soria-Comas, J. (2016). *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*. Morgan & Claypool Publishers.

Respecto a la generalización aplicada a los datos originales, en la figura 2 se muestra la jerarquía de generalización para el atributo Edad. Los valores correspondientes al nivel 0 son los valores originales. En la generalización aplicada a este atributo hemos llegado hasta el nivel 2 en la jerarquía, en el cual trabajamos con rangos de 10 elementos.

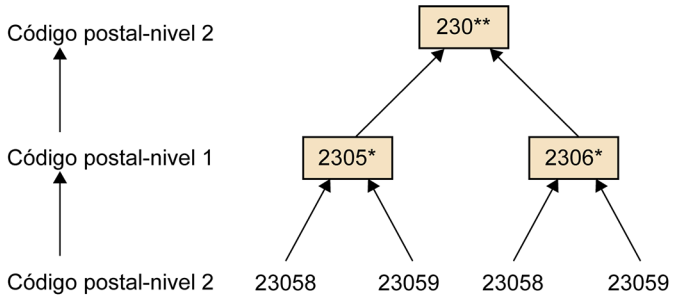
Figura 2. Jerarquía de generalización para el atributo Edad



Fuente: elaboración propia (adaptado de Domingo-Ferrer, 2016)

Para el atributo Código postal, hemos utilizado la jerarquía de generalización que se muestra en la figura 3. En la generalización aplicada a este atributo hemos llegado hasta el nivel 2 en la jerarquía, el cual engloba los cuatro códigos postales existentes en los datos originales.

Figura 3. Jerarquía de generalización para el atributo Código postal



Fuente: elaboración propia (adaptado de Domingo-Ferrer, 2016)

Ahora, partiendo de esta tabla protegida, imaginemos que tenemos un atacante que está intentando descubrir la enfermedad que le han diagnosticado a un cierto individuo. Imaginemos que este atacante sabe que el individuo en cuestión tiene un registro vinculado en la tabla protegida, que tiene 34 años y que vive en el barrio con código postal 23059. Con la combinación de datos que el atacante posee, en la tabla aparecen cuatro registros diferentes que podrían corresponder al individuo buscado. El atacante no tiene manera de saber exactamente cuál de esos cuatro registros corresponden a su objetivo. Como resultado, la única información que consigue el atacante es que su objetivo ha sido diagnosticado con una de las enfermedades que aparecen en ese grupo de cuatro registros.

Un factor muy importante para tener en cuenta al aplicar este modelo de privacidad es el valor del parámetro *k*. Aquí entran en juego dos factores: la utilidad de los datos protegidos y el nivel de privacidad que queremos dar a los *record owners*. Por el lado de la utilidad, como podemos ver en la tabla protegida, haber utilizado un k=4 nos ha obligado a utilizar un cierto nivel de generalización (pérdida de información) precisamente para poder construir grupos de cuatro registros indistinguibles. Conseguir una protección k=2 hubiese

sido más sencillo y hubiese requerido menos generalización, con lo cual, la tabla protegida retendría más información, sería más útil. De la misma manera, conseguir una protección k=6 hubiese requerido generalizar aún más los cuasi-identificadores, generando una tabla protegida con menos información y menos útil. Por el lado de la privacidad, tenemos la probabilidad de reidentificación de 1/k que garantiza aplicar *k*-Anonimidad. En este caso, un k=2 daría al atacante un 50 % de probabilidades de reidentificar al individuo en los datos protegidos. Por el contrario, un k=6 daría al atacante un 16 % de probabilidades de tener éxito en su ataque. Como podemos ver, a mayores valores de *k*, mayor garantía de anonimidad y mayor pérdida de utilidad en los datos protegidos; y viceversa. Como conclusión, el parámetro *k* se debería seleccionar buscando el equilibrio entre ambos factores. No existe una respuesta concreta sobre qué valor de *k* es el mejor, simplemente dependerá de las circunstancias.

La principal ventaja del modelo *k*-Anonimidad, y la razón por la que se explica su éxito respecto a otros modelos existentes en la literatura, es que proporciona una noción muy intuitiva de la limitación del riesgo de revelación: si ocultamos un cierto registro dentro de un conjunto de *k* registros indistinguibles, tenemos una probabilidad del 1/k de reidentificar dicho registro; es un concepto intuitivo y fácil de entender. No obstante, esta protección resulta insuficiente cuando los registros dentro del grupo *k*-anónimo tienen un valor similar para el atributo confidencial. Por ejemplo, siguiendo con la tabla protegida bajo *4*-Anonimidad anterior, tenemos un primer conjunto de cuatro registros en el cual las enfermedades diagnosticadas son cánceres de diferentes órganos; si el atacante sabe que uno de esos registros corresponde a un cierto individuo, puede inferir sin lugar a error que este sufre cáncer. En este caso sencillo podemos ver cómo el *k*-anonimato brinda protección contra la revelación de identidad: el atacante no puede saber qué registro es el que está vinculado con el individuo, solo lo puede «acertar» con una probabilidad de 1/4. No obstante, esta protección es insuficiente para evitar el riesgo de revelación del atributo cuando los valores del atributo confidencial son similares en todos los registros: el atacante no necesita reidentificar el registro para saber que el individuo sufre de cáncer.

En este sentido, en la literatura se han propuesto dos ataques para el modelo de privacidad *k*-Anonimidad que explotan la falta de variabilidad en el atributo confidencial:

- Ataque de homogeneidad (en inglés, *homogeneity attack*): Tal y como se ha comentado, si todos los registros en un grupo *k*-anónimo comparten el mismo valor para el atributo confidencial, la *k*-Anonimidad no proporciona ninguna protección contra la revelación de atributos.

- Ataque con conocimiento de antecedentes (en inglés, *background knowledge attack*): Este ataque puede ocurrir en grupos de *k* registros donde hay poca variabilidad en los valores del atributo confidencial y el atacante tiene algún tipo de conocimiento adicional sobre el individuo objeto del ata-

que. Por ejemplo, en el segundo grupo de la tabla protegida bajo *4*-Anonimidad, tenemos tres registros con «Gripe H1N1» y un registro con «Cáncer colon»; la baja variabilidad ya hace que el atacante sin información adicional tenga un 50 % de probabilidades de acertar la enfermedad que el usuario padece (en lugar del 25 % que teóricamente ofrece la *4*-Anonimidad), pero además, si el atacante tiene alguna información extra que, por ejemplo, le permite descartar el cáncer, sabrá con un 100 % de certeza que el individuo padece «Gripe H1N1»

6.2. l-Diversidad

En un intento por mitigar los problemas que la *k*-Anonimidad sufre a la hora de ofrecer protección contra la revelación del atributo, los autores en «l-diversity: privacy beyond *k*-anonymity» proponen el modelo de privacidad conocido como *l*-Diversidad.

Lectura recomendada

Machanavajjhala, A.; Kifer, D.; Gehrke, J.; Venkitasubramaniam, M. (2007). «l-diversity: privacy beyond *k*-anonymity». *ACM Transactions on Knowledge Discovery from Data* (vol. 1, núm. 1).

El objetivo de este modelo es exigir un nivel mínimo de diversidad para los valores de atributo confidencial en cada uno de los grupos de registros *k*-anónimos. Este nivel mínimo de diversidad requerido lo marca el parámetro *l*. En particular, **la condición necesaria para que un grupo de registros cumpla la *l*-diversidad es que haya al menos *l* valores diferentes bien representados en el atributo confidencial**. Esta condición la deben cumplir todos los grupos de registros indistinguibles de la tabla protegida.

La indicación de que tiene que haber *l* valores diferentes «bien representados» es bastante vaga y, en la práctica, ha llevado a varias definiciones complementarias. Por ejemplo, la definición de *l*-Diversidad más sencilla sería la conocida como *Distinct l-Diversity*, la cual simplemente pide que en el grupo de registros anonimizado aparezcan *l* valores distintos para el atributo confidencial. Otra definición más compleja sería la de *Entropy l-Diversity*, en la cual se calcula la entropía (esto se podría explicar como la «incertidumbre» o el «caos» que un elemento o conjunto de elementos proporciona; a más «caos», es que hay más variabilidad) conjunta que proporcionan los valores del atributo confidencial, dicha entropía conjunta tiene que ser igual o superior al umbral que marca el parámetro *l*.

Centrándonos en la definición de *Distinct l-Diversity* y en la tabla de ejemplo protegida bajo *4*-Anonimidad, podemos ver que el segundo grupo solo cumpliría *2*-Diversidad, dado que solo hay dos valores posibles en el atributo confidencial: «Gripe H1N1» y «Cáncer de colon»; respecto a los otros grupos de registros: el primer conjunto cumpliría *3*-Diversidad, y el tercer conjunto cumpliría *4*-Diversidad. En total, la tabla anonimizada en su composición actual solo garantizaría *2*-Diversidad (el valor más bajo de todos los grupos).

Si quisiésemos generar una versión de esa tabla protegida bajo *4*-Diversidad, deberíamos aplicar más generalización al atributo Edad (con la pérdida de información que eso representa). De esta manera podríamos tener la siguiente tabla anonimizada dejando solo dos opciones para la Edad (mayor de 35 años y menos de 35 años) y obteniendo dos grupos de registros en lugar de tres.

Tabla 17.Tabla protegida bajo *4*-Diversidad

DNI	Edad	Código postal	Diagnóstico
*	<35	230**	Cáncer pulmón
*	<35	230**	Cáncer páncreas
*	<35	230**	Cáncer pulmón
*	<35	230**	Cáncer colon
*	<35	230**	Cáncer colon
*	<35	230**	Gripe H1N1
*	>35	230**	Gripe H1N1
*	>35	230**	Cáncer colon
*	>35	230**	Gripe H1N1
*	>35	230**	Gripe H1N1
*	>35	230**	VIH
*	>35	230**	Cáncer pulmón

Como ya hemos visto, la *l*-Diversidad intenta mitigar el riesgo de revelación del atributo al requerir un mínimo nivel de variabilidad en los valores del atributo confidencial para cada grupo de registros. No obstante, este modelo no resulta completamente satisfactorio ya que es vulnerable a dos ataques bien documentados en la literatura (Domingo-Ferrer, 2016):

Lectura recomendada

Domingo-Ferrer, J.; Sánchez, D.; Soria-Comas, J. (2016). *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*. Morgan & Claypool Publishers.

- Ataque de asimetría (en inglés, *skewness attack*): Para dar la menor cantidad de información posible sobre el atributo confidencial vinculado a un individuo en concreto, *l*-Diversidad fuerza que en cada grupo de registros aparezcan un número mínimo de valores diferentes. En este sentido, la situación ideal para este modelo de privacidad sería que, en cada grupo de registros, cada valor posible tuviese la misma frecuencia de aparición; esto es, si tenemos cuatro valores posibles para un atributo, lo ideal sería que estos cuatro valores estuviesen representados bajo la misma proporción en todos los grupos de registros de la tabla protegida (o sea, al 25 % cada valor). De esta forma, un atacante no ganaría ninguna información sobre el atributo confidencial protegido de un individuo en concreto: cualquiera de los cuatro valores podría corresponder a dicha persona.
  Esta situación ideal, aunque en apariencia sea perfecta para la privacidad de los *record owners*, puede ser contraproducente para ellos si la proporción utilizada en los grupos protegidos para los valores de la variable confiden-

cial es muy asimétrica respecto a la proporción de aparición de dichos valores en el conjunto de datos original (o en el mundo en general). Para visualizar este problema, imaginemos una tabla de pacientes de un hospital donde el atributo confidencial registra la presencia o ausencia de una enfermedad determinada. Supongamos que en la tabla de datos original la incidencia de la enfermedad en cuestión es del 1 %; así pues, el 99 % de individuos vinculados a dicha tabla están libres de dicha enfermedad. Si protegemos esa tabla bajo *2*-Diversidad, lo que estaremos intentando es que en cada grupo de registros protegidos la proporción de aparición de la enfermedad sea del 50 %. De esta manera, si un atacante es capaz de vincular un cierto individuo con uno de los grupos de registros protegidos, aun no pudiendo reidentificar el registro concreto, inferirá que el individuo tiene un 50 % de probabilidades de sufrir la enfermedad, lo cual es un serio problema teniendo en cuenta que las probabilidades reales de sufrir esa enfermedad en la tabla original son del 1 %. En otras palabras, el solo hecho de aparecer en la tabla protegida con *l*-Diversidad se ha convertido en un problema para los *record owners,* dado que el modelo de privacidad ha «forzado» una sobrerrepresentación de un valor en concreto del atributo confidencial en los datos protegidos.

- Ataque de similitud (en inglés, *similarity attack*): Como ya se ha dicho, *l*-Diversidad busca como ideal que los valores posibles del atributo confidencial queden representados bajo la misma proporción en todos los grupos de registros de la tabla protegida. Un problema significativo de la aplicación práctica de este modelo es que a la hora de evaluar la diversidad de valores en un grupo de registros anonimizado no se tiene en cuenta la distancia semántica entre los valores, simplemente se considera que, si los valores son distintos, entonces hay diversidad. Nos podemos dar cuenta del problema que esto supone al ver la tabla protegida bajo *4*-Diversidad: en el primer grupo tenemos cuatro enfermedades teóricamente diferentes (tienen nombres diferentes), pero tres de esas enfermedades son «cáncer», de hecho, de los seis registros, cinco están vinculados con la enfermedad «cáncer» (son valores a la práctica similares); si sabemos que un individuo está oculto en ese grupo de registros, podemos inferir que tiene cáncer con un 83 % de probabilidad, lo cual nos indica que la *l*-Diversidad no está teniendo éxito a la hora de proteger su privacidad.

6.3. t-Proximidad

Este modelo de privacidad es un refinamiento de la *l*-Diversidad. En el caso de la *l*-Diversidad se intentaba representar todos los valores del atributo confidencial con igual proporción en cada grupo de registros protegidos, lo cual conlleva los problemas que ya hemos comentado. **En el caso de la *t*-Proximidad, la condición que hay que cumplir es que, en cada grupo de registros protegidos, los posibles valores del atributo confidencial aparezcan en la misma proporción en la cual aparecían en la tabla original (o lo más parecida posible, según el parámetro *t*).**

Lectura recomendada

Sobre el modelo de privacidad:

Li, N.; Li, T.; Venkatasubramanian, S. (2007). «*t*-closeness: privacy beyond *k*-anonymity and *l*-diversity». *ICDE* (págs. 106-115). IEEE.

Volviendo al ejemplo introducido al explicar el ataque de asimetría de la *l*-Diversidad, en ese caso la *t*-Proximidad buscará que, en cada grupo de registros, se mantenga la proporción de aparición de la enfermedad de 1 % positivos y 99 % negativos. De igual forma, la *t*-Proximidad no tiene problemas relacionados con la distancia semántica entre valores. Para visualizar esto, imaginemos que el grupo anonimizado donde cinco de los seis registros apuntaban a «cáncer» estuviese protegido bajo *t*-Proximidad (en lugar de *l*-Diversidad). *t*-Proximidad «fuerza» que las proporciones se mantengan en los grupos protegidos respecto a la tabla original; por lo tanto, si *t*-Proximidad ha generado un grupo protegido donde el 83 % de registros apuntan a distintos tipos de «cáncer», tenemos la certeza de que en la tabla original el 83 % de los registros también apuntan a distintos tipos de dicha enfermedad, y además, tenemos la certeza de que las proporciones de distintos tipos de cáncer también se mantienen; por lo tanto, un atacante no puede aprender nada nuevo simplemente por saber que un individuo está vinculado a un cierto grupo de registros protegidos.

El parámetro *t* de este modelo es un umbral que define la distancia máxima entre la distribución de valores en la tabla original y la distribución de valores en los grupos protegidos que estamos dispuestos a tolerar. Entendemos esta distancia entre distribuciones como cuánto de similares son. En la definición de este modelo de privacidad no se indica una métrica de distancias de distribuciones en concreto, aunque generalmente se utiliza la métrica *Earth Mover's Distance (EMD)* (Rubner; Tomasi; Guibas, 2000). Cabe decir que la situación ideal sería buscar una protección de los datos bajo *t=0*, lo cual implicaría que los datos protegidos han conseguido preservar exactamente las proporciones existentes en los datos originales; no obstante, para llegar a este nivel de anonimización ideal, generalmente implicará aplicar una fuerte generalización y eliminación a los datos, con la consiguiente pérdida de utilidad.

Lectura recomendada

Rubner, Y.; Tomasi, C.; Guibas, L. J. (2000). «The earth mover's distance as a metric for image retrieval». *International Journal of Computer Vision* (vol 40, núm. 2, págs. 99-121).

7. Anonimización de otros tipos de datos

En los apartados anteriores hemos tratado la anonimización de microdatos almacenados en bases de datos estadísticas. En este tipo de bases de datos, la información está perfectamente estructurada y siempre podemos encontrar un conjunto de atributos cuasi-identificadores y atributos confidenciales fijo, relativamente pequeño y con unos valores concretos y bien definidos. Por ejemplo, en las tablas de ejemplo que hemos venido utilizando se podían ver dos o tres atributos cuasi-identificadores que correspondían a elementos muy concretos, como Edad, Sexo o Código postal, los cuales tienen unos valores posibles específicos y claros. En esas tablas también podíamos ver un atributo confidencial muy claramente diferenciado del resto, dado que reflejaba una información objetivamente confidencial, como podía ser un diagnóstico médico o el salario. Dependiendo de la tabla en cuestión, el número de atributos de un tipo u otro puede variar, pero en cualquier caso, sabemos que una tabla en concreto siempre tendrá un determinado número de atributos para cada uno de sus registros; y también podremos distinguir un atributo cuasi-identificador de un atributo confidencial.

Nos hemos centrado en este tipo de estructura de datos dado que es la más típica a la hora de gestionar la información y publicarla para su uso por parte de terceros (Domingo-Ferrer, 2016); no obstante, cabe destacar que, aun siendo la forma más típica de manejar datos, existen otras opciones que merecen cierta atención. Estas otras opciones, aun trabajando igualmente sobre tablas (las tablas son una forma habitual de gestionar datos en la informática), tienen importantes diferencias respecto a lo que hemos visto hasta ahora. En particular, veremos que en ciertos tipos de datos puede haber registros que tengan más atributos que otros, o puede pasar que no sepamos diferenciar un atributo confidencial de un cuasi-identificador.

Lectura recomendada

Domingo-Ferrer, J.; Sánchez, D.; Soria-Comas, J. (2016). *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*. Morgan & Claypool Publishers.

También es importante destacar que, aun tratándose de otras formas de estructurar la información, los métodos de anonimización de microdatos más habituales que hemos visto, como la generalización o la eliminación, siguen teniendo aplicación a estos otros tipos de datos; y los modelos de privacidad vistos anteriormente, como la *k*-Anonimidad, también pueden ser utilizados aplicando algún tipo de modificación. Por lo tanto, el lector debe ver este apartado como una extensión de lo que ya se ha visto, en el que se tratan particularidades que se pueden encontrar en el campo de la publicación de datos.

En concreto, a continuación, revisaremos la problemática asociada a la anonimización de datos transaccionales, de movimiento de objetos y de datos textuales, los cuales, aun siendo menos frecuentes que los microdatos utilizados en bases de datos estadísticas, son suficientemente comunes y delicados desde el punto de vista de la privacidad como para ser considerados.

Lectura recomendada

Benjamin, C. M. Fung; Ke Wang, Rui Chen; Philip, S. Yu (2010). «Privacy-Preserving Data Publishing: A Survey of Recent Developments». *ACM Computing Surveys* (vol. 42, núm. 4, art. 14).

7.1. Datos transaccionales

Los datos transaccionales se caracterizan por tener un conjunto de atributos de número variable, y por la dificultad de distinguir atributos cuasi-identificadores de los confidenciales, hasta el punto de considerarlos cuasi-identificadores y confidenciales simultáneamente. El ejemplo típico de datos transaccionales es el de registros de compradores que han adquirido una serie de ítems (sea en una tienda en línea como Amazon o una compra realizada en una tienda física como un supermercado).

Lectura recomendada

Terrovitis M.; Mamoulis N.; Kalnis, P. (2008). «Privacy-preserving anonymization of set-valued data». *Proceedings of the VLDB Endowment* (vol. 1, págs. 115-125).

Un ejemplo gráfico sobre este tipo de datos y los problemas de privacidad que plantean se puede encontrar en «Privacy-preserving anonymization of set-valued data». De esta manera, consideremos una base de datos de un supermercado que almacena las compras realizadas por los diferentes compradores. Consideremos también a la persona Bob que ha comprado en el supermercado café, pan, queso gouda, leche, té y una bombilla. Esta compra ha generado en el supermercado un registro transaccional con un cierto identificador («12345» en la tabla de ejemplo) de la compra, y el conjunto de ítems comprados. Dicho registro se almacena en la base de datos junto con los registros transaccionales generados por otros clientes:

Tabla 18.

ID	Compra
12345	{café, pan, queso gouda, leche, té, bombilla}
34543	{leche, lejía}
55435	{galletas, zumo de naranja, pan}

Lo primero que podemos ver en la tabla de datos original es que el atributo Compra es muy diferente a los atributos que nos encontrábamos en las tablas de microdatos. El valor que puede tener este atributo es totalmente variable, tanto en número de ítems como en la combinación de ítems seleccionado: podemos tener valores para el atributo Compra que sean únicos, en otras palabras, puede ocurrir que una persona compre una combinación de ítems que no haya comprado nadie más. Esto representa un fuerte contraste respecto a los atributos como Sexo o Diagnóstico que veíamos en los microdatos; en esos atributos el valor correspondía a un conjunto bastante limitado de opciones (concretamente, Sexo solo admitía dos opciones), y el valor era obligatoriamente único. Con los datos transaccionales, en cambio, podemos tener un atributo con un valor, o varios, y estos pueden corresponder a un conjunto

de opciones enorme, y que es relativamente fácil que aumente con el tiempo o que cambie (un supermercado trae nuevos productos y elimina otros periódicamente).

Siguiendo con el ejemplo, imaginemos que Bob va con su compra en el autobús, y un atacante puede ver parte de su compra: café, pan y bombilla. Posteriormente, el supermercado publica la base de datos transaccionales eliminando los atributos identificadores (este seguiría siendo el primer paso en cualquier proceso de anonimización, sean microdatos o cualquier otro tipo de datos):

Tabla 19.

ID	Compra
*	{**café**, **pan**, queso gouda, leche, té, **bombilla**}
*	{leche, lejía}
*	{galletas, zumo de naranja, **pan**}

El atacante puede comprobar la lista de ítems que él sabe que Bob compró con todos los registros de la tabla publicada, y si esa combinación solo aparece en un único registro, tendrá la certeza al 100 % de que ese registro corresponde a Bob (revelación de identidad) y podrá conocer la lista completa de ítems que este ha comprado (revelación de atributo). Es importante en este caso ver cómo el atributo Compra actúa simultáneamente de atributo confidencial (los elementos comprados son el elemento confidencial de los datos) y de atributo cuasi-identificador, puesto que es la combinación de valores de este atributo el que lleva a la reidentificación del registro.

Cabe destacar que este tipo de datos, y su problemática asociada, aparecen siempre que tenemos atributos que pueden tomar como valores un conjunto variable y múltiple de elementos; no están limitados a la compra de ítems en tiendas. Por ejemplo, los *query logs* de AOL que vimos al principio de este documento son otro ejemplo típico de datos transaccionales. En concreto, una consulta como «*dog cornea treatment*» enviada al motor de búsqueda AOL, se registraría como:

Tabla 20.

Seudónimo	Consulta
2178	{*dog, cornea, treatment*}

Vemos cómo tenemos el atributo Consulta, el cual toma como valor un conjunto de elementos variable y múltiple, en este caso, formado por los sustantivos que formaban la consulta. Se puede ver la semejanza de este tipo de datos con el ejemplo de la compra en el supermercado.

Una forma habitual de proteger este tipo de datos se describe en «Privacy-preserving anonymization of set-valued data» y pasa por aplicar generalización y eliminación, tal y como se hacía en los conjuntos de microdatos. En el ejemplo que hemos venido tratando, la generalización que podríamos realizar es sustituir «café», «leche» y «zumo de naranja» por un ítem más general llamado «bebida desayuno». Respecto a la eliminación, podríamos eliminar «bombilla», asumiendo para el ejemplo que es un elemento muy «raro» en la base de datos original (muy poca gente lo compra). La tabla resultante quedaría así:

Lectura recomendada

Terrovitis M.; Mamoulis N.; Kalnis, P. (2008). «Privacy-preserving anonymization of set-valued data». *Proceedings of the VLDB Endowment* (vol. 1, págs. 115-125).

Tabla 21.

ID	Compra
*	{**bebida desayuno**, **pan**, queso gouda, leche, té}
*	{**bebida desayuno**, lejía}
*	{galletas, **bebida desayuno**, **pan**}

Si el atacante accede a esta tabla, con la información parcial que posee: café, pan y bombilla, y sabiendo que café equivale a «bebida desayuno», verá dos registros diferentes que podrían corresponder a Bob, con lo cual ya se estaría reduciendo el riesgo de revelación. Aplicando más generalización y eliminación se podría reducir aún más dicho riesgo, a costa de reducir la utilidad de los datos protegidos (a semejanza de lo visto con los microdatos).

Cabe destacar que un modelo de privacidad que se plantea en la literatura para proteger datos transaccionales es la k^m-Anonimidad, una variación de la *k*-Anonimidad, en la cual, asumiendo que el atacante conoce un máximo de *m* ítems de una transacción específica, se evitará que pueda reidentificar dicha transacción en un grupo de *k* transacciones protegidas.

Lectura recomendada

Sobre k^m-Anonimidad:

Terrovitis M.; Mamoulis N.; Kalnis, P. (2008). «Privacy-preserving anonymization of set-valued data». *Proceedings of the VLDB Endowment* (vol. 1, págs. 115-125).

7.2. Datos de movimiento de objetos

Los datos de movimiento de objetos y personas están relacionados directamente con los servicios basados en localización, en inglés, «*Location-Based Services* (LBS)», servicios que trabajan con las posiciones físicas que ocupan en ciertos instantes de tiempo los objetos/individuos y con cierta información de contexto o personal. Cabe decir que una secuencia de múltiples posiciones físicas visitadas en ciertos instantes de tiempo corresponde a una trayectoria. En vista de estas características, podemos concluir que los datos de movimiento de los objetos o personas tienen las siguientes particularidades: dependen de la localización, dependen del tiempo, y se generan en grandes cantidades.

Lectura recomendada

Benjamin, C. M. Fung; Ke Wang, Rui Chen; Philip, S. Yu (2010). «Privacy-Preserving Data Publishing: A Survey of Recent Developments». *ACM Computing Surveys* (vol. 42, núm. 4, art. 14).

Un ejemplo gráfico sobre este tipo de datos y los problemas de privacidad que plantean sería el siguiente (Benjamin *et al.*, 2010): un hospital quiere publicar una tabla de datos de pacientes que contiene: 1) un identificador del paciente; 2) sus trayectorias en la ciudad donde residen; y 3) la enfermedad que se les ha diagnosticado. La tabla con los datos originales quedaría de la siguiente forma:

Tabla 22.

DNI	Trayectoria	Diagnóstico
37713522Z	A1 → D2 → B3 → E4 → F6 → C7	VIH
36689764P	B3 → E4 → F6 → E8	Gripe H1N1
67457676C	B3 → C7 → E8	Gripe H1N1
56564576X	D2 → F6 → C7 → E8	Alergia
15434434V	D2 → C5 → F6 → C7	VIH
76533777B	C5 → F6 → E9	Cáncer colon

La trayectoria de un individuo es una secuencia de múltiples parejas de localización e instante de tiempo. La localización queda representada como una letra que, por ejemplo, puede equivaler a una posición específica GPS, o una calle, o un barrio, etc. El instante de tiempo queda representado como un número que codifica una *fecha+hora* determinada, que, por ejemplo, puede equivaler a 14/08/2019+20:07. Dado que la cantidad de localizaciones es variable y diferente para cada registro, tenemos una situación similar al caso de los datos transaccionales. Aunque en este caso el atributo confidencial (el diagnóstico) y el atributo cuasi-identificador (la trayectoria) están, *a priori*, diferenciados.

Siguiendo con el ejemplo, imaginemos que un atacante tiene acceso a una versión de la tabla anterior en la que la única medida de protección aplicada ha sido eliminar el DNI de los pacientes. Imaginemos también que el atacante sabe que su objetivo, Alice, ha visitado E en el instante 4, y C en el instante 7.

Tabla 23.

DNI	Trayectoria	Diagnóstico
*	A1 → D2 → B3 → **E4** → F6 → **C7**	VIH
*	B3 → **E4** → F6 → E8	Gripe H1N1
*	B3 → C7 → E8	Gripe H1N1
*	D2 → F6 → **C7** → E8	Alergia
*	D2 → C5 → F6 → **C7**	VIH
*	C5 → F6 → E9	Cáncer colon

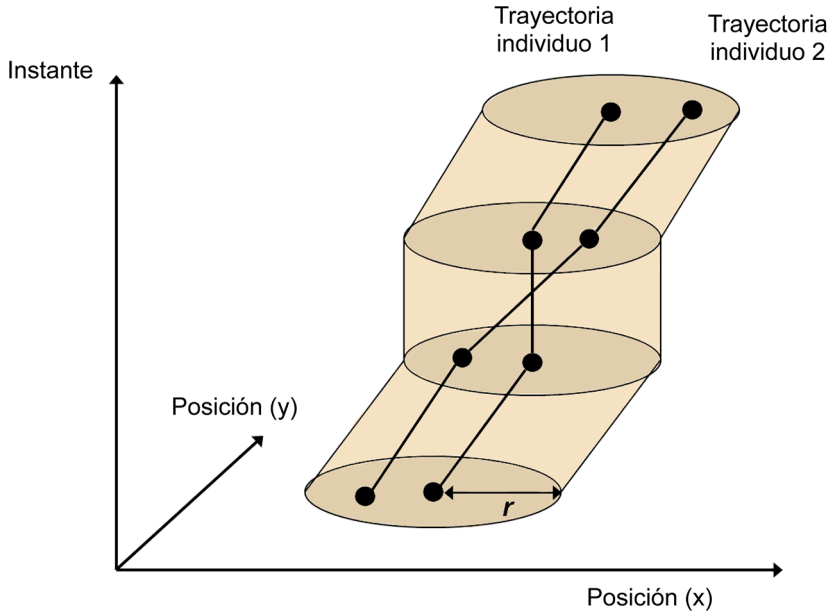
Con esta información parcial, el atacante es capaz de reidentificar el registro perteneciente a Alice, dado que solo una trayectoria de la tabla contiene esas dos localizaciones en los instantes indicados. Como resultado, tenemos una revelación de identidad que, además, lleva a revelar también el diagnóstico de Alice.

Imaginemos ahora que, en lugar de Alice, el objetivo del atacante es Bob, y que sabe que Bob ha estado en la localización D en el instante 2, y en la localización F en el instante 6. En este caso, el atacante encuentra tres registros que pueden corresponder a Bob (el primero, el cuarto, y el quinto). En esta situación, el atacante no puede re-identificar el registro vinculado a Bob, pero puede inferir que Bob tiene VIH con un 67 % de probabilidad. En este caso, tenemos un claro riesgo de revelación del atributo, que puede acabar en un 100 % de éxito en caso de que el atacante tenga algún conocimiento adicional sobre el tipo de enfermedad que sufre Bob.

Como resultado, podemos ver que los datos de movimiento de objetos son ligeramente diferentes a los datos que hemos visto hasta ahora, pero también podemos ver que tienen importantes semejanzas con datos transaccionales y con microdatos. De hecho, comparten los mismos problemas respecto a los riesgos de revelación y el efecto del conocimiento de antecedentes que pueda tener el atacante. De la misma forma a como se ha intentado anonimizar microdatos y datos transaccionales, en el caso de los datos de movimiento, el proceso de anonimización más habitual también pasa por aplicar generalización y eliminación a los puntos de las trayectorias bajo una adaptación de la *k*-Anonimidad. En particular, la idea básica en este caso es crear un área de incertidumbre, según un cierto radio *r,* alrededor de las posiciones físicas mediante *generalización* (por ejemplo, si la localización real es una calle, podríamos generalizarla a un barrio), y, de esta manera, conseguir que dentro de esa área de incertidumbre coincidan *k* individuos diferentes en un cierto instante de tiempo.

La figura 4 representa visualmente dos trayectorias originales correspondientes al movimiento de dos individuos diferentes protegidas mediante *2*-Anonimidad para datos de movimiento de objetos. La generalización de los datos originales (por ejemplo, generalizar las coordenadas exactas GPS al barrio de la ciudad que engloba dichas posiciones GPS) queda representada como un área de incertidumbre alrededor de los puntos de localización, de forma que esa área sustituye las posiciones exactas y hace que los dos individuos aparezcan en los datos protegidos ocupando el mismo lugar en el mismo instante de tiempo. De esta forma se consigue la anonimización de sus trayectorias (asumiendo la pérdida de información/utilidad correspondiente).

Figura 4. Anonimización de datos de movimiento de individuos



Fuente: elaboración propia (adaptado de Benjamin *et al.*, 2010).

7.3. Datos textuales

Los tipos de datos anteriores no eran tipos totalmente estructurados como los microdatos, dado que teníamos atributos que podían tomar un número múltiple e indefinido de valores y, en ciertas circunstancias, podía ser complicado distinguir los atributos cuasi-identificadores de los confidenciales. Aun así, estos datos seguían manteniendo suficiente estructura como para ser representados mediante tablas, al igual que hacíamos con los microdatos. De la misma forma, la anonimización de esos tipos de datos, que podríamos llamar «semi-estructurados», acaba siendo bastante similar a la protección que aplicamos a los microdatos, esto es, usamos los mismos métodos de anonimización (generalización y eliminación), y la *k*-Anonimidad (más o menos adaptada) sigue siendo el modelo de privacidad habitual.

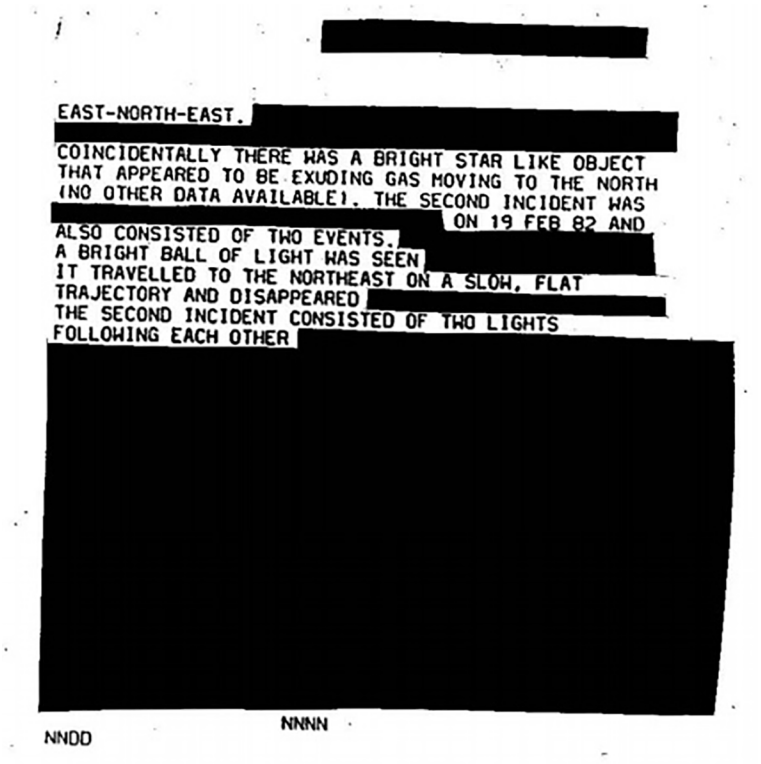
En contraste con todo lo visto anteriormente, tenemos los datos textuales, entendiéndolos como documentos de texto en el cual aparecen referencias a individuos y a datos personales suyos que se pueden considerar confidenciales. Estos datos no pueden ser representados como tablas dado que son texto libre, estos datos no tienen atributos y tampoco hay un conjunto limitado de valores. En la literatura, a estos tipos de datos se les llama «no-estructurados», y se les considera muy difíciles de tratar dada la complejidad de evaluar si un cierto texto (una palabra, una frase, o un párrafo entero) representa una amenaza a la privacidad de algún individuo.

Lectura recomendada

Sánchez, D.; Batet, M.; Viejo, A. (2013). «Automatic General-Purpose Sanitization of Textual Documents». *IEEE Transactions on Information Forensics and Security* (vol. 8, núm. 6, págs. 853-862).

La anonimización de datos textuales se ha basado tradicionalmente en redactar aquellos elementos del texto que se podían considerar peligrosos. Redactar es eliminar los elementos elegidos del texto, sustituyéndolos por una marca negra. Podemos ver un ejemplo del resultado de redactar un documento en la figura 5.

Figura 5. Documento redactado



El proceso de redacción tradicionalmente es un método manual, esto es: un experto tiene que leer el texto y discernir qué elementos son peligrosos y cuáles son inocuos. Aparte del propio conocimiento del experto, también existen guías sobre qué elementos deben ser eliminados. Por ejemplo, en el contexto médico tenemos las reglas *HIPAA (Health Insurance Portability and Accountability Act)*, propuestas por Estados Unidos en 1996, que indican exactamente qué elementos identifican directamente a una persona (nombre, número de seguridad social, etc.) y, por lo tanto, deben ser eliminados de cualquier texto previamente a su publicación. Cabe destacar que, aun con las diferencias entre datos textuales y microdatos, podemos ver que, al final, el comportamiento a la hora de anonimizar es similar: en los microdatos eliminábamos los atributos identificadores de las tablas, en los datos textuales eliminamos cualquier palabra o frase que pueda identificar a una persona.

Lectura recomendada

Department of Health and Human Services, Office of the Secretary (2000). *The Health Insurance Portability and Accountability Act of 1996*. Tech. Rep. Federal Register 65 FR 82462.

El problema de la anonimización manual de documentos utilizada tradicionalmente es doble: 1) anonimizar manualmente todos los documentos que se publican en entornos abiertos como Internet a día de hoy ya es inabarcable, hacerlo en el futuro cuando absolutamente todo se publique en formato digital y se generen informes textuales fácilmente de casi cualquier cosa se prevé

imposible; y 2) como ya vimos al hablar de microdatos, los métodos basados en eliminar información destruyen de forma significativa la utilidad de los datos protegidos.

Para solventar el primer problema, se han propuesto soluciones que intentan automatizar el proceso de anonimizar textos. La idea básica en estos casos es analizar los textos originales con herramientas informáticas de procesado de lenguaje natural de forma que se puedan extraer los elementos relevantes del texto, por ejemplo, los sustantivos. Una vez se tienen los elementos relevantes identificados, se puede proceder de diversas maneras, la más habitual y sencilla (Chakaravarthy *et al*. 2008) sería consultar los elementos encontrados en bases de datos para evaluar su grado de confidencialidad. Por ejemplo, en un documento médico, primero eliminaríamos los nombres de personas, después cualquier tipo de número, y finalmente podríamos eliminar todos aquellos sustantivos que coincidan con nombres de enfermedades, tratamientos y síntomas, utilizando una base de datos especializada. Otras opciones en la literatura se basan en evaluar la cantidad de información que cada elemento relevante del texto proporciona y eliminar aquellos que son demasiado informativos, por el riesgo de revelación que estos puedan tener (Sánchez *et al*. 2013).

Lectura recomendada

Chakaravarthy, V. T.; Gupta, H.; Roy, P.; Mohania, M. (2008). «Efficient techniques for document sanitization». *Proceedings of the ACM Conf. Information and Knowledge Management* (págs. 843-852).

Lectura recomendada

Sánchez, D.; Batet, M.; Viejo, A. (2013). «Automatic General-Purpose Sanitization of Textual Documents». *IEEE Transactions on Information Forensics and Security* (vol. 8, núm. 6, págs. 853-862).

El problema principal de estas herramientas automáticas es que el lenguaje es ambiguo, puede contener errores ortográficos, y está vivo (cada año el vocabulario que usamos aumenta con nuevos conceptos). Como resultado, para un software puede ser difícil detectar de forma efectiva, primero, si una parte de un texto es relevante o no y segundo, si esa parte es peligrosa o no para la privacidad de algún individuo. Por poner un ejemplo, puede ser sencillo para un software encontrar el nombre de enfermedad «HIV» en un texto y eliminarla; pero si el texto habla también de los síntomas de dicha enfermedad, y el software no los elimina por considerarlos inocuos, un atacante con cierto conocimiento que lea el texto resultante anonimizado puede acabar infiriendo que este trata sobre «HIV».

Respecto al problema de la información que se pierde al eliminar datos, en la literatura se ha considerado usar solo la eliminación para los identificadores directos, mientras que el resto de los elementos textuales (que podrían actuar como cuasi-identificadores y/ atributos confidenciales) se protegen mediante generalización, a semejanza de lo que se hacía con los microdatos. La idea básica en este caso es tener jerarquías de generalización/especialización suficientemente amplias como para abarcar todos los conceptos del lenguaje (puesto que un texto libre puede tratar cualquier tema) o, al menos, los conceptos que se vayan a tratar en la tipología de documentos que queremos anonimizar (si los documentos son médicos, con una jerarquía de conceptos médicos podría ser suficiente). Al proceso de generalizar los datos textuales (en lugar de eliminarlos) para retener la mayor cantidad de información posible se le denomina *sanitizar documentos*.

8. Integración de la privacidad en el desarrollo de productos tecnológicos

En este apartado, primero detallaremos los principios relativos al tratamiento de los datos personales que deben cumplir los responsables de dicho tratamiento y, por lo tanto, todo aquel nuevo producto que se quiera desarrollar y que trabaje con datos personales. A partir de esos principios, introduciremos una serie de conceptos que emanan directamente de ellos como son: la protección de datos desde el diseño y por defecto; las *privacy enhancing technologies* (tecnologías de mejora de la privacidad); y las evaluaciones de impacto en la protección de datos (EIPD).

8.1. Principios relativos al tratamiento de los datos personales

La normativa actual aplicable a la protección de datos personales configura una serie de principios que deben ser respetados cuando se produzca el tratamiento de los datos. De esta manera, cualquier producto que trate con datos personales y, por lo tanto, tenga que tomar medidas para la protección de la privacidad de los individuos debería cumplir con los siguientes principios recogidos en el art. 5 RGPD:

Lectura recomendada

Agencia Española de Protección de Datos (AEPD) (2019). *Protección de Datos: Guía para el Ciudadano* [en línea]. <https://www.aepd.es/media/guias/guia-ciudadano.pdf>

- Principio de licitud, lealtad y transparencia: Cuando se recolectan datos personales estos deben ser tratados de manera lícita, leal y transparente. De conformidad con este principio, se debe proporcionar toda la información necesaria sobre el objeto y fines del tratamiento, sus consecuencias y posibles riesgos para el interesado, tanto si los datos han sido recolectados directamente de su propietario, o han sido cedidos por una tercera entidad.

- Principio de limitación de la finalidad: Los datos personales se recolectan para unos fines determinados, explícitos y legítimos. De conformidad con este principio, la finalidad del tratamiento de los datos personales debe estar claramente definida, así como permitida por el ordenamiento jurídico.

- Principio de minimización de datos*:* Los datos personales recolectados se limitarán a lo mínimo necesario en relación con los fines para los que son tratados. De conformidad con este principio no es posible recabar y guardar o tratar datos simplemente por si pudiesen ser útiles en un futuro.

- Principio de exactitud: Los datos personales almacenados deben ser exactos y actuales. De conformidad con este principio se deben proporcionar medidas para poder actualizar o suprimir los datos.

- Principio del plazo de conservación: los datos personales almacenados serán mantenidos por el plazo de tiempo mínimo necesario para cumplir

con sus fines. De conformidad con este principio, la conservación de datos debe limitarse a las finalidades para las cuales se han recabado dichos datos. Una vez cumplidas estas finalidades, los datos deben ser borrados o, al menos, desprovistos de todo elemento que permita identificar a los interesados.

- Principio de integridad y seguridad*:* Los datos personales serán tratados de manera que se garantice su adecuada seguridad, incluyendo la protección contra el tratamiento no autorizado o ilícito y contra su pérdida, destrucción o daño accidental, aplicando las medidas técnicas y de organización apropiadas. De conformidad con este principio, aquellas entidades que traten con datos personales deben actuar proactivamente con el objetivo de protegerlos frente a cualquier riesgo que amenace su seguridad. En este sentido, las *privacy enhancing technologies* (tecnologías de mejora de la privacidad) son métodos que ayudan a mitigar las amenazas de seguridad y proteger los datos de los individuos.

- Principio de responsabilidad proactiva: Los responsables y encargados de tratamiento deben cumplir los principios anteriores y deben ser capaces de demostrar dicho cumplimiento. De conformidad con este principio, cualquier producto que trate con datos personales debe aplicar los principios adicionales llamados protección de datos desde el diseño y protección de datos por defecto. En la literatura, estos principios también se conocen como *privacy by design* y *privacy by default*.

8.2. Protección de datos desde el diseño y por defecto

Tal y como indica la Agencia Española de Protección de Datos, en el RGPD se hace referencia a dos principios para la implementación efectiva de la responsabilidad proactiva como son los de protección de datos desde el diseño y protección de datos por defecto (art. 25 RGPD).

Lectura recomendada

Agencia Española de Protección de Datos (AEPD) (2019). *Medidas de protección de datos desde el diseño y por defecto* [en línea]. <https://www.aepd.es/reglamento/ cumplimiento/privacidad-por -defecto.html>

El principio de protección de datos desde el diseño tiene como objetivo que la protección de datos se encuentre presente desde el inicio de la concepción de un proyecto (esto es, su etapa de diseño). Esta línea de actuación se debe traducir en la implementación de las medidas técnicas y organizativas necesarias para aplicar de forma efectiva los principios de protección de datos que garanticen el correcto tratamiento de los datos personales.

Por ejemplo, esta línea de actuación en una supuesta tienda en línea de aplicaciones para *smartphones* nos llevaría a diseñar el software a cargo de las operaciones de compra de forma que aplique la anonimización de los datos personales de los clientes (utilizando alguna de las técnicas que hemos hablado en este documento, por ejemplo, *k*-Anonimidad basada en generalización y

eliminación) desde el momento que estos realizan las compras, de forma que ante cualquier uso posterior (autorizado o no) de los datos adquiridos, estos ya estuviesen correctamente protegidos.

Un punto importante para tener en cuenta sobre el desarrollo de productos que tengan entre sus tareas el tratamiento de datos personales es la necesidad de demostrar que estos se han desarrollado teniendo en cuenta la protección de datos desde el diseño. La piedra angular de esta demostración es la Evaluación de Impacto en la Protección de Datos (EIPD), una herramienta que analiza las actividades de tratamiento de datos previstas, sus riesgos y de la cual se obtiene un informe con las medidas que deben aplicarse para poder garantizar la protección de los datos.

Lectura recomendada

Agencia Española de Protección de Datos (AEPD) (2019). *Guía práctica para las Evaluaciones de Impacto en la Protección de los Datos sujetas al RGPD* [en línea]. <https://www.aepd.es/media/guias/guia-evaluaciones-de-impacto-rgpd.pdf>

Respecto al principio de protección de datos por defecto, este se refiere a que solo se deben adquirir, tratar y almacenar aquellos datos personales que sean estrictamente necesarios para los fines del producto desarrollado. Como en el caso de la privacidad desde el diseño, aplicar privacidad por defecto se traduce en la implantación de medidas técnicas y organizativas que permitan actuar en las siguientes operaciones del tratamiento de datos, teniendo en cuenta los principios comentados anteriormente y que deben ser cumplidos:

- Recogida de datos: recabando la mínima cantidad de datos personales posibles en función de los productos y servicios seleccionados por el usuario.
- Tratamiento de los datos: cualquier proceso que utilice datos personales debe acceder a los mínimos datos posibles necesarios para cumplir con su tarea.
- Conservación: se debe implementar una política de conservación de datos que permita eliminar aquellos datos que no sean estrictamente necesarios.
- Accesibilidad: limitar el acceso por parte de terceros a los datos personales.

8.3. Privacy enhancing technologies (PET)

Las *privacy enhancing technologies* (PET) son tecnologías de mejora de la privacidad cuyo objetivo es proteger la privacidad de los usuarios de tecnología; esto es, son tecnologías que protegen la confidencialidad de datos de los individuos.

Lectura recomendada

Wang, Y.; Kobsa, A. (2008). *Privacy-enhancing technologies, Handbook of Research on Social and Organizational Liabilities in Information Security* (págs. 352-375).

Esta es una categoría de herramientas muy general y, por lo tanto, engloba a todas las tecnologías que se han creado para proteger la privacidad de las personas de muy diversas maneras. Por ejemplo, existen PET que anonimizan la navegación web de los individuos, existen PET que cifran las comunicaciones entre entidades para que un atacante externo no pueda acceder a las transmisiones o modificarlas de alguna manera; existen PET que proporcionan con-

trol a los usuarios de una red social sobre quién puede acceder a ciertos datos personales o mensajes; o existen PET que ofuscan la localización física de una persona de forma que un servicio basado en localización no pueda conocer su localización exacta en un determinado instante de tiempo.

Los PET, por tanto, engloban muchas tecnologías diferentes que protegen la privacidad de los individuos de diferentes maneras, y pueden ser utilizadas por diferentes actores durante la ejecución de un servicio. Por ejemplo, como hemos dicho, un usuario de servicios basados en localización podría usar una PET controlada por él mismo para que la entidad que ofrece el servicio no conozca nunca su localización exacta; o podría ser dicha entidad la que utilizase una PET para proteger los datos de localización que obtiene de sus usuarios; o, el usuario y la entidad podrían usar un PET para comunicarse de forma cifrada y segura entre ellos evitando la acción de cualquier atacante externo.

La gran diversidad de las PET existentes en la literatura hace complicado clasificarlas o destacar unas soluciones sobre otras. A continuación describiremos algunas tipologías de PET por su relevancia y cotidianidad, pero no pretendemos hacer una clasificación formal, ni indicar que estas sean las PET más relevantes.

- Securización de comunicaciones: En las comunicaciones entre dos entidades en entornos inseguros, por ejemplo, un usuario que quiere comprar un ítem en una tienda en línea y tiene que proporcionar sus datos bancarios, es habitual utilizar la tecnología TLS/SSL para cifrar las comunicaciones entre el explorador web y el servidor web.

- Anonimización de la navegación web*:* Cada vez que un usuario accede a una página web envía al servidor correspondiente una gran cantidad de datos, los cuales incluyen las conocidas *cookies*, pero también la configuración del explorador web utilizado. De esta manera, la navegación web deja un rastro que permite el seguimiento de los individuos. Una herramienta famosa para anonimizar la navegación web es la red TOR. Esta red hace de intermediario entre el explorador web y el servidor web que ofrece la página web, de forma que, para solicitar la página web, la red TOR crea una combinación de máquinas situadas en la red que serán las que finalmente pidan la página web al servidor, en lugar del usuario. Como resultado, el servidor web nunca sabrá quién solicitó la página web en realidad.

- Sistemas de gestión de la identidad: Cada vez que un individuo quiere usar un servicio, lo normal es que este deba autenticarse al servidor que lo ofrece. Un proceso de autenticación implica que el servidor debe conocer ciertos datos privados del individuo (por ejemplo, una contraseña), de forma que el individuo pueda demostrar su identidad revelando datos que solo él debería conocer. Con la gran cantidad de servicios tecnológicos que existen en la actualidad, el hecho de que todas las entidades que los ofrecen tuviesen que conocer datos privados de los individuos podría ser un

problema importante de privacidad y seguridad. Para evitar esto, existen tecnologías como OpenID que nos permite simplemente compartir datos privados con una única entidad y que esta nos «represente» delante de todas las demás.

- Sistemas de control de acceso: Los sistemas de control de acceso permiten a los individuos definir exactamente mediante políticas de acceso qué entidad puede actuar (leer, usar, borrar, etc.) sobre un cierto recurso (texto, imagen, aplicación, etc.). Esto es muy habitual por ejemplo en las redes sociales, donde un usuario puede publicar algo y hacer que solo sus amigos puedan leerlo, o comentar sobre ello; o también podría hacer que un cierto recurso publicado fuese accesible para todo el mundo. De esta manera, la configuración de privacidad de una red social es una PET, aunque muy sencilla. Sistemas de control de acceso más complejos y con más funcionalidades serían por ejemplo los basados en ABAC (esto es, control de acceso basado en atributos), aplicados mediante una arquitectura XACML con su respectivo lenguaje de políticas de acceso.

- Anonimización de microdatos: Hemos dedicado la mayor parte de esta documentación a explicar cómo se protege la privacidad de los individuos en entornos de publicación de datos. En este sentido, el conjunto de técnicas necesarias para generar una tabla de microdatos protegida bajo *k*-Anonimidad son PET.

- Transferencias de datos anónimas mediante BlockChain: En la actualidad, es muy habitual que las transferencias de información importante entre dos individuos requieran una tercera entidad, teóricamente de confianza, que dé fe de la transacción que está ocurriendo y gestione sus efectos. Estaríamos hablando, por ejemplo, de transacciones de dinero electrónico entre un comprador y una tienda en línea, donde hay un banco que gestiona la transferencia de dinero de una cuenta a otra en su entorno seguro; o estaríamos hablando también de algún tipo de contrato legal con ciertas obligaciones para las dos partes, que debe ser verificado por un notario. La existencia de esta tercera entidad que lo sabe todo y lo gestiona todo representa en sí misma un problema de privacidad. La PET que viene a solucionar este problema es el BlockChain.
  El BlockChain es una base de datos de transferencias de información totalmente distribuida donde son los individuos quienes «escriben» sobre ella y verifican que lo que otros han escrito sea correcto. Así pues, en este entorno no hay una entidad central que lo sabe todo y lo gestiona todo, en su lugar, hay una comunidad de individuos que se han dado unas reglas comunes de actuación y en la cual, mientras la mayoría de los individuos se comporte de forma honesta, todo funcionará correctamente.
  El BlockChain en sí no es más que una especie de «pizarra pública» donde los individuos anotan todas las transacciones que se han realizado entre ellos desde el inicio del sistema. Un ejemplo de transacción para el caso de la sustitución de los bancos sería: «Individuo 1234 transfiere diez monedas

a Individuo 2223». Para que la comunidad diese por buena esta transacción, lo que haría sería comprobar en la pizarra (el histórico de transacciones realizadas desde el primer día) que el individuo «1234» realmente tiene diezmonedas, para ello, la comunidad comprobaría que entre las transferencias que ha recibido y que ha enviado aún le quedan diez monedas. Si «1234» tiene esas 10 monedas, la comunidad da la nueva transferencia por buena y, a partir de ese momento, el individuo «2223» puede usar esas monedas para hacer otras nuevas transferencias.

La «pizarra» de la que hablamos tiene todo un conjunto de tecnologías criptográficas detrás para que las transferencias se realicen de forma segura y confiable (lo cual incluye el uso de firmas digitales). Un factor interesante de cara a la privacidad es que los individuos no usan su identidad real sino un identificador que se genera a partir de unas claves criptográficas y que los individuos pueden cambiar cuando deseen. De esta manera, un observador no puede conocer la identidad real del individuo «1234». No obstante, la privacidad ofrecida de esta manera puede romperse fácilmente si en algún momento tenemos que vincular la identidad real con el identificador utilizado en el BlockChain, por ejemplo, si una compra requiere que nos identifiquemos.

Finalmente, cabe destacar que la finalidad del BlockChain es conseguir que una comunidad de individuos iguales pueda gestionar transferencias entre ellos ocultando sus identidades reales. El contenido de las transferencias no se cifra dado que toda la comunidad tiene que poder verlas y verificar su validez; sí se aplica criptografía para garantizar su integridad y autenticidad, esto es: solo el propietario de las claves criptográficas necesarias puede demostrar que es el «Individuo 1234» y nadie puede alterar una transferencia de las que aparecen en la pizarra pública.

8.4. Evaluación de impacto en la protección de datos (EIPD)

La aplicación del principio de protección de datos desde el diseño requiere que el responsable del tratamiento de datos personales considere desde el inicio, en la fase de diseño, las acciones preventivas suficientes para poder identificar, evaluar y tratar los riesgos asociados al tratamiento de datos personales, y así, poder asegurar los principios de protección de los datos garantizando los derechos y libertades de los interesados.

Para cubrir este requerimiento tenemos la herramienta Evaluación de Impacto en la Protección de Datos Personales (EIPD), la cual permite evaluar de manera anticipada cuáles son los riesgos a los que están expuestos los datos personales en función de las actividades que se llevarán a cabo con los mismos, y permite establecer una respuesta a dichos riesgos adoptando las medidas de protección necesarias para reducirlos hasta un nivel de riesgo aceptable.

Lectura recomendada

Agencia Española de Protección de Datos (AEPD) (2019). *Guía práctica para las Evaluaciones de Impacto en la Protección de los Datos sujetas al RGPD* [en línea]. <https://www.aepd.es/media/guias/guia-evaluaciones-de-impacto-rgpd.pdf>

El proceso de realizar una EIPD consta de tres secciones que, a su vez, se desglosan en diferentes tareas:

1) Contexto
a) Describir el ciclo de vida de los datos.
b) Analizar la necesidad y proporcionalidad del tratamiento

2) Gestión de riesgos
a) Identificar amenazas y riesgos
b) Evaluar los riesgos
c) Tratar los riesgos

3) Conclusión
a) Plan de acción e informe de conclusiones

El ciclo de vida de los datos se divide en las siguientes etapas: captura de datos, almacenamiento, uso/tratamiento, cesión a terceros, y destrucción. En esta tarea del EIPD, se debe poder responder las siguientes preguntas para cada etapa del ciclo de vida:

- Quiénes llevan a cabo los siguientes roles: interesados, responsable del tratamiento, encargado del tratamiento y terceras partes.
- Qué sistemas de información hay involucrados.
- Qué flujos de datos hay entre los diferentes sistemas de información y los diferentes roles.
- Quién tiene acceso a los datos y con qué finalidad.
- Cuál es la base legitimadora de las actividades de tratamiento (Consentimiento expreso, relación contractual, interés legítimo, etc.).

En la tarea analizar la necesidad y proporcionalidad de las actividades de tratamiento se responderán las siguientes cuestiones:

- Qué se va a hacer con los datos y con qué finalidad.
- Qué datos se van a tratar. Si son necesarios todos ellos. Quién es el afectado por los datos que hay que tratar. Si se le ha informado correctamente.

Estas cuestiones están directamente relacionadas con los principios de: licitud, lealtad y transparencia; minimización de datos y limitación de la finalidad. La proporcionalidad que habrá que aplicar en el tratamiento de datos se basa en evaluar si la finalidad que se persigue se puede conseguir por otros medios, por ejemplo: utilizando otros datos o haciendo uso de otros métodos menos invasivos, por ejemplo, modificando otros datos que se tengan para evitar recolectar nuevos.

En la tarea Identificar amenazas y riesgos, primero se identifican las amenazas, teniendo en cuenta las operaciones que se realizan en todo el ciclo de vida de los datos, y la clasificación de amenazas según su tipología: 1) acceso ilegitimo

a datos; 2) modificación no autorizada de los datos; y 3) eliminación de los datos. De esta manera, en la etapa de «almacenamiento» del ciclo de vida de los datos podríamos tener como una amenaza de tipo «acceso ilegitimo a datos» un «ataque intencionado realizado por un *hacker*», o un «acceso intencionado por parte de personal no autorizado»; o podríamos tener como amenaza de tipo «eliminación de datos» un «error humano que provoque borrado de datos».

Una vez se identifican las amenazas, estas se relacionan con riesgos y su posible impacto. Por ejemplo, la amenaza «acceso intencionado por parte de personal no autorizado» en la etapa de ciclo de vida «almacenamiento» puede producir como riesgo la «vulneración de los derechos y libertades» de los propietarios de los datos, y el impacto sería «daño moral, físico o material».

La tarea de identificar amenazas, riesgos y su impacto se puede procedimentar siguiendo los listados de riesgos asociados al cumplimiento normativo que ofrece la Agencia Española de Protección de Datos.

La tarea de *evaluar los riesgos* requiere, primero, estimar la probabilidad de que el riesgo se materialice, y estimar el nivel de impacto en caso de que este finalmente ocurra. La probabilidad de riesgo y el nivel de impacto se clasifican en 4 niveles numéricos: probabilidad/impacto despreciable (valor 1); probabilidad/impacto limitado (valor 2); probabilidad/impacto significativo (valor 3); y probabilidad/impacto máximo (valor 4). Con los valores estimados, se puede calcular el nivel de riesgo de la amenaza (llamado, **riesgo inherente**) multiplicando el valor de probabilidad de riesgo y el valor de impacto; según el resultado numérico obtenido, el nivel de riesgo de la amenaza se clasificará en cuatro niveles: bajo (1-2), medio (3-6), alto (7-9), y muy alto (10-16).

En la tarea de tratar los riesgos se definirán las medidas necesarias para tratar aquellas amenazas con niveles de riesgo por encima de lo aceptable. Existen tres medidas principales para tratar el riesgo: reducirlo mediante medidas de control; transferirlo a otra organización, por ejemplo, contratando a una aseguradora que afronte las posibles consecuencias materiales; o anularlo completamente, lo cual implica no realizar el tratamiento/uso de datos al cual está vinculada la amenaza.

Entre las medidas de control que buscar reducir el riesgo tenemos: 1) medidas organizativas, por ejemplo, definir procedimientos para ejercer los derechos de los interesados; 2) medidas legales, por ejemplo, cláusulas para recogida de consentimientos expresos; y 3) medidas técnicas, que son aquellas que velan por la seguridad de los activos de información, esto son las *privacy enhancing technologies*, por ejemplo, anonimización bajo *k*-Anonimidad, securizar comunicaciones bajo SSL/TLS, etc.

Tras la aplicación de las medidas de control, se debe calcular de nuevo el nivel de riesgo de la amenaza (llamado, **riesgo residual**) en cuestión, para poder evaluar y documentar de qué manera se ha reducido el riesgo inherente.

La tarea final corresponde al plan de acción e informe de conclusiones. En esta etapa se documentará el resultado de la EIPD junto con el plan de acción que incluya las medidas de control que habrá que implantar para gestionar los riesgos identificados. Si el resultado de la EIPD es no favorable, se debe analizar la posibilidad de incluir medidas de control adicionales que permitan reducir el nivel de exposición al riesgo hasta un nivel aceptable. Si el resultado es favorable, la actividad de tratamiento/uso de los datos personales se puede llevar a cabo siempre y cuando las medidas de control incluidas en el plan de acción hayan sido implantadas.

Cabe destacar que, para cumplir con el principio de protección de datos desde el diseño, es necesario considerar el plan de acción resultante de esta tarea durante la fase de definición del nuevo producto tecnológico que tiene que utilizar datos personales.

Como nota final, para detalles específicos sobre cómo realizar correctamente una EIPD, el lector debería consultar la «Guía práctica para las Evaluaciones de Impacto en la Protección de los Datos sujetas al RGPD» de la Agencia Española de Protección de Datos.

Lectura recomendada

Agencia Española de Protección de Datos (AEPD) (2019). *Guía práctica para las Evaluaciones de Impacto en la Protección de los Datos sujetas al RGPD* [en línea]. <https://www.aepd.es/media/guias/guia-evaluaciones-de-impacto-rgpd.pdf>

Resumen

En este módulo hemos dado una descripción completa de las amenazas a la privacidad y los mecanismos de protección vinculados con la publicación, principalmente de microdatos; pero también con la publicación de otros tipos de datos, como los transaccionales, de movimiento de objetos o textuales. También hemos explicado cómo se tiene en cuenta la protección de la privacidad en el desarrollo de nuevos productos tecnológicos que tratan con datos personales, explicando conceptos relevantes a este tema, como la privacidad desde el diseño y por defecto.

En el primer apartado hemos hablado de los tipos de datos habituales que se utilizan al publicar información en entornos abiertos como Internet.

En el segundo apartado hemos hablado de los dos objetivos principales a la hora de proporcionar privacidad en la publicación de datos: el anonimato y la confidencialidad. Nos hemos centrado en la diferencia entre datos anonimizados y datos seudonimizados y, finalmente, hemos explicado cuál es el problema que se debe resolver al realizar un proceso de anonimización de datos.

En el tercer apartado hemos descrito el riesgo de revelación de identidad y el riego de revelación de atributo. Riesgos que definen directamente la estrategia de protección a seguir para minimizarlos.

En el cuarto apartado hemos explicado los diferentes métodos de anonimización de datos, distinguiendo entre métodos basados en perturbación, no-perturbación, y generación de datos sintéticos.

En el quinto apartado hemos hablado sobre la importancia de preservar la utilidad de los datos protegidos y hemos dado explicaciones sobre el nivel de retención de la utilidad obtenido según el método de anonimización utilizado.

En el sexto apartado hemos explicado qué es un modelo de privacidad, y hemos detallado el modelo de *k*-Anonimidad, el más célebre en la literatura, y sus extensiones.

En el séptimo apartado nos hemos centrado en la anonimización de datos transaccionales, de movimiento de objetos y textuales; siendo los dos primeros similares a los microdatos, pero con ciertas peculiaridades a las cuales hay que dar atención. El caso de los datos textuales es bastante más diferente, aunque las medidas de protección aplicadas para este tipo de datos al final son similares.

Finalmente, en el octavo apartado, hemos explicado la integración de la privacidad en el desarrollo de productos tecnológicos partiendo de los principios relativos al tratamiento de datos personales establecidos por el Reglamento general de protección de datos (RGDP). En este apartado hemos hablado de conceptos relevantes, como la protección de datos desde el diseño y por defecto, las *privacy-enhancing technologies,* y finalmente, hemos explicado someramente cómo se realiza una evaluación de impacto en la protección de datos (EIPD), la cual es una herramienta que nos permite demostrar documentalmente la aplicación de la privacidad desde el diseño en el desarrollo de tecnologías que traten con datos personales.

Bibliografía

Agencia Española de Protección de Datos (AEPD) (2019). *Guía práctica para las Evaluaciones de Impacto en la Protección de los Datos sujetas al RGPD*. [en línea]. <https://www.aepd.es/media/guias/guia-evaluaciones-de-impacto-rgpd.pdf>

Agencia Española de Protección de Datos (AEPD) (2019). *La k-Anonimidad como medida de la privacidad*. [en línea]. <https://www.aepd.es/media/notas-tecnicas/nota-tecnica-kanonimidad.pdf>

Agencia Española de Protección de Datos (AEPD) (2019). *Medidas de protección de datos desde el diseño y por defecto*. [en línea]. <https://www.aepd.es/reglamento/cumplimiento/privacidad-por-defecto.html>

Agencia Española de Protección de Datos (AEPD) (2019). *Protección de Datos: Guía para el Ciudadano*. [en línea]. <https://www.aepd.es/media/guias/guia-ciudadano.pdf>

Barbaro M.; Zeller T. (2006). «A face is exposed for AOL searcher no. 4417749». *New York Times* (agosto).

Benjamin, C. M. Fung; Ke Wang, Rui Chen; Philip, S. Yu (2010). «Privacy-Preserving Data Publishing: A Survey of Recent Developments». *ACM Computing Surveys* (vol. 42, núm. 4, art. 14).

Chakaravarthy, V. T.; Gupta, H.; Roy, P.; Mohania, M. (2008). «Efficient techniques for document sanitization». *Proceedings of the ACM Conf. Information and Knowledge Management* (págs. 843-852).

Dalenius, T.; Reiss S. P. (1978). Data-swapping: a technique for disclosure control, Proceedings of the ASA Section on Survey Research Methods (págs. 191-194).

Defays, D.; Anwar, M. N. (1998). «Masking microdata using micro-aggregation». *Journal of Official Statistics* (vol. 14, núm. 4, págs. 449-461).

Department of Health and Human Services, Office of the Secretary. (2000). *The Health Insurance Portability and Accountability Act of 1996*. Tech. Rep. Federal Register 65 FR 82462.

Domingo-Ferrer, J.; Sánchez, D.; Soria-Comas, J. (2016). *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*. Morgan & Claypool Publishers.

Greenberg, B. (1996). *Rank swapping for masking ordinal microdata*. Washington DC : U.S. Bureau of the Census.

Hundepool, A.; Domingo-Ferrer, J.; Franconi, L. *et al.* (2012). *Statistical Disclosure Control*. Wiley.

Latanya, S. (2000). «Uniqueness of Simple Demographics in the U.S. Population». *LIDAP-WP4*. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh PA.

Li, N.; Li, T.; Venkatasubramanian, S. (2007). «*t*-closeness: privacy beyond *k*-anonymity and *l*-diversity». *ICDE* (págs. 106-115). IEEE.

Machanavajjhala, A.; Kifer, D.; Gehrke, J.; Venkitasubramaniam, M. (2007). «l-diversity: privacy beyond *k*-anonymity». *ACM Transactions on Knowledge Discovery from Data* (vol. 1, núm. 1).

Parlamento Europeo (2016). *Reglamento (UE) 2016/679, general de protección de datos (RGPD)*. [en línea]. <https://www.boe.es/doue/2016/119/L00001-00088.pdf>

Rodríguez, M. (2017). *Semantic Perturbative Privacy-preserving Methods for Nominal Data*. Tesis doctoral. Universitat Rovira i Virgili.

Rubin, D. B. (1993). «Discussion: statistical disclosure limitation». *Journal of Official Statistics* (vol. 9, págs. 462-468).

Rubner, Y.; Tomasi, C.; Guibas, L. J. (2000). «The earth mover's distance as a metric for image retrieval». *International Journal of Computer Vision* (vol 40, núm. 2, págs. 99-121).

Samarati, P.; Latanya, S. (1998). «Protecting privacy when disclosing information: *k*-anonymityand its enforcement through generalization and suppression». *Technical report* . SRI International.

Sánchez, D.; Batet, M.; Viejo, A. (2013). «Automatic General-Purpose Sanitization of Textual Documents». *IEEE Transactions on Information Forensics and Security* (vol. 8, núm. 6, págs. 853-862).

Sánchez, D.; Viejo, A. (2017). «Personalized privacy in open data sharing scenarios». *Online Information Review* (vol. 41, núm. 3).

Soria-Comas, J.; Domingo-Ferrer, J. (2015). «Big data privacy: challenges to privacy principles and models». *Data Science and Engineering* (pág. 1-8).

Terrovitis M.; Mamoulis N.; Kalnis, P. (2008). «Privacy-preserving anonymization of set-valued data». *Proceedings of the VLDB Endowment* (vol. 1, págs. 115-125).

US Federal Trade Commission. (2014). *Data brokers, a call for transparency and accountability.*

Wang, Y.; Kobsa, A. (2008). «Privacy-enhancing technologies», Handbook of Research on Social and Organizational Liabilities in Information Security (págs. 352-375).

Warren; Brandeis (1890). «The Right to Privacy». *Harvard Law Review* (vol. 193).

Willenborg, L.; De Waal T. (2001). *Elements of Statistical Disclosure Control*. Nueva York: Springer-Verlag.