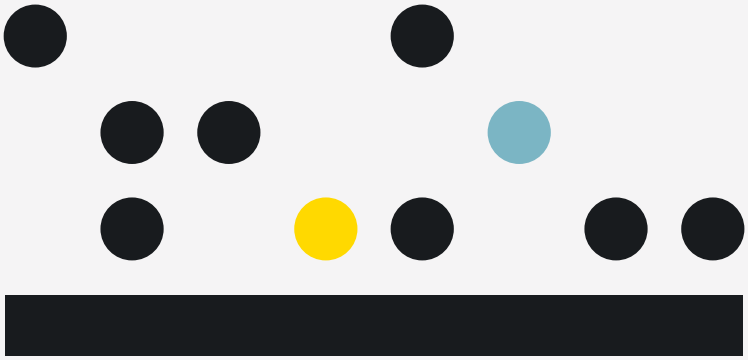
Anónimos pero no tanto



Datos

Daniel Yankelevich

Abril 2021

fundar

Anónimos pero no tanto

Cómo hacer una gestión de datos eficiente sin poner en riesgo la privacidad

Daniel Yankelevich

Índice

Anónimos pero no tanto

Cómo hacer una gestión de datos eficiente sin poner en riesgo la privacidad

fundar

Introducción

En 2005 la Universidad de Harvard inició un estudio de alto impacto sobre el genoma humano, denominado Personal Genome Project, para el cual reclutó a 2500 donantes voluntarios de material genético. A ellos se les aclaró desde el inicio que la información que proporcionaran iba a mantenerse en el anonimato: en efecto, a partir del material genético se puede determinar la predisposición a enfermedades o a distintas condiciones, además de que, en el cuestionario, los voluntarios daban información sobre alcoholismo y uso de drogas, entre otros datos personales, todos contenidos sensibles cuya difusión podía impactar negativamente en la vida de esas personas[1].

El equipo del Personal Genome Project borró nombres, números de documento y otros datos de los donantes considerados "personales". Como se quería conservar cierta información sobre distribución geográfica, se enmascararon los domicilios y se conservó solo la información del código postal. El director del proyecto expresó entonces sus reparos sobre el riesgo de que alguien pudiera reidentificar los registros y acceder a los datos privados asociados a cada persona. Ese riesgo se materializó mucho más rápido de lo esperado, cuando el laboratorio de Data Privacy de Harvard, dirigido por Latanya Sweeney, en un experimento que intentaba poner a prueba la confiabilidad de estos procedimientos, logró reidentificar un 43% de los datos de una muestra de donantes que tomó. La fecha de nacimiento, el código postal y el género resultaron suficientes para identificar los nombres de las personas (consultando información electoral y otras fuentes) (Sweeney et al., 2013). De hecho, en otro trabajo, Sweeney (2000) estimó que un 87% de la población de los Estados Unidos podía identificarse en forma inequívoca en función de esos tres datos.

Este caso hace surgir varias preguntas que abordaremos en este trabajo. ¿Es este un caso singular o es posible reidentificar registros a través de algunos pocos datos si se consultan bases externas? ¿Qué conjuntos de datos, así como ocurrió con la fecha de nacimiento y el código postal, permiten desanonimizar una base de datos? ¿Se pueden formular recomendaciones claras e identificar mejores prácticas para anonimizar un grupo de datos y proteger así la identidad y privacidad de las personas?

> ¿Es posible reidentificar registros a través de algunos pocos datos si se consultan bases externas? ¿Qué conjuntos de datos permiten desanonimizar una base de datos? ¿Se pueden formular recomendaciones claras e identificar mejores prácticas para anonimizar un grupo de datos y proteger así la identidad y privacidad de las personas?

La Asociación Internacional de Auditores de Sistemas (Isaca) dio inicio en 2021 a su programa de certificación de profesionales en privacidad (Certified Data Privacy Solutions Engineer) y no es casualidad: en las últimas décadas, la privacidad se ha convertido en un tema sensible. La gestión de datos no es sólo una cuestión de buen criterio, sentido común o mejores regulaciones: también requiere un trabajo técnico de alto nivel para ser llevada adelante adecuadamente.

En este contexto, en empresas, organismos públicos y semipúblicos, e incluso en el extenso mundo de los datos abiertos, hoy existen varios incentivos para no compartir datos y casi ninguno para hacerlo. Hacer pública alguna información o difundirla, hasta dentro de una organización, implica responsabilidades de varios tipos, incluso penales, que hacen que la solución más frecuentemente

1 "Harvard Professor Re-Identifies Anonymous Volunteers In DNA Study", Forbes, 25/03/2013, disponible en: https://www.forbes.com/sites/adamtanner/2013/04/25/harvard-professor-re-identifies-anonymous-volunteers-in-dna-study/?sh=6ff2e4bb92c9

tomada por quienes están a cargo de la gestión de datos sea mantenerlos en secreto. Sostendremos aquí que esa estrategia no es una solución al riesgo que siempre implica difundir datos, sino que es el peor camino que se puede elegir. Cuando no se comparte información se le quita valor y se sabotea la posibilidad de utilizarla para tomar mejores decisiones. Por la misma razón, tampoco es una estrategia inteligente "agregar" los datos (es decir, publicar un promedio o la suma de todos ellos) en forma descuidada como única opción, otro de los procedimientos muchas veces utilizados para tratar de minimizar los riesgos de la difusión. En efecto, así mostrada, esa información no tiene ninguna utilidad si se trata de hacer análisis sofisticados.

Para abordar este tema es útil diferenciar dos conceptos: desidentificar y anonimizar. Desidentificar es eliminar elementos que asocien un registro a una persona individual, como códigos de identificación personal, códigos de dispositivos (direcciones IP, MAC) e identificadores biométricos. Anonimizar, en tanto, es eliminar la posibilidad de asociar registros con los individuos a los que esos registros se refieren. No se trata de acciones en un modelo binario, sino que en general existe un continuo entre registros identificados y registros completamente anonimizados.

> La anonimización de grupos de datos requiere una visión técnica y un cuidadoso análisis. Proponemos una metodología de alto nivel para encontrar soluciones al problema de compartir información sin violar la privacidad y de garantizar la anonimización de las bases de datos.

Como mostraremos en este artículo, la anonimización de grupos de datos requiere una visión técnica y un cuidadoso análisis. Aquí propondremos una metodología de alto nivel para encontrar soluciones al problema de compartir información sin violar la privacidad y de garantizar la anonimización de las bases de datos.

Las primeras dos preguntas que planteamos más arriba se pueden responder analizando algunos casos en los que se logró desanonimizar datos con características muy diferentes usando técnicas también distintas.

Netflix

En 2006 Netflix propuso un concurso que desembocaría finalmente en las actuales competencias de ciencia de datos, el sitio Kaggle.com[2] y las contrataciones "abiertas". Desde hacía tiempo, Netflix venía trabajando en su algoritmo de recomendación que, según las películas y series que una persona había visto, intentaba predecir qué otras películas o series le gustaría ver. Mejorar este algoritmo era complejo y caro. La competencia propuesta tenía las siguientes reglas: podía inscribirse cualquier grupo de ciencias de datos de cualquier lugar del mundo; cada grupo recibiría la base de datos con las valoraciones que hacían los clientes de las películas que habían visto (por ejemplo, el cliente 11234 le había puesto cinco estrellas a *El Señor de los anillos*, tres estrellas a *Piratas del Caribe* y ninguna estrella a *Diario de una pasión*, después de haber visto las tres películas), excepto por una pequeña porción de datos que Netflix se "guardaba" para sí mismo. Obviamente, no se daba a los concursantes información personal de los usuarios (como nombre o región geográfica en que

2 Kaggle es una plataforma que permite a organizaciones de todo tipo publicar problemas de ciencia de datos que necesitan resolver con el formato de competencias. Grupos de analistas y científicos de todo el mundo pueden inscribirse, por lo general las competencias tienen un premio en efectivo para el ganador y la organización obtiene una solución a su problema.

vivían), sino sólo un número arbitrario que los identificaba. Cada día, los equipos que concursaban podían proponer a lo sumo un algoritmo, que se validaba contra esa porción de datos "reservados" y se le informaba el error (cuánto se había equivocado el algoritmo al predecir preferencias sobre ese grupo de datos que Netflix no había compartido con los equipos). Al finalizar el período de la competencia, aquel equipo que mejor desempeño hayan tenido dentro de aquellos que hubieran mejorado el algoritmo original de Netflix en al menos un 10%, ganaría 1 millón de dólares como premio (Bennett y Lanning, 2007).

Finalmente, dos equipos lograron exactamente el mismo error, pero de acuerdo con las reglas en ese caso el primero en haber sido propuesto resultaba ganador; uno fue subido al sitio 20 minutos antes que el otro, y se hizo acreedor al millón de dólares.

Esta competencia de Netflix dio origen no sólo a una forma de relación con grupos de ciencia de datos sino también a nuevas técnicas y tecnología para sistemas de recomendación. Sin embargo, el final no es feliz. Entre los más de 2500 equipos de todo el mundo que se inscribieron para resolver el problema, algunos usaron los datos para otros *experimentos*. Narayanan y Shmatikov, dos investigadores de la Universidad de Texas, por ejemplo, mostraron que con la ayuda de una base de datos externa (en este caso, IMDb, un sitio online con información de películas), podían identificar registros correspondientes a un grupo de personas incluidas en la base publicada por Netflix para la competencia, aun con muy poca información sobre ellas (Narayanan y Shmatikov, 2008). Al identificarlas se revelaban además preferencias políticas, sexuales, y otro tipo de información personal. Además, a partir de esta experiencia los investigadores presentaron una serie de estrategias para reidentificar datos personales que podían aplicarse en situaciones similares.

¿Cuán grave es hacer públicas las preferencias personales sobre películas? Lo suficiente como para que el caso llegara a los tribunales: en diciembre de 2009 un grupo de usuarios inició una demanda colectiva contra Netflix. El caso se cerró en una negociación privada, pero fue lo suficientemente complejo como para que Netflix decidiera suspender nuevas competencias, incluso algunas que ya habían sido anunciadas.

Taxis

Como parte de su política de datos abiertos, en 2013, la ciudad de Nueva York hizo pública la información de todos los viajes en taxi en la ciudad del año anterior, previamente anonimizados. Esta base incluía datos tales como inicio y fin de cada viaje, costo, propina y los datos encriptados (en realidad, mapeados mediante un hash[3]) de las patentes e identificación del conductor. Esta base parecía no contener información que pudiera violar la privacidad de los pasajeros. Sin embargo, Anthony Tockar, un estudiante australiano de posgrado en temas de datos, combinó esa información con otra fácilmente disponible sobre celebridades (fotos de paparazzis y notas periodísticas sobre espectáculos). De esta forma, logró identificar viajes de algunos famosos (en particular, de Bradley Cooper y Jessica Alba), y mediante algunas comparaciones adicionales identificó viajes regulares, domicilios y más información ("todos los jueves visita tal domicilio por 50 minutos, ¿será su terapeuta?"). Incluso se permitió hacer bromas sobre las propinas que dejaba Bradley Cooper (Mitnick, 2017). Esto funcionó con celebridades porque existen fotos en revistas y sitios *online* pero, sin llegar a ese nivel de exposición mediática, todos dejamos un rastro digital en nuestra vida cotidiana, por ejemplo al usar la tarjeta SUBE en el transporte público.

3 Los datos estaban procesados y presentados usando una función hash, que permite mapear datos de longitud variable a un formato fijo y se puede usar para encriptar. Las funciones hash, por lo general, funcionan en un solo sentido, es decir, no se puede volver atrás desde el mapeo al dato original.

Unos años después, Vijay Pandurangan, un especialista canadiense en ingeniería de software, demostró que el encriptado de las licencias de los taxis no era útil, ya que era completamente reversible. Además, publicó en internet un detalle de su análisis, que era reproducible[4]. Pandurangan notó que el encriptado de las licencias respondía a un algoritmo conocido (el hash MD5). Como los números de licencia tienen un formato fijo (seis dígitos, y casi todas comienzan con el mismo número), sencillamente probó encriptar todos los números posibles con ese algoritmo y ver cuáles coincidían con el encriptado. Suena como mucho trabajo, pero con una computadora de potencia media es completamente posible. En otras palabras: si uno usa funciones de encriptado en forma descuidada, en realidad no está asegurando los datos, sino que la cadena se rompe por el eslabón más débil.

Antenas

Como vemos, existen casos de reidentificación en contextos bastante heterogéneos. En una línea de trabajo completamente distinta, un grupo de investigadores (De Montjoye et al., 2013) analizó las rutas de movilidad en una muestra de 1.500.000 personas durante 15 meses usando la información de las antenas con las que se conecta cada celular. Estos patrones permiten identificar a las personas en forma casi unívoca: usando cuatro puntos tomados en rangos de una hora y con la misma resolución de las antenas de los proveedores, se puede identificar en forma única al 95% de los individuos.

Zang y Bolot analizaron los datos de celulares para determinar los caminos usuales de las personas sólo con la información de sus llamadas. En un trabajo (Zang y Bolot, 2011) examinaron un conjunto de datos masivo (30.000 millones de registros correspondientes a 25 millones de teléfonos celulares en los Estados Unidos) para demostrar que, a partir de los registros de llamadas, se podían inferir las locaciones más comunes de los usuarios de estos celulares. Esto permite asociar llamadas a lugares donde se encuentran las personas y con alta probabilidad a sus comportamientos (por ejemplo, las dos principales locaciones suelen ser domicilio y trabajo, pero se pueden identificar otras, casas de amigos, familiares, colegios, clubes, etc.).

Redes sociales

Como último ejemplo, veremos que es también posible utilizar información de estructura — datos no asociados a una persona, sino a una red a la que pertenece, como quiénes son sus amigos, con quiénes se conecta— para identificar alguien en una red social, incluso si usa un seudónimo. En un experimento, se utilizó un algoritmo para mostrar que un tercio de los usuarios que tienen cuenta en Twitter y Flickr podían ser reidentificados con un error de sólo 12%, mediante el uso de información de topología de la red (Narayanan y Shmatikov, 2009).

En las redes sociales, la pertenencia a determinados grupos (o incluso los "likes" a ciertas fotos o artículos) puede ser útil para reidentificar: un grupo de investigadores demostró que es posible utilizar un método para acceder a la historia de un navegador en un dispositivo y a partir de ahí identificar a las personas que acceden a un determinado sitio malicioso[5] (Wondracek et al., 2010). Este método fue probado en la red Xing, una red social alemana con unos 20 millones de usuarios, pero puede aplicarse en otros contextos.

4 "On Taxis and Rainbows. Lessons from NYC's improperly anonymized taxi logs", V. Pandurangan, 21/06/2014. Disponible en https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1

5 Se conoce con este nombre a sitios web que parecen inofensivos o que simulan ser otros sitios para que quien entre en ellos termine bajando un virus, troyano o similar.

Mejores prácticas

Como dijimos, es usual que, frente a la necesidad de anonimizar datos, el camino elegido sea no compartirlos.

En efecto, vemos cada vez más, y sobre todo en la administración pública, una tendencia a armar "silos de datos": grupos separados y autónomos que disponen de información y la usan pero no la comparten, por lo que quedan aislados entre sí. Con la intención de no incurrir en problemas de violación de la privacidad, se termina perdiendo el valor y la mayor riqueza de la utilización de datos, que es cruzar información de diferentes fuentes.

No compartir datos es equivalente a apagar una computadora como medida de seguridad frente a un potencial virus informático. Dado que, como dijimos, no compartir datos puede tener un impacto más negativo que hacerlo, en este artículo proponemos una metodología para compartir o publicar datos teniendo los necesarios cuidados para preservar la intimidad de las personas involucradas.

> Vemos una tendencia a no compartir datos. Con la intención de no incurrir en problemas de violación de la privacidad, se termina perdiendo el valor y la riqueza de la utilización de datos, que es cruzar información de diferentes fuentes. No compartir datos es equivalente a apagar una computadora como medida de seguridad frente a un potencial virus informático.

En los ejemplos presentados más arriba se pueden identificar algunas características del problema que venimos describiendo:

- Las cuestiones de privacidad y reidentificación pueden aparecer en distintos contextos y con datos muy heterogéneos.

- Sin embargo, se pueden considerar algunas prácticas en común: cruzar bases de datos (la base supuestamente anonimizada con datos públicos, por ejemplo), ignorar que las decisiones que afectan la privacidad son de diseño y que por eso deben analizarse antes de publicar datos; subestimar a quien puede tener acceso a los datos e intentar reidentificarlos (es decir, pensar que nadie se va a tomar la molestia de hacer cierto trabajo que complejo pero posible).

- La puesta en práctica de soluciones parciales o inadecuadas (*hashing*, eliminar columnas, etc.) que terminaron resultando insuficientes.

- La cuestión no es sólo regulatoria, es técnica y requiere una visión acorde.

Por todo lo dicho, queda claro que la solución no es sencilla o única y que depende del contexto[6]. Desde el área de datos de Fundar proponemos una metodología de aproximación al problema que se basa en cuatro pasos.

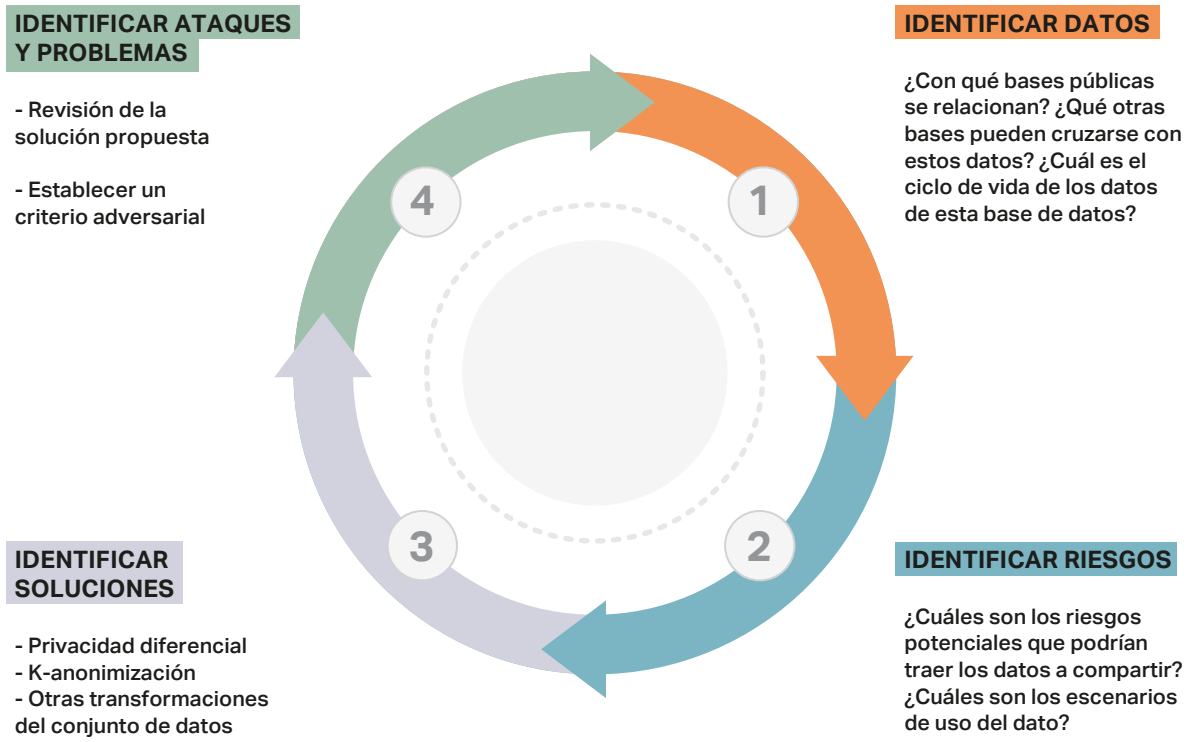
---

6 Esto no quiere decir que no pueda realizarse un análisis sistemático y una propuesta de soluciones, tal como puede verse en El Emam et al. (2009). Un excelente compendio de técnicas para analizar datos manteniendo privacidad puede encontrarse en Aggarwal y Yu (2008).

Mejores prácticas

Metodología de gestión eficiente de datos sin poner en riesgo la privacidad

Gráfico 1



Fuente: Elaboración propia.

1- Identificar datos

Más allá de qué tipo de datos se trata, es necesario preguntarse: ¿con qué bases públicas se relacionan? ¿Qué otras bases pueden cruzarse con estos datos? ¿Cuál es el ciclo de vida de los datos de esta base?

También individualizar aquellos campos que pueden actuar como identificadores y que deben ser borrados, modificados o eliminados. Por ejemplo, son identificadores: códigos, códigos de aparatos (direcciones IP, MAC) e identificadores biométricos. Este paso es necesario pero no resulta completamente obvio que se realice en forma sistemática. A modo de ejemplo, en 2007 el Grupo de trabajo del Artículo 29 de la Comisión Europea, que monitorea la protección de datos, reclamó a Google y Yahoo que anonimizaran las direcciones IP de sus usuarios con frecuencia . Las empresas respondieron que en efecto anonimizan estas direcciones, pero mantienen los primeros dígitos. Como se desprende de los ejemplos que presentamos, en dominios muy reducidos enmascarar la información parcialmente resulta inadecuado y facilita la reidentificación, con lo cual el dato de una dirección IP parcialmente enmascarada sin dudas debe considerarse como un potencial identificador.

Adicionalmente, es necesario entender qué información es confidencial o sensible y puede recuperarse en función de uno o más campos de estos datos. Por ejemplo, podría ocurrir que la identidad de la persona (su número de documento, nombre o un dato biométrico) sea el dato confidencial y se elimine de la base. Pero, como vimos en el primer ejemplo, en algunos casos puede recuperarse

usando fecha de nacimiento, código postal y género. De la misma forma, puede ser que no demos información en forma directa, pero que una o más variables puedan combinarse para recuperar el dato crítico con determinada probabilidad o confianza.

Finalmente, en esta etapa es necesario identificar qué propiedad de los datos se quiere conservar. Si los datos van a ser usados para construir un modelo predictivo basado en preferencias individuales, esas preferencias no pueden eliminarse ya que sería lo mismo que no compartir los datos. Entender para qué hacen falta los datos y cuál es su valor es clave para encontrar soluciones adecuadas. De otra manera, lo que terminamos compartiendo no son datos que luego se convierten en información que brinda valor o ayuda a la toma de decisiones, sino información aislada que no genera ningún valor. En cierta forma es lo mismo que no compartirlos[7].

2- Identificar riesgos

¿Cuáles son los riesgos potenciales que podrían traer los datos a compartir? En el caso de los taxis, ¿identificar a los pasajeros? ¿A los choferes? Saber qué películas mira una persona, ¿puede brindar información delicada sobre ella? Hacer públicos los registros hospitalarios, por ejemplo, puede afectar mucho la vida de una persona y generar problemas personales y laborales. En el caso de Netflix, quedó claro que incluso información que en principio se consideraría poco sensible puede serlo: en la demanda presentada se describe con detalle por qué la información sobre preferencias en películas puede llevar a inferencias sobre inclinaciones sexuales o asociaciones políticas cuya difusión puede dañar la vida de las personas involucradas.

Por eso, analizar los riesgos potenciales de la publicación, cruce y reidentificación de una base de datos es importante a la hora de priorizar y entender los escenarios de uso.

En muchos casos, como hemos mostrado, el riesgo surge al cruzar los datos con otras bases. El análisis no debe hacerse sólo de la base que se pretende publicar, sino de esa base en un contexto. Un punto importante: los datos desidentificados no están protegidos por ninguna regulación. Es decir que se puede publicar una base desidentificada y aun así estar en una situación de riesgo. Algunas jurisdicciones han comenzado a tomar medidas que tienen en cuenta estos riesgos: por ejemplo, desde 2009, el estado de California considera el código postal de una persona como dato personal en la ley que protege esos datos, pero estas soluciones son siempre parciales y acotadas y no tienen en cuenta el contexto de uso.

Por otro lado, la privacidad no es una preocupación general de la población. Varios estudios han mostrado que los jóvenes se preocupan mucho menos por la privacidad que los mayores, y que un porcentaje importante de la población está dispuesta a ceder su información a cambio de beneficios. A la par, una nueva noción de privacidad está emergiendo: muchos jóvenes están acostumbrados a no compartir su información en internet, e incluso usan navegadores seguros y privados y borran información personal o que permita identificar su IP.

En este contexto, esta identificación de riesgos no es absoluta ni objetiva: está asociada a una preocupación de un grupo de ciudadanos o stakeholders. Más allá de que el derecho a la privacidad debe ser algo uniforme y no depender de la conciencia de un determinado grupo o persona, resulta

---

7 De hecho, en muchos casos se publica o comparte información "agrupada" para evitar problemas de privacidad, que luego no puede usarse para toma de decisiones con el nivel de granularidad adecuado. Esta propuesta apunta a compartir datos para lograr mejores resultados manteniendo la privacidad necesaria, no evitando el compartir. Este balance ha sido estudiado con detalle, ver por ejemplo Aggarwal y Yu (2008) y Dwork (2006).

importante entender el alcance que el riesgo de pérdida de privacidad pueda tener en cada situación. Los datos pueden ser considerados sensibles por diferentes razones: por revelar origen racial y étnico, opiniones políticas, convicciones religiosas, filosóficas o morales, afiliación sindical o información referente a salud o vida sexual.

3- Identificar soluciones

Al tratarse de un tema relativamente nuevo y con ramificaciones en diferentes disciplinas, no existe un *compilado de buenas prácticas*, aunque sí hay propuestas de técnicas para anonimizar en algunos contextos, o para compartir datos sin revelar información personal. Los trabajos de privacidad diferencial[8] brindan un marco teórico para entender las posibilidades y límites de la anonimización estadística.

En muchos casos, buscar anonimizar completamente un conjunto de datos es muy difícil, costoso o incluso imposible. De hecho, Dwork y Roth (2014), demuestran "la ley fundamental de recuperación de la información": responder con precisión un número grande (o no limitado) de consultas a una base de datos termina destruyendo cualquier noción de privacidad.

Una alternativa a la privacidad diferencial es la k-anonimización (k-anonymity, ver Sweeney (2002)) un conjunto de datos cumple con ser k-anónimo para un número k, si el mayor grado de individualización que logro usando los datos es un grupo de k individuos. Para dar un ejemplo: si la única información que brindo sobre un grupo de individuos es que asisten a determinada escuela, ese conjunto de datos es k-anonimizado para el número mínimo de personas de una escuela. Es decir, si la escuela con menos estudiantes tiene 34, usando esta información, aun en el caso de informar que el individuo asiste a esa escuela, queda "agrupado" dentro de esos 34 alumnos, no puedo diferenciarlo. Por lo tanto, este conjunto es 34-anónimo.

De acuerdo con el tipo de datos y los riesgos identificados en el paso anterior, se puede elegir un grado de k-anonimización que sea razonable lograr y a la vez que sea adecuado para la situación de riesgo identificada. La k-anonimización es un concepto, no una técnica[9].

Más allá de los algoritmos, que proveen un modo sistemático de tratar los datos, existen técnicas que pueden considerarse a la hora de transformar un conjunto de datos para evitar potenciales reidentificaciones:

- Eliminación. Siempre es posible eliminar una o más columnas o variables.

- Generalización. Agrupar datos, brindar información sólo a nivel de grupo (distribución estadística y parámetros generales), cambiar su nivel de detalle (granularidad). Por ejemplo, reemplazar el ingreso de un grupo familiar por el decil al que pertenece.

- *Hashing* y encriptado. Aplicar a un dato una función "de un solo sentido", es decir, una función cuya inversa es muy costosa de calcular. Por ejemplo, reemplazar una descripción con un código, que permita identificar coincidencias pero no diferenciar entre ellas; o reemplazar un identificador

---

8 La privacidad diferencial es una técnica que permite que cada consulta realizada a una base de datos no se responda con información exacta, sino que se introduzca cierta cantidad de ruido para evitar que, mediante múltiples consultas, se pueda usar una estrategia de reidentificación. Ver Dwork (2006) y Dwork y Roth (2014).

9 Existen numerosos algoritmos y técnicas para transformar un conjunto de datos (Simi et al, 2017), entre ellos algoritmos como Incognito, Saramati, Data-fly.

con una versión encriptada, que permita identificar a los individuos pero a partir del cual no pueda recuperarse la versión original.

- Distorsionar los datos ("agregar ruido"), mezclar, confundir. Esto suele ser usado en gráficos, pero se puede aplicar en varios contextos. Algunas técnicas agregan ruido aunque mantienen las propiedades relevantes de los datos originales, lo que permite aplicar técnicas de análisis de datos sin revelar información[10].

4- Identificar ataques y problemas

El último paso es la revisión de la solución propuesta en modalidad de testeo, considerando posibles ataques o problemas. Por ejemplo, como mencionamos en el ejemplo de los taxis, una solución de *hashing* en un dominio reducido puede no ser útil. Si bien es muy difícil "volver atrás" de un código de *hashing* arbitrario, en dominios reducidos puede no serlo. A modo de ejemplo, de un archivo de claves encriptadas no es posible recuperar las claves originales, pero con la información adicional de que las claves son números de seis cifras, volver atrás es sumamente sencillo. Otros potenciales problemas pueden estar dados por los datos, por ejemplo si son conjuntos de datos de alta dimensionalidad (donde hay muchas variables asociadas a cada registro, lo que aumenta la probabilidad de combinar la base con datos externos y entre sí), la homogeneidad de los datos[11] y la posible predicción de un valor sensible (en este caso, un valor que permite identificar inequívocamente a un individuo, organización, grupo o elemento que se quiere anonimizar) en función de valores no sensibles[12].

En algunos casos, incluso si no se puede identificar a quién pertenece cada registro, el hecho de poder determinar la pertenencia a una base es de por sí un problema.

El factor clave en este paso es establecer un criterio adversarial. Es decir, se debe considerar que se enfrenta a un adversario y revisar la solución desde el punto de vista de un posible ataque. Además, esta visión es dinámica: aplicar una solución o una técnica seguramente genere un cambio de estrategia en el potencial ataque. Esta situación dinámica es una característica común de los planteos adversariales. En un planteo adversarial se considera que existe un adversario (puede ser una persona o un grupo de personas) que intentará romper todos los intentos que hagamos por anonimizar los datos. Es decir, cada vez que usemos alguna técnica, irá mejorando sus ataques para hacer frente a cada nueva técnica que usemos. De esta manera, el problema de la reidentificación no se reduce a un problema estático, sino que existe una dinámica en la cual el adversario puede intentar varias jugadas y nosotros deberemos responder también en forma dinámica[13].

---

10 Ver por ejemplo el marco de trabajo SuLQ (Blum et al., 2005).

11 El problema de la homogeneidad se da cuando un valor sensible es idéntico (o cambia poco) en muchos registros. En esos casos, aun siendo k-anónimo uno puede predecir ese valor para todo el grupo y por lo tanto para cada individuo.

12 Para estos casos, existen técnicas específicas, que extienden el concepto de k-anonimización, como las nociones de l-diversidad y t-cercanía (l-diversity, t-closeness, ver Li et al., 2007), cuya descripción va mucho más allá del alcance de este artículo. En El Emam (2009) se puede ver la aplicación de algunas técnicas estadísticas útiles a la hora de realizar el análisis.

13 En el caso de reidentificación, este concepto se asocia a "ataques de reconstrucción", formalizados en Dwork y Roth (2014).

Mejores prácticas

Metodología de gestión eficiente de datos sin poner en riesgo la privacidad

Tabla 1

Identificar datos	¿Con qué bases públicas se relacionan estos datos? ¿Con qué otras bases pueden cruzarse? ¿Cuál es el ciclo de vida de los datos de esta base de datos? ¿Qué campos de mi base de datos pueden actuar como identificadores y cuáles deben ser borrados, modificados o eliminados? ¿Qué información es confidencial o sensible y puede recuperarse en función de uno o más campos de esta base de datos?
Identificar riesgos	¿Cuáles son los riesgos potenciales de compartir estos datos?¿Cuáles son los escenarios de uso de los datos?
Identificar soluciones	Algunas buenas prácticas - Privacidad diferencial - K-anonimización - Eliminación de variables - Generalización - *Hashing* y encriptado - Distorsión de datos
Identificar ataques y problemas	Revisión de la solución propuesta. Establecer un criterio adversarial para anticipar problemas.

Fuente: Elaboración propia.

Por supuesto, aplicar esta metodología requiere un análisis mucho más detallado y una descripción de las técnicas y algoritmos mencionados. Como hemos intentado demostrar en este trabajo, la solución "no compartir" es la peor estrategia, aunque en apariencia sea la más efectiva. Agrupar los datos, como también observamos aquí, impide adoptar procedimientos de inteligencia artificial o llevar adelante un análisis avanzado de la información. No hay una solución técnica única que funcione en todos los casos. La estrategia más conveniente, entonces, es conocer y aplicar metodologías de trabajo que están disponibles y que permiten el uso de los datos para la mejor toma de decisiones, sin poner en riesgo la privacidad o hacer pública información que no debe ser revelada. Buscar alternativas para compartir los datos con un nivel de detalle suficiente como para poder utilizar algoritmos avanzados es clave a la hora de poner en valor la información y obtener de ella todo lo que tiene para dar.

Referencias

- Aggarwal, C y P. Yu (2008). "A General Survey of Privacy-Preserving Data Mining Models and Algorithms". En *Privacy-Preserving Data Mining – Models and Algorithms*, Springer.

- Bennett, J. y S. Lanning (2007). "The Netflix Prize". Proceedings of KDD Cup and Workshop 2007, workshop coorganizado por ACM SIGKDD y Netflix, KDD-2007, San José, California.

- Blum, A., Dwork, C., McSherry, F. y Nissim, K. (2005) "Practical Privacy: The SuLQ Framework", in 24th ACM SIGMOD *International Conference on Management of Data / Principles of Database Systems*, Baltimore.

- De Montjoye, Y-A., Hidalgo, C., Verleysen, M. et al. (2013). "Unique in the Crowd: The privacy bounds of human mobility". *Sci Rep 3*, 1376. Disponible en https://doi.org/10.1038/srep01376

- Dwork, C. (2006). "Differential Privacy". En: Bugliesi M., Preneel B., Sassone V., Wegener I. (eds) Automata, Languages and Programming. ICALP 2006. Lecture Notes in Computer Science, vol 4052. Springer, Berlin, Heidelberg. Disponible en: https://doi.org/10.1007/11787006_1

- Dwork, C. y A. Roth (2014). "The Algorithmic Foundations of Differential Privacy". Foundations and Trends in Theoretical Computer Science. Vol. 9, no. 3–4, pp. 211-407. Disponible en: https://www.nowpublishers.com/article/Details/TCS-042.

- El Emam, K., Dankar, F., Vaillancourt, R., Roffey, T y Lysyk, M. (2009). "Evaluating the Risk of Re-identification of Patients from Hospital Prescription Records". Canadian Journal of Hospital Pharmacy, CJHP – Vol. 62, No. 4.

- Li, N., Li, T. y Venkatasubramanian, S. (2007). "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity". IEEE 23rd International Conference on Data Engineering, pp. 106–115.

- Mitnick, K. (2017). The Art of Invisibility: *The World's Most Famous Hacker Teaches You How to Be Safe in the Age of Big Brother and Big Data*, Londres, Hachette.

- Narayanan, A. y V, Shmatikov, V. (2008). "Robust de-anonymization of large sparse datasets". In *Proceedings - 2008 IEEE Symposium on Security and Privacy,* SP (pp. 111-125). [4531148] (Proceedings - IEEE Symposium on Security and Privacy). Disponible en: https://doi.org/10.1109/SP.2008.33

- Narayanan, A. y V. Shmatikov (2009). "De-anonymizing Social Networks," 30th IEEE Symposium on Security and Privacy, Oakland, California, pp. 173-187. Disponible en: https://www.cs.utexas.edu/~shmat/shmat_oak09.pdf.

- Simi, M.S. et al. (2017). "An Extensive Study on Data Anonymization Algorithms Based on K-Anonymity", IOP Conf. Ser.: Mater. Sci. Eng. 225 012279.

- Sweeney, L. (2000). "Simple Demographics Often Identify People Uniquely", Data Privacy Working Paper 3, Carnegie Mellon University, Pittsburgh. Disponible en https://dataprivacylab.org/projects/identifiability/paper1.pdf

- Sweeney, L (2002). "k-anonymity: a model for protecting privacy". International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), pp. 557-570.

- Sweeney, L., Abu, A., and Winn, J. (2013). "Identifying Participants in the Personal Genome Project by Name". White Paper 1021-1, Data Privacy Lab, Universidad de Harvard.

- Wondracek, G., Holz, T., Kirda, E. y Kruegel, C. (2010). "A Practical Attack to De-anonymize Social Network Users", IEEE Symposium on Security and Privacy, Berkeley/Oakland, California, pp. 223-238. Disponible en: https://ieeexplore.ieee.org/document/5504716

- Zang, H. y J. Bolot (2011). "Anonymization of location data does not work: A large-scale measurement study". Conferencia Internacional Mobile computing and networking 17, 145–156.

-

Acerca del autor

Daniel Yankelevich

Director del Área de Datos de Fundar.

Informático, PhD de la Universidad de Pisa, realizó su postdoctorado en Carolina del Norte, EEUU. Docente universitario, con trayectoria en el sector privado y en proyectos de investigación.

Modo de citar

Yankelevich, Daniel (2021). Anónimos pero no tanto: cómo hacer una gestión de datos eficiente sin poner en riesgo la privacidad. Buenos Aires: Fundar. Disponible en https://www.fund.ar/