

http://elpais.com/elpais/2015/01/29/ciencia/1422520042_066660.html

Cuatro compras con la tarjeta bastan para identificar a cualquier persona

Los patrones de uso de las tarjetas permiten descubrir la identidad del 90% de una muestra de 1,1 millones de personas anónimas, según demuestra un estudio del MIT

- **Eres lo que ‘te gusta’**

MIGUEL ÁNGEL CRIADO 29 ENE 2015 - 19:59 CET



Los datos anónimos de las compras permiten identificar a las personas, en especial a mujeres y a los que tienen más ingresos. / Yves-Alexandre de Montjoye/MIT

Cuatro gestos tan triviales como pagar el billete de metro, la comida del mediodía, las zapatillas en una tienda deportiva o las entradas al cine permiten identificar a casi cualquier persona. Aunque no se sepa el nombre o el número de cuenta, un estudio con datos de compra de 1,1 millones de personas revela la identidad en más del 90% de los casos. Es el poder de los metadatos y el *big data*.

Cuando estalló el *caso Snowden*, en Estados Unidos se produjo un gran escándalo con uno de los programas de espionaje de la NSA que recopilaba millones de llamadas telefónicas. Las autoridades estadounidenses aclararon enseguida que no espiaban el contenido de las conversaciones en sí sino metadatos como quién llamaba a quién, a qué hora o durante cuánto tiempo. Google o Facebook también los usan para mejorar sus servicios u ofrecer publicidad más personalizada. En principio, la agregación de este tipo de datos de forma anónima en grandes bases

no plantearía una gran amenaza a la privacidad de las personas. Esa presunción se ha demostrado ahora falsa.

Un grupo de investigadores del [Media Lab](#) del Instituto Tecnológico de Massachusetts (MIT) ha diseñado un par de algoritmos matemáticos que permiten identificar a una persona basándose en sus hábitos de compra. Consiguieron que un gran banco de un país de la OCDE (por razones obvias, no dicen ni qué banco ni qué país) les dejara aplicar sus algoritmos a una base de datos con las transacciones de pagos electrónicos de 1,1 millones de clientes en unas 10.000 tiendas durante los meses de enero y marzo de 2014.

Las bases con inmensas cantidades de datos anonimizados sirven para mostrar
la unicidad del ser humano

"Con un promedio de cuatro transacciones, el día y la tienda, basta para identificar de forma exclusiva a las personas en el 90% de los casos", dice el investigador del MIT y coautor de la investigación, Yves-Alexandre de Montjoye. "La lógica que subyace en esto reside en que muchas personas compran algo en una determinada tienda (Mango, por ejemplo) un día determinado (pongamos, ayer). Sin embargo, solo algunas de ellas también comprarán en determinado H&M ese mismo día. Y aún menos irán a comer al día siguiente en la misma zona. En cuanto sepas cuatro lugares o tiendas y días, el 90% de las veces hay una y solo una persona en toda la base de datos que compra algo en cuatro lugares en esos cuatro días", explica.

De los archivos del banco usados para el estudio, los investigadores solo disponían del día (renunciaron a registrar la hora, lo que habría afinado aún más sus resultados) y el comercio donde se realizó la compra. Como es lógico, cada transacción lleva también un identificador de usuario en forma de una combinación alfanumérica de 8 dígitos, que permite a la entidad pasar el cobro al que hizo la compra. A ellos les sirvió para confirmar sus porcentajes de acierto.

Hay un cuarto metadato que parecería irrelevante pero, en realidad, da pistas extra para identificar a las personas. Se trata del importe de la compra. Los autores del estudio, publicado en [Science](#), agruparon los importes en intervalos. No necesitaron el precio exacto para ampliar las posibilidades de volver a identificar al comprador. "Analizamos también qué pasaba si sabemos el precio aproximado de lo que has comprado. Por ejemplo, unos 30 dólares en Mango, 20 en H&M, 7 para la comida... Esto, en la práctica, hace a las personas incluso más únicas. Aquí, con solo tres puntos (tiendas, día y precio) tienes un 94% de posibilidades de volver a identificar a un individuo", comenta el investigador francés.

Es la paradoja de este tiempo. Las bases con inmensas cantidades de datos anonimizados sirven para mostrar la unicidad del ser humano. Como dice de Montjoye, no se trata tanto de la predictibilidad de los humanos, "si no más sobre cómo nuestra conducta (y nuestros patrones de compra) nos hacen únicos en comparación con los demás".

Las mujeres y los ricos, más identificables

Los autores del estudio también investigaron el efecto del género o el nivel de ingresos en la probabilidad de la reidentificación. Aunque las mujeres solo eran el 24% de la muestra analizada, comprobaron que ellas son, de promedio, 1,2 veces más identificables. El mismo fenómeno sucede con el nivel de ingresos. Los compradores con mayor promedio de gasto tienen 1,7 veces más probabilidades de ser identificados.

El objetivo del estudio no era explicar por qué las compras permiten identificar a las personas, pero los investigadores jugaron con un par de posibles variables para explicar estas diferencias. Vieron que la forma en la que uno reparte el tiempo entre las distintas tiendas era el mejor predictor para saber si el comprador era mujer o de altos ingresos. Estos dos grupos muestran un patrón de mayor diversidad a la hora de comprar que los hombres o las personas de menores ingresos.

Los metadatos pueden ser datos personales", dice el director de la AEPD

Aunque los científicos del MIT sostienen que habrá que hacer más estudios en otros ámbitos, ellos apuestan por que los datos de navegación en internet, los movimientos bancarios o los datos de transporte y movilidad también tienen un alto grado de unicidad y permitirían de forma inequívoca distinguir a una persona de otra. En un país como España, donde según las [estadísticas del Banco de España](#) hay casi 70 millones de tarjetas de crédito y débito con las que se hacen operaciones de unos 100.000 millones de euros, la sola idea de se pueda identificar a alguien por el uso que hace de su tarjeta, espanta.

Por fortuna, los autores del estudio tuvieron que firmar un acuerdo de confidencialidad con el banco para poder hurgar en su base de datos, que se suponen a buen recaudo. Pero, como concluyen en su trabajo, el problema fundamental que revela esta investigación es que las leyes sobre [privacidad](#) descansan sobre una premisa que ellos han demostrado que no es cierta. Por muy dura que sea la norma, esta solo es aplicable a los datos personales, es decir, aquellos que permiten identificar a un individuo. Los más obvios son su nombre, su

cara, su dirección o su teléfono. Pero, ¿qué pasa con los metadatos como la compra de unas zapatillas en una tienda determinada?

"Los metadatos pueden ser datos personales y muchas veces lo son", recuerda el director de la [Agencia Española de Protección de Datos](#), José Luis Rodríguez. "Para que no sean datos personales tienen que ser anónimos, con una disociación irreversible", añade. Si, como en esta investigación, se puede hacer el camino inverso desde los metadatos a la identidad de la persona, entonces sí se le aplicaría la legislación sobre privacidad. Para Rodríguez, el problema de fondo es que "en la medida en que existe cada vez más información disponible, se debilita la anonimización porque hay más posibilidades de combinar y, por lo tanto, de identificar o individualizar a la persona".

Tecnología y privacidad, condenados a entenderse

M.Á.C

La revista [Science](#) incluye hoy un especial sobre las conflictivas relaciones entre la tecnología y privacidad. A lo largo de una serie de artículos se repasan nuevas amenazas como el reconocimiento facial o las cada vez más retorcidas maneras que encuentran empresas y gobiernos para aprovecharse de los datos de los ciudadanos. También aparece un revelador artículo sobre el derecho al olvido.

La investigadora del Instituto Tecnológico de Worcester [Susan Landau](#) mantiene en uno de los artículos que las personas han perdido la capacidad de proteger sus datos personales y su privacidad. Apoyadas en la facilidad que tienen las máquinas para establecer conexiones entre los datos, las empresas y gobiernos tienen cada vez más fácil recopilar grandes cantidades de información y sacarle provecho. Menciona, por ejemplo, una vieja investigación suya que demostró que un internauta medio necesitaría 244 horas para leerse todas las políticas de privacidad que hay en las páginas que se visitan. Para ella, los viejos métodos para proteger la privacidad ya no sirven.

El director de la AEPD, José Luis Rodríguez, no cree que haya que tirar todo a la basura y renunciar al derecho a la privacidad porque sea cada vez más complicado ejercerlo. "Mantener una esfera de privacidad es imprescindible para el desarrollo de la persona", recuerda.

En lo que sí está de acuerdo es en que los riesgos son cada vez mayores. Por eso, coincide con Landau en que, además de una legislación firme, hacen falta soluciones tecnológicas que protejan los datos personales. "No es admisible que la tecnología evolucione solo por el lado de la recopilación de los datos, debería ir pareja con sistemas que los protejan", dice.

En otro de los trabajos, se revisa el impacto que está teniendo el llamado derecho al olvido, tras la resolución de la Justicia europea sobre un caso español. En el artículo, el profesor de la Universidad de Georgetown, Abraham Newman, desmonta dos mitos en los que se apoyan quienes critican la mera existencia del derecho al olvido en internet.

Por un lado, niega que la desindexación de información personal de los buscadores dañe a la libertad de expresión y el derecho a la información. Al fin y al cabo, la información no se borra, solo se oculta de los ojos de Google. Por el otro, rechaza, como ha sostenido Google, que el trabajo de eliminar unos miles de enlaces pueda dañar la salud económica de la compañía. Y da un dato: En los primeros cinco meses de aplicación de la resolución judicial, el buscador ha revisado 180.000 peticiones de retirada, aceptando el 40%. Mientras, en un solo mes, el de diciembre pasado, tuvo que atender peticiones de retirada de nueve millones de enlaces por posible infracción de derechos de autor.