Robin Worrall

LA DOBLE HUELLA DEL *BIG DATA*

PRIVACIDAD Y SOSTENIBILIDAD, DOS GRANDES RETOS DE INTERNET

Xavier Duran

El 5 de febrero del 2021 los periodistas de investigación Charlie Warzel y Stuart A. Thompson publicaron en la web de *The New York Times* un artículo donde explicaban que alguien les había proporcionado datos de miles de seguidores de Donald Trump que el 6 de enero habían protagonizado una invasión violenta del Capitolio en Washington (Warzel y Thompson, 2021). Y eso, a pesar de que los números de móvil no llevan asociada una identidad, había permitido trazar el origen de muchas personas hasta el lugar de donde venían e incluso averiguar sus nombres, direcciones y cuentas de redes sociales.

Dos años antes, a los dos autores ya les habían llegado datos de localizaciones de más de doce millones de ciudadanos norteamericanos, lo que permitía rastrear sus movimientos. Pero en el caso del Capitolio, cada localización iba asociada a un código ID –o *mobile advertising identifier*–, que es único para cada usuario y está ligado a su móvil inteligente o a su tableta. Solo había que cruzar este ID con otras bases de datos para obtener una inmensa cantidad de información sobre cada persona.

«En 2018 Facebook tenía datos personales para llenar, de media, 400.000 documentos de Word con cada usuario»

Es posible que mucha gente considere útil e incluso necesaria la posibilidad de utilizar un ID para perseguir delitos. Pero, tal como explican los dos periodistas, este ID lo usan numerosas empresas, instituciones y entidades, incluidos bancos y fondos de inversión, que pueden obtener, así, una gran información sobre cualquier ciudadano.

Nuestra actividad diaria lo hace posible. Casi todo el mundo navega por Internet y hace búsquedas y una gran mayoría de ciudadanos realizan, en algún momento, una compra a distancia. Pero, aunque no se haga nada de esto, todos estamos en muchas bases de datos, desde padrones municipales a entidades deportivas o culturales, pasando por servicios médicos, entidades de ahorro y archivos fiscales.

Y muchos, además, tenemos actividad en las redes sociales. Y esto da muchísima información. Según explicaba Mat Travizano en septiembre de 2018 a *Entrepreneur* (Travizano, 2018), Facebook tenía en aquel momento datos personales para llenar, de media, 400.000 documentos de Word con cada usuario. Y Google podía llenar cerca de tres millones para cada internauta.

¿Cómo se ha llegado a esta situación en la que estas empresas tienen tanta información? ¿Y qué impacto tiene en nuestra privacidad y seguridad?

LOS DATOS, MODELO DE NEGOCIO

Para comprenderlo tenemos que ir a finales de los años noventa del siglo xx. Habían surgido multitud de empresas tecnológicas, las llamadas *puntocom*. Las expectativas de crecimiento eran altas y, probablemente, también demasiado optimistas. Y el capital riesgo se lanzó. Eso hizo que subieran de forma meteórica, en bolsa, unas empresas que, en la mayor parte, no tenían ni siquiera un modelo de negocio. Ganaban cuota de mercado, pero con muchos servicios gratuitos y, por lo tanto, no generaban beneficios. Hacia el año 2000, la situación de la mayoría de estas empresas era muy delicada, por no decir desesperada, y tenían muchas más deudas y promesas de futuro que realidades.

Y fue entonces cuando empresas como Google vieron qué potencial tenían para monetizar sus servicios: los datos. A pesar de que los móviles inteligentes no estaban muy extendidos y la capacidad de los ordenadores y de los algoritmos para procesar grandes cantidades de datos eran limitadas, el uso de la información como materia prima funcionó. Así, Google ingresó 19 millones de dólares en 2000, pero el 2001 ya eran 86 millones; en 2002, 440 millones; en 2003, 1.500 millones, y en 2004 ingresó 3.200 millones de dólares. Un aumento del 3.590 % en cuatro años. En 2020 ingresó 146.920 millones solo en publicidad.

La clave tenía que ser la publicidad personalizada. Con todos los datos que había acumulado, Google podía averiguar los intereses de los usuarios. Y si una empresa quería enviar anuncios, Google podía, sin revelar identidades, dirigirla de forma que, en vez de hacer campañas indiscriminadas, ya se podía ir a los grupos más susceptibles de hacer caso a aquellos anuncios. Ofrecer equipación de golf a los que se interesan por el golf, viajes al Caribe a personas que suelen buscar información sobre las playas del Caribe y estancias en *campings* a los que se acaban de comprar una caravana.

De entrada, esto no parece malo. Mucha gente pensaba, y todavía piensa, que si recibe publicidad personalizada

le hacen un favor, porque le eligen la que le puede interesar. Y si aun así no la quiere, no pierde nada. Algunas molestias, a lo sumo.

El problema aparece cuando la información ya no sirve solo para enviar publicidad personalizada, sino para decidir si se permite a aquel usuario contratar un seguro médico o de vida, si es lo bastante solvente como para concederle un crédito o una hipoteca o incluso si merece la pena contratarlo para un trabajo.

Tener datos es tan valioso que muchas empresas ya lo sitúan como un objetivo, al margen de si fabrican neveras o coches o si alquilan vehículos o apartamentos. En noviembre del 2018, el CEO de Ford anunció el objetivo de monetizar y vender los datos que recogía de los cien millones de personas que conducían sus vehículos (Sadowski, 2020, p. 31). Aparte de sus hábitos y rutas, quizás podría averiguar si conducían de forma lo bastante prudente o por carreteras suficientemente seguras, lo que interesa mucho a compañías de seguros.

Sobre esto ya hemos hablado en MÈTODE en otra ocasión (Duran, 2018), pero las posibilidades de saber cosas de nosotros crecen de forma desmesurada. Además, Google, Facebook y Amazon, que se reparten la mayor parte del mercado publicitario digital mundial –se calcula que en 2018 entre Google y Facebook controlaban el 84 % del mercado mundial exceptuando China–, saben quiénes somos, porque para utilizar sus servicios nos tenemos que inscribir y dar como mínimo nuestro correo electrónico.

Con datos del 2018, Amazon tuvo 6.720 millones de dólares de ingresos en publicidad. Parece poco si lo comparamos con los 38.370 millones de Facebook y los 83.680 de Google. Pero pensamos que, en teoría, la mayor parte de sus ingresos tienen que ser por vender productos, no por anuncios. Pero la publicidad le da más margen que las ventas.

DEDUCCIONES PELIGROSAS E IDENTIFICACIÓN CON UNA FOTO

Y ya no hablemos, a pesar de su importancia, de la filtración o venta directa de datos. A principios de abril de 2021, se filtraron datos personales de más de 553 millones de usuarios de Facebook. Más allá de eso, con millones de datos y algoritmos cada vez más sofisticados, se pueden hacer muchas más deducciones –que pueden ser acertadas o no– e incidir en aspectos más íntimos. Es lo que le pasó a Ángel Cuevas, profesor de la Universidad Carlos III de Madrid. Estaba en una reunión de trabajo en Barcelona y recibió a través de Facebook publicidad que le invitaba a conectar con la comunidad gay y reservar un apartamento «con gente como tú».

Cuevas pensó cómo interpretaba Facebook si era o no homosexual si él no había dado nunca información sobre su orientación –se supone que Facebook lo hacía, acertadamente o no, a partir de otros datos–. Pero, además, se preguntó por qué la empresa permitía que algunos anunciantes le enviaran publicidad basada en su hipotética orientación sexual.

A partir de aquí, él y su equipo investigaron y descubrieron que en Arabia Saudí, por ejemplo, unas 500.000 personas tenían la etiqueta de homosexual. En países donde eso puede significar condena de prisión o incluso de muerte eso es muy peligroso, porque, aunque Facebook no revele identidades, en comunidades relativamente pequeñas identificar individuos concretos no es tan difícil (García et al., 2018). De hecho, una práctica para estudiantes avanzados de tratamiento de datos puede ser tomar de forma aleatoria una persona cualquiera de todo el mundo y tratar de averiguar el máximo de cosas sobre ella (Véliz, 2020, p. 65).

Es cierto que el Reglamento Europeo de Protección de Datos, en pleno vigor desde mayo del 2018, proporciona protección a los ciudadanos, y que para aceptar que desde una web tengan acceso a nuestros datos tenemos que dar el consentimiento explícito. Pero también lo es que las condiciones de los famosos «acepto» son a menudo muy

Tyler Merbler

Anete Lusina

Para buena parte de la población, utilizar las redes sociales o hacer una compra en línea ya forma parte de la rutina. Todas estas actividades que realizamos en Internet proporcionan mucha información a empresas que han convertido los datos en materia prima con la que conseguir beneficios económicos millonarios.

«Tener datos es tan valioso que muchas empresas ya lo sitúan como un objetivo, al margen de si fabrican neveras o coches o si alquilan vehículos o apartamentos»

Gracias al *mobile advertising identifier*, un código de identificación único para cada usuario vinculado a su dispositivo, los datos de miles de los seguidores de Donald Trump que participaron en el asalto del Capitolio en enero de 2021 acabaron en manos de dos periodistas de investigación de *The New York Times*. Solo cruzando este código con otras bases de datos se obtuvo una inmensa cantidad de información sobre cada persona.

complicadas y que si queremos acceder a un servicio o a una web tendemos a decir que sí a todo lo que nos proponen. Incluso se han sancionado entidades bancarias por faltas graves, como no facilitar al cliente la posibilidad de utilizar los servicios sin acceder a proporcionar datos. Además, empresas no europeas quizás no siempre respeten este reglamento. En abril de 2021, en el Reino Unido se anunció una demanda contra TikTok, una *app* china para compartir vídeos cortos, por recopilar, supuestamente de forma ilegal, datos personales de millones de niños, como números de teléfono, ubicación de la conexión e, incluso, datos biométricos. En 2019, la firma china recibió una multa de 5,7 millones de dólares de la Comisión Federal del Comercio de Estados Unidos por un mal uso de los datos.

Además, cuando entramos en una web puede ser que, en realidad, estemos cediendo datos a numerosas webs con las que la primera tiene acuerdos. Son las famosas *cookies* de terceros, que permiten seguir nuestro rastro y usar los datos. Esto y otros muchos problemas con la privacidad están expuestos en el reportaje «Tot el que sabem de tu», emitido en noviembre de 2020 en el programa *30 minuts* de TV3 (Duran et al., 2020).

En el mismo reportaje se explica el caso de la artista e investigadora Joana Moll. Cuando desarrollaba su proyecto *The dating brokers* ("Los marchantes de citas") descubrió, de forma casual, que en Internet se vendían datos de clientes de páginas de contactos. Moll compró un millón de perfiles de usuarios de todo el mundo por 136 euros. Incluía unos 600.000 perfiles de hombres y unos 300-400.000 perfiles de mujeres, con direcciones de correo electrónico, nombres de usuario, fechas de nacimiento, orientación sexual, descripciones muy detalladas del físico y la personalidad, si tenían hijos, si fumaban, si tomaban drogas...

Pero, además, descubrió que estas webs pertenecían a empresas que, a su vez, formaban parte de grandes grupos. Como se daba el consentimiento para ceder datos a todas las empresas del grupo, alguien que hubiera entrado en una web de contactos podría haber acabado proporcionando datos sensibles a más de 700 empresas.

Las posibilidades del uso de datos personales y los riesgos que significan son numerosos (Sadowski, 2020; Véliz, 2020). Y probablemente no hay lugar donde esconderse, como demuestra el caso de Clearview, también expuesto en el reportaje anterior. Se trata de una aplicación creada en los Estados Unidos que, a partir de la imagen de una persona cualquiera y gracias al reconocimiento facial, proporciona otras fotografías de esta misma persona, incluso de años atrás, que están en Internet, y las direcciones de las webs donde se encuentran. A partir de aquí, el usuario puede ir a las webs y elaborar un perfil sobre cualquier individuo.

La base de datos de Clearview tiene unos tres mil millones de imágenes capturadas sin conocimiento de los que salen en ellas. Ya ha recibido varias demandas, pero si bien lo que digan los jueces es importante, lo que queremos destacar es que cada vez hay más tecnologías que permiten encontrar montones de información de cualquier persona, por discreta que sea, porque todos podemos estar etiquetados en fotografías, a menudo sin saberlo.

EL NEGOCIO Y LAS EMISIONES DE LOS SERVIDORES

Antes hemos explicado que Amazon obtiene una buena parte de sus ingresos por publicidad. De hecho, de los 280.000 millones que ingresó en 2019 –con 11.500 millones de beneficios–, el 50,37 % provenían de las ventas *online*. Eso significa poco más de la mitad; la otra no viene de estas ventas. Aproximadamente un 6 % provenía de tiendas físicas, pero más importante era el 19,17 % de ventas de terceros. Amazon les proporciona toda la infraestructura digital y se queda un porcentaje importante de las ventas. De hecho, en las ventas propias Amazon ajusta mucho el precio y establece márgenes pequeños, pero tiene una estrategia: cobra rápidamente de sus clientes y paga a plazo más largo a sus proveedores. Así genera mucha liquidez.

En 2019 también, un 12,48 % de los ingresos provenían de Amazon Web Services (AWS). Se trata de un servicio para que las empresas puedan tener sus archivos y programas en la nube y no tengan que gastar en servidores o bases de datos propios.

Este parece uno de los grandes negocios del futuro. Se calcula que en 2019 en la nube había 45 zettabytes de datos –45.000 trillones de bytes, equivalente a más de 7.000 millones de años de vídeo de alta definición– y que en 2025 habrá 175 zettabytes. Esto significará un mercado de cerca de 600.000 millones de euros. Y actualmente AWS controla algo más del 40 % del mercado. Azure, de Microsoft, tenía, en 2019, el 29,4 % y Google Cloud, el 3 %.

Los grandes servidores de Amazon muestran la gran visión de su propietario y director ejecutivo, Jeff Bezos. A pesar de que hablar de nube hace pensar a mucha gente que los datos se pasean por la atmósfera hasta que alguien los captura en su ordenador, esta nube la forman estructuras físicas, que guardan programas y datos, y miles de kilómetros de fibra óptica por donde viajan. Quizás enviemos un correo electrónico al vecino del lado y cuando le llega ha pasado por un servidor que está cerca del Polo Norte. Los servidores –ordenadores donde funcionan programas accesibles desde diferentes puntos de la red– buscan en cada momento los caminos más adecuados por la red de fibras y cuando nos bajamos una canción o hacemos una compra *online* no sabemos muy bien por qué partes del planeta han pasado los bits que lo han permitido.

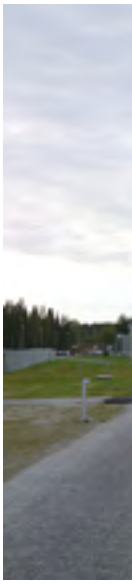
Pero eso significa un gran consumo de energía y muchas emisiones de CO_2, aunque no sean tan evidentes como los humos que salen por las chimeneas. Cuando tecleamos en el ordenador, en la tableta o en el móvil, lo tenemos puesto en marcha y consumiendo. Y también está consumiendo el dispositivo donde alguien recibirá el mensaje y el servidor de la web que estamos consultando. Y están consumiendo los servidores, puestos en marcha las 24 horas del día, porque Internet nunca duerme. No somos conscientes de ello, pero si Internet está siempre disponible y en cualquier momento podemos movernos por millones de webs o tener actividad en cualquier red social es porque estos grandes servidores están activos.

Dar el servicio es un buen negocio porque muchas empresas, incluso grandes, ya no gastan en infraestructuras propias sino que las alquilan –el propio Netflix es cliente de AWS.

Pero al margen de quién haga dinero, el negocio es nefasto para el medio ambiente. Los procesadores han ganado en eficiencia, pero la cantidad de información crece de forma exponencial. Algunos pronósticos apuntan que de cara al 2030 el conjunto de las tecnologías de la información y la comunicación pasaría a consumir el 21 % de la electricidad a escala global (Stern, 2020). Y casi un 40 % del consumo energético de los centros de datos se debe a su refrigeración. Por eficientes que sean, los servidores, con miles de procesadores, se calientan y este calor se tiene que disipar.

Una opción es instalar los servidores en zonas muy frías. Facebook tiene uno en Luleå, en el nordeste de Suecia, donde además puede disponer de grandes cantidades de energía hidroeléctrica y ahorrar costes y emisiones de CO_2.

Aun así, el consumo es muy elevado. Y si hablamos de emisiones equivalentes de CO_2, seremos conscientes del impacto. Algunos estudios calculan unas emisiones anuales de Internet de mil millones de toneladas, equivalentes a un 2,8 % de las emisiones totales –más que el sector de la aviación, responsable de un 2 %.



Cristopher Down

¿CUÁNTO CO_2 EMITE UN CORREO ELECTRÓNICO?

Según un estudio de la compañía energética británica OVO, los británicos envían diariamente 64 millones de correos electrónicos innecesarios, de esos que solo dicen «hola» o «gracias» o equivalentes (Tweedale, 2021). Cada correo hace emitir un gramo de CO_2. Por lo tanto, los correos que se podrían ahorrar son responsables de 23.475 toneladas de dióxido de carbono; en otros lugares se habla de 16.433 toneladas. Parecen muchas emisiones y equivalen a 22 vuelos de ida y vuelta entre Londres y Nueva York. Pero como las emisiones anuales totales británicas fueron de 435,2 millones de toneladas en 2019,

Cottonbro

A pesar de que el Reglamento Europeo de Protección de Datos ofrece protección a la ciudadanía, las empresas de fuera de Europa no siempre respetan este marco regulativo. En abril de 2021, el Reino Unido anunció una demanda contra TikTok, una *app* china de vídeos cortos muy popular entre los más jóvenes, por recopilar datos personales de millones de niños, como números de teléfono, ubicación de la conexión e, incluso, datos biométricos.

Los servidores que permiten a Internet funcionar de forma permanente son grandes estructuras físicas que tienen un gran consumo de energía y muchas emisiones de CO_2. Casi un 40 % del consumo energético de estos centros se debe a la necesidad de refrigerar los servidores funcionando a pleno rendimiento. Una solución que algunas empresas han encontrado es instalar los servidores en zonas muy frías; en la imagen, el centro de datos de Facebook en Luleå (Suecia).

dejar de enviar correos innecesarios solo las reduciría en un 0,0037 % (si bien todo suma, claro está).

Pero debemos tener en cuenta muchas prácticas más. Un estudio de la Universidad de Bristol estimaba que en 2016 el visionado de videos de Youtube produjo 11,13 millones de toneladas de CO_2 (Preist et al., 2019). Comparado con las emisiones mundiales, unos 35.000 millones de toneladas, vuelve a parecer poco. Pero equivalen a las producidas por una ciudad como Frankfurt o como Glasgow o por países como Luxemburgo o Zimbabue el mismo año. Y a YouTube tenemos que añadir toda la música descargada y todas las series, películas y documentales vistos en plataformas. Y todos los juegos dentro del llamado *gaming*. O bien hacer una búsqueda en un buscador, traducir un texto, enviar fotografías o presentaciones... Cuanto más complejo el material, más bits y más emisiones.

El problema no son solo las emisiones actuales, sino las perspectivas de crecimiento. Según cálculos de la Agencia Internacional de la Energía, los bitcoines provocan tantas emisiones como Nigeria o Uruguay. Para la plataforma Digiconomist todavía son más y superan las de Colombia y Bangladés. Según un artículo publicado a principios de abril en *Nature Communications* (Jiang et al., 2021), al ritmo actual, en China, en 2024, todo el proceso que rodea a las transacciones y validaciones en bitcoines producirá tantas emisiones de gases de invernadero como Italia o Chequia. Y en el ámbito interno, las emisiones se situarían en uno de los diez primeros lugares de 182 ciudades y 42 sectores industriales de China.

Eso es porque hacer del bitcóin una moneda virtual segura requiere una serie de cálculos complejos para garantizar la fiabilidad y mantener, al mismo tiempo, la privacidad, basándose en el llamado *blockchain* ("cadena de bloques"). Pero si pensamos que el bitcóin es una más de las monedas virtuales y que ahora solo representa el 0,4 % del dinero en circulación, podremos suponer el impacto que puede tener dentro de unos años.

¿Existen soluciones que no pasen por utilizar menos las herramientas digitales? De entrada, quizás tendremos que aprender a no derrochar recursos y hacer un uso más racional de ellos. Al mismo tiempo, podemos esperar que los tecnólogos encuentren soluciones más sostenibles y que la energía venga cada vez más de fuentes renovables. Si buscamos en Google «green Internet» veremos que el tema despierta interés: da 5.360.000.000 de resultados. La búsqueda sola ya ha producido más emisiones y visitar unas cuantas de estas webs todavía generará más. Si en alguna existen soluciones viables y eficientes, podemos dar las emisiones por buenas.

REFERENCIAS

Duran, X. (2018). Todo lo que saben de nosotros: ¿Se puede navegar con privacidad en el océano del *big data*? *Mètode*, *99*, 4–9. **https://metode.es/revistas-metode/article-revistes/todo-lo-que-saben-de-nosotros.html**

Duran, X., Bonet, X. (autores), & Solà, C. (director). (2020, 8 de noviembre). Tot el que sabem de tu [episodio de programa de televisión]. En C. Fernández (Productor) *30 minuts*. España: Corporació Catalana de Mitjans Audiovisuals. **https://www.ccma.cat/tv3/alacarta/30-minuts/tot-el-que-sabem-de-tu/video/6067763/**

García, D., Mitike Kassa, Y., Cuevas, A., Cebrián, M., Moro, E., Rahwan, I., & Cuevas, R. (2018). Analyzing gender inequality through large-scale Facebook advertising data, *PNAS*, *115*(27), 6958–6963. **https://doi.org/10.1073/pnas.1717781115**

Jiang, S., Li, Y., Lu, Q., Hong, Y., Guan, D., Xiong, Y., & Wang, S. (2021). Policy assessments for the carbon emission flows and sustainability of Bitcoin blockchain operation in China. *Nature Communications*, *12*, 1938. **https://doi.org/10.1038/s41467-021-22256-3**

Preist, C., Schien, D., & Shabajee, P. (2019). Evaluating sustainable interaction design of digital services: The case of YouTube. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. **https://doi.org/10.1145/3290605.3300627**

Sadowski, J. (2020). *Too smart. How digital capitalisme is extracting data, controlling our lives, and taking over the world*. MIT Press.

Stern, E. (2020, 27 de octubre). Núvols i aire fred per a la computació. *Divulcat*. **https://www.enciclopedia.cat/divulcat/nuvols-i-aire-fred-per-a-la-computacio**

Travizano, M. (2018, 28 de septiembre). The tech giants get rich using your data. What do you get in return? *Entrepreneur*. **https://www.entrepreneur.com/article/319952**

Tweedale, A. (2021). The carbon footprint of the internet: What's the environmental impact of being online? *OVO Blog*. **https://www.ovoenergy.com/blog/green/the-carbon-footprint-of-the-internet.html**

Véliz, C. (2020). *Privacy is power*. Bantam Press.

Warzel, C., & Thompson, S. A. (2021). They stormed the Capitol. Their apps tracked them. *The New York Times*. **https://www.nytimes.com/2021/02/05/opinion/capitol-attack-cellphone-data.html**

XAVIER DURAN. Químico y periodista científico, redactor especializado en ciencia y tecnología en los servicios informativos de TV3. Entre sus últimos libros està *El imperio de los datos* (Publicaciones de la Universitat de València, 2018) y *La ciencia en la literatura* (Universidad de Barcelona, 2018).