Técnicas formales de privacidad de datos:

¿Está el Servel protegiendo nuestra privacidad?

MATÍAS TORO

Doctor en Ciencias Mención Computación por la Universidad de Chile. Posdoctorante del Departamento de Ciencias de la Computación de la misma Universidad. Líneas de investigación: lenguajes de programación, *type-and-effect systems, gradual typing, security typing* y privacidad diferencial.

mtoro@dcc.uchile.cl

Nuestros datos personales están cada día repartidos en más y más aplicaciones o sistemas. Estos datos pueden (o no) contener información sensible o confidencial, como por ejemplo, información médica, financiera, de citas, o de preferencias artísticas. Esto genera un gran problema: el uso de los datos podría filtrar parte de esta información sensible. Informalmente, evitar que estos datos sensibles se hagan públicos o conocidos por terceros se llama *privacidad de publicación de datos*, que por simplicidad, llamaremos *privacidad*. En particular, en un análisis sobre una base de datos, se dice que la privacidad de un individuo es violada, si se aprende algo nuevo al agregar el individuo a la base de datos.

> Decimos que un análisis de datos *preserva la privacidad* si se aprende algo "útil" del análisis y al mismo tiempo no se viola la privacidad de ningún individuo.

Existen varias técnicas para preservar la privacidad, y dentro de las más importantes se encuentran la *anonimización* y la *privacidad diferencial*. En este artículo veremos un ejemplo práctico de violación de privacidad y propondremos cómo intentar revertirla usando técnicas de anonimización.

El caso del plebiscito

Recientemente, el Servicio Electoral (Servel) dispuso de un listado público con información acerca de la votación del plebiscito constitucional [1]. Cada elemento del listado corresponde a una persona, donde se indican algunos datos de ella como por ejemplo, su país de nacimiento, sexo, comuna, edad, y si votó o no votó en el plebiscito (llamado "votaron"). El *dataset* tiene 14.855.719 filas y cuenta con la información de todo el universo votante (ver Figura 1).

	Cédula	Circunscripción	Comuna	DV	Edad	Nacionalidad	País Domicilio	País Nacimiento	Partido	Provincia	Rango Edad	Región	Sexo	Sufragio	Voto Exterior	N° de registros	Votaron
0	0	Coyhaique	Coyhaique	0	22	chilena	Chile	Chile	[130] Federación Regionalista Verde Social	Coyhaique	20-24	De Aysén del General Carlos Ibáñez del Campo	F	sufragó	Nacional	1	1.0
1	0	El Puerto	Valparaíso	0	89	chilena	Chile	Chile	Sin Partido	Valparaíso	80 o +	De Valparaíso	F	no sufragó	Nacional	1	NaN
2	0	Iquique	Iquique	0	99	chilena	Chile	Chile	Sin Partido	Iquique	80 o +	De Tarapacá	M	no sufragó	Nacional	1	NaN
3	0	Río Tranquilo	Rio Ibáñez	0	22	chilena	Chile	Chile	Sin Partido	General Carrera	20-24	De Aysén del General Carlos Ibáñez del Campo	M	sufragó	Nacional	1	1.0
4	0	El Puerto	Valparaíso	0	95	chilena	Chile	Chile	Sin Partido	Valparaíso	80 o +	De Valparaíso	M	no sufragó	Nacional	1	NaN

Figura 1. Datos de participación electoral del plebiscito 2020.

df[(df['Pais Nacimiento'] == 'Irlanda') & (df['Sexo'] == 'masculino') & (df['Comuna'] == 'Santiago') & (df['Edad'] >= 30) & (df['Edad'] <= 39)]

	Cédula	Circunscripción	Comuna	DV	Edad	Nacionalidad	País Domicilio	País Nacimiento	Partido	Provincia	Rango Edad	Región	Sexo	Sufragio	Voto Exterior	N° de registros	Votaron
1009708	0	Parque Almagro	Santiago	0	35	extranjero	Chile	Irlanda	Sin Partido	Santiago	35-39	Metropolitana de Santiago	M	sufragó	Nacional	1	1.0
1766909	0	El Centro	Santiago	0	36	extranjero	Chile	Irlanda	Sin Partido	Santiago	35-39	Metropolitana de Santiago	M	no sufragó	Nacional	1	NaN
14046592	0	El Centro	Santiago	0	34	extranjero	Chile	Irlanda	Sin Partido	Santiago	30-34	Metropolitana de Santiago	M	no sufragó	Nacional	1	NaN

Figura 2. Solo tres personas calzan con los criterios del filtro.

df.groupby(['Edad', 'Pais Nacimiento', 'Sexo' , 'Comuna']).filter(lambda x: x['Número de registros'].count() == 1)

	Cédula	Circunscripción	Comuna	DV	Edad	Nacionalidad	País Domicilio	País Nacimiento	Partido	Provincia	Rango Edad	Región	Sexo	Sufragio	Voto Exterior	N° de registros	Votaron
48	0	Chanco	Chanco	0	95	chilena	Chile	Chile	Sin Partido	Cauquenes	80 o +	Del Maule	M	no sufragó	Nacional	1	NaN
133	0	Santa Bárbara	Santa Bárbara	0	115	chilena	Chile	Chile	Sin Partido	Biobío	80 o +	Del Biobío	F	no sufragó	Nacional	1	NaN
194	0	Arica	Arica	0	133	chilena	Chile	Chile	Sin Partido	Arica	80 o +	De Arica y Parinacota	M	no sufragó	Nacional	1	NaN
209	0	Aysén	Aysén	0	61	extranjero	Chile	Japón	Sin Partido	Aysén	60-64	De Aysén del General Carlos Ibáñez del Campo	M	no sufragó	Nacional	1	NaN
212	0	San Joaquín	San Joaquín	0	67	extranjero	Chile	Argentina	Sin Partido	Santiago	65-69	Metropolitana de Santiago	M	no sufragó	Nacional	1	NaN
	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
14840749	0	0	0	0	79	extranjero	Chile	Uruguay	Sin Partido	Santiago	75-79	Metropolitana de Santiago	M	sufragó	Nacional	1	1.0
14840751	0	Isla de Maipo	Isla de Maipo	0	81	extranjero	Chile	Estados Unidos	Sin Partido	Talagante	80 o +	Metropolitana de Santiago	M	sufragó	Nacional	1	1.0
14842069	0	Renaico	Renaico	0	97	chilena	Chile	Chile	Sin Partido	Malleco	80 o +	De la Araucanía	M	no sufragó	Nacional	1	NaN
14842084	0	Quirihue	Quirihue	0	99	chilena	Chile	Chile	Sin Partido	Itata	80 o +	De Ñuble	M	no sufragó	Nacional	1	NaN
14851455	0	O'Higgins	O'Higgins	0	21	chilena	Chile	Chile	Sin Partido	Capitán Prat	20-24	De Aysén del General Carlos Ibáñez del Campo	M	sufragó	Nacional	1	1.0

65585 rows × 17 columns

Figura 3. Hay 65.585 grupos de edad, país de nacimiento, sexo y comuna con un solo elemento.

Con la intención de proteger la privacidad de cada persona, las columnas **rut** y **nombre** fueron eliminadas/ofuscadas del listado público. Estas columnas son llamadas *identificadoras* ya que identifican únicamente a cualquier individuo. Las columnas país de nacimiento, sexo, comuna y edad, las llamaremos *cuasi-identificadores* (porque "casi" identifican a un individuo), y la columna "votaron" la llamaremos *atributo sensible* ya que es un valor que se desea proteger.

> Una columna se llama *identificadora* si permite identificar unívocamente a cualquier individuo. Se llama *cuasi-identificadora* si "casi" permite identificar a un individuo.

Lamentablemente, eliminar las columnas rut y nombre no es suficiente para proteger la privacidad de **todos** los individuos; podemos *re-identificar* a ciertos individuos usando *datos auxiliares*. Con datos auxiliares, nos referimos a un listado de datos que contienen las columnas identificadoras y algunas cuasi-identificadoras. Esta información es usada comúnmente por atacantes que buscan asociar columnas identificadoras a datos sensibles. A continuación mostraremos lo que se llama *ataque de asociación de registros* para averiguar si un amigo nuestro votó o no votó en el plebiscito.

Sabemos que nuestro amigo es irlandés, su sexo es masculino, vive en Santiago y en el 2019 tenía entre 30 y 39 años. Si filtramos el listado usando estos datos, podemos ver que solo tres personas cumplen con esas condiciones (ver Figura 2).

Adicionalmente, podemos aprender que hay un 33,33% de probabilidad de que nuestro amigo haya votado en el plebiscito. Peor aún, si conseguimos su edad real en el 2019 (35), podemos saber con 100% de probabilidad que nuestro amigo sí votó en el plebiscito.

De hecho, si agrupamos los datos por los cuasi-identificadores, es decir, contamos cuánta gente hay por cada conjunto de valores de atributos distintos, podemos ver que hay 65.585 individuos que potencialmente podemos re-identificar unívocamente (ver Figura 3).

```
df.groupby(['Edad', 'Pais Nacimiento', 'Sexo' , 'Circunscripcion', 'Comuna', 'Nacionalidad', 'Pais Domicilio', 'Partido', 'Provincia', 'Region']).filter(lambda x: x['Número de registros'].count() == 1)
```

	Cédula	Circunscripción	Comuna	DV	Edad	Nacionalidad	País Domicilio	País Nacimiento	Partido	Provincia	Rango Edad	Región	Sexo	Sufragio	Voto Exterior	N° de registros	Votaron
48	0	Chanco	Chanco	0	95	chilena	Chile	Chile	Sin Partido	Cauquenes	80 o +	Del Maule	M	no sufragó	Nacional	1	NaN
133	0	Iquique	Iquique	0	97	chilena	Chile	Chile	[3] Unión Democrata Ind.	Iquique	80 o +	De Tarapacá	F	no sufragó	Nacional	1	NaN
194	0	El Golf	Las Condes	0	98	chilena	Chile	Chile	[2] Partido Demócrata Cristiano	Santiago	80 o +	Metropolitana de Santiago	F	no sufragó	Nacional	1	NaN
209	0	Santa Bárbara	Santa Bárbara	0	115	chilena	Chile	Chile	Sin Partido	Biobío	80 o +	Del Biobío	F	no sufragó	Nacional	1	NaN
212	0	Temuco centro	Temuco	0	99	chilena	Chile	Chile	[2] Partido Demócrata Cristiano	Cautín	80 o +	De la Araucanía	M	no sufragó	Nacional	1	NaN
	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
14840749	0	Los Lagos	Los Lagos	0	21	chilena	Chile	Chile	[2] Partido Demócrata Cristiano	Valdivia	20-24	De los Ríos	M	sufragó	Nacional	1	1.0
14840751	0	Natales	Natales	0	21	chilena	Chile	Chile	[37] Evolución Política	Última Esperanza	20-24	De Magallanes y de la Antártica Chilena	M	sufragó	Nacional	1	1.0
14842069	0	Torres del Paine (C. Castillo)	Torres del Paine	0	21	chilena	Chile	Chile	[8] Humanista	Última Esperanza	20-24	De Magallanes y de la Antártica Chilena	M	sufragó	Nacional	1	1.0
14842084	0	Natales	Natales	0	22	chilena	Chile	Chile	[8] Humanista	Última Esperanza	20-24	De Magallanes y de la Antártica Chilena	M	sufragó	Nacional	1	1.0
14851455	0	Huara	Huara	0	21	chilena	Chile	Chile	[1] Renovación Nacional	Del Tamarugal	20-24	De Tarapacá	M	sufragó	Nacional	1	1.0

257349 rows × 17 columns

Figura 4. Si agrupamos los datos por más columnas, podemos obtener más grupos de un elemento.

Por ejemplo, si sabemos que una persona está inscrita en Aysén, tenía 61 años el 2019, es japonés y tiene sexo masculino, sabemos con certeza que no votó: no hay otra persona con esas características en el conjunto de datos.

Peor aún, si consideramos más columnas, la cantidad de personas que potencialmente podríamos re-identificar aumentará aún más, llegando hasta un máximo de 257.349 personas (ver Figura 4).

¿Qué podemos hacer entonces?

No todo está perdido, por suerte existen garantías formales que nos ayudarán a obtener un mejor nivel de privacidad, donde la más básica se conoce como *k-anonimato* [2].

> Se dice que un conjunto de datos satisface *k-anonimato* si y solo si cada grupo de cuasi-identificadores aparece al menos *k* veces en el conjunto de datos.

Intuitivamente, esto quiere decir que si un atacante cuenta con información auxiliar, el individuo que busca re-identificar estará "mezclado" con otros *k*-1 individuos con las mismas características.

El conjunto de datos del Servel claramente cumple 1-anonimato, y sería bueno encontrar una serie de transfor-

	Edad	Nacionalidad	Sexo	Comuna	count
93171	138	chilena	masculino	Freire	1
80551	104	chilena	masculino	Rauco	1
80554	104	chilena	masculino	Renca	1
80557	104	chilena	masculino	Rinconada	1
80560	104	chilena	masculino	Romeral	1
...	...	...	...	...	...
10697	29	chilena	femenino	Puente Alto	5123
9041	27	chilena	masculino	Puente Alto	5143
12059	30	chilena	masculino	Puente Alto	5209
11042	29	chilena	masculino	Puente Alto	5214
10037	28	chilena	masculino	Puente Alto	5246

93172 rows × 5 columns

Figura 5. En gris se indican algunos de los grupos de personas con un solo elemento.

	Rango Edad	pnac	Sexo	Region	count
39	18-19	Extranjero	femenino	De los Ríos	1
739	70-74	Extranjero	femenino	De Aysén del General Carlos Ibáñez del Campo	3
803	75-79	Extranjero	femenino	De Aysén del General Carlos Ibáñez del Campo	6
99	20-24	Extranjero	femenino	De Aysén del General Carlos Ibáñez del Campo	6
43	18-19	Extranjero	femenino	De Ñuble	6
...	...	...	...	...	...
287	35-39	Chile	masculino	Metropolitana de Santiago	257260
143	25-29	Chile	femenino	Metropolitana de Santiago	277837
207	30-34	Chile	femenino	Metropolitana de Santiago	282399
159	25-29	Chile	masculino	Metropolitana de Santiago	283588
223	30-34	Chile	masculino	Metropolitana de Santiago	288647

896 rows × 5 columns

Figura 6. Datos agrupados por columnas generalizadas.

maciones a los datos que hagan que los datos publicados cumplan *k*-anonimato para al menos *k*>1 (entre más grande el *k*, mejor).

Como ejercicio, consideraremos solo las columnas edad, nacionalidad, sexo y comuna. Si agrupamos por estas columnas y contamos los distintos valores, podemos ver que hay muchos grupos con un solo miembro (ver Figura 5).

Como primer intento de anonimización podemos *generalizar* las columnas edad, nacionalidad y comuna. La generalización de una columna consiste en reemplazar sus valores por otros valores que corresponden a una categoría más amplia. Para la columna edad, generalizamos en rangos de edad de 5 en 5, para nacionalidad generalizamos las nacionalidades extranjeras como 'extranjera', y para comuna generalizamos publicando solo la región de ella. Al hacer esto obtenemos los grupos de la Figura 6.

Mejoraron los grupos, pero los datos siguen siendo 1-anónimos ya que queda 1 grupo con 1 integrante. Podemos solucionar esto haciendo una unión de los grupos de 18 a 19 con los de 20-24, obteniendo así datos que son 3-anónimos.

Claramente esta no es la solución final, ya que hay más quasi-identificadores en los datos publicados y no sabemos con certeza qué atributos conoce el atacante. Adicionalmente, aumentar la cantidad de atributos dificulta más la tarea de mejorar el *k*-anonimato, ya que los grupos serán más específicos, y por ende se requerirá generalizar aún más, disminuyendo la utilidad de los datos.

Otra solución que es independiente de los datos auxiliares que podría tener un atacante, es usar una técnica más reciente llamada *privacidad diferencial* [3].

> *La privacidad diferencial* consiste en perturbar los resultados agregando ruido aleatorio.

Esta técnica se especializa en publicar datos agregados y no microdatos (un subset de los datos originales), pero genera garantías de privacidad formales más fuertes. De hecho, esta técnica iba a ser usada para publicar los datos del censo del 2020 de Estados Unidos que se postergó por la pandemia por COVID-19 [4].

Conclusión

Existen muchos otros casos de estudio emblemáticos acerca de violación de privacidad. Por ejemplo, en el 2007 Netflix realizó una competencia abierta para mejorar su sistema de recomendación. Para ello se proveyó de un dataset anonimizado, donde los identificadores fueron reemplazados por otros aleatorios. El problema fue que a estos datos se les hizo un ataque de asociación de registro con data pública de IMDB (Internet Movie DataBase), de-anonimizando así a muchos individuos, revelando sus preferencias de películas y series. Otro caso emblemático ocurrió en 1997 en Estados Unidos, donde un gobernador aprobó la liberación de registros médicos anonimizados de funcionarios públicos. Dos días después el gobernador recibió un correo con todos sus registros médicos, obtenidos mediante un ataque de asociación con el padrón electoral, cruzando código postal, fecha de nacimiento y género.

Vimos que existen técnicas formales que permiten evitar estos escándalos, y publicar datos asegurando la privacidad de los individuos. Pero aún existen desafíos, como por ejemplo (1) que la privacidad de datos aún no está interiorizada en el colectivo, y (2) no se sabe (ni se puede predecir) la información auxiliar con la que va a contar un atacante. Afortunadamente, estas técnicas están ganando terreno en grandes compañías, como el caso de la privacidad diferencial, que está siendo usada por Google y Microsoft para telemetría, LinkedIn para análisis de marketing, y Apple para la recolección de datos personales de usuarios. ■

REFERENCIAS

[1] https://www.servel.cl/wp-content/uploads/2021/06/VW_VOTARON_2020PLEB_Datos_completos.zip.

[2] P. Samarati and L. Sweeney. 1998. “Protecting Privacy When Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression”. Technical Report SRI-CSL-98-04.

[3] C. Dwork and A. Roth. 2014. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science 9, 3–4 (2014), 211–407.

[4] S. Ruggles, C. Fitch, D. Magnuson, and J. Schroeder. 2019. “Differential Privacy and Census Data: Implications for Social and Economic Research.” AEA Papers and Proceedings 109:403–408.