Privacidad Diferencial: introducción

Guillermo Navarro-Arribas

June 13, 2023

Privacidad diferencial (*differential privacy*) es un termino que posiblemente habrás visto asociado a la privacidad de datos o información. En los últimos años ha tomado mucha relevancia, primero en el mundo académico y posteriormente a nivel comercial gracias a que grandes empresas como Google[1] o Apple[2] anunciaron el uso de privacidad diferencial para proteger los datos de sus usuarios. Pero ¿qué es la privacidad diferencial? ¿Como nos garantiza la privacidad o anonimato de datos de usuarios? ¿Es realmente efectiva?

1 Consultando una base de datos

Vamos a suponer que tenemos una base de datos que llamaremos D con información sobre ciertos usuarios. Esta base de datos contiene información confidencial sobre los usuarios como pueden ser datos personales o fiscales. Como ejemplo y de forma muy simplificada, supongamos que la base de datos contiene los siguientes registros (entendemos por registro cada fila):

nombre	edad	salario
Ataúlfo	43	100
Sigerico	15	300
Walia	50	200
Teodoredo	51	100
Turismundo	16	500
Tedorico II	40	300
Eurico	14	700
Alarico II	49	300
Gesaleico	69	500
Amalarico	17	200

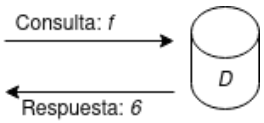
Para obtener información de la base de datos permitimos que otros usuarios hagan consultas y obtengan las correspondientes respuestas.

Por ejemplo, supongamos que queremos permitir que se pueda obtener la consulta:

"número de registros/individuos mayores de edad". Es decir, el número de individuos con edad igual o superior a 18. Denotaremos esta consulta como la función f, y el resultado de aplicar la consulta a la base de datos D, como $f(D)$.

> **Ejemplo 1** $f(D) =$ núm. de usuarios en D con edad ≥ 18

Si hacemos la consulta a la base de datos D anterior obtendremos que $f(D) = 6$.



Podríamos pensar que esta consulta f es segura, ya que no obtenemos información sensible de ningún usuario en concreto. Obtenemos información agregada, que no nos esta revelando datos específicos de ningún usuario. Sin embargo y como veremos, esta consulta aparentemente inocente puede acarrear problemas de privacidad y revelar información sensible sobre los usuarios de la base de datos.

2 ¿Que pasa si tenemos versiones diferentes de la base de datos?

Supongamos que tenemos dos versiones diferentes de la base de datos anterior que denotamos como D_1 y D_2. Estas dos bases de datos son idénticas a excepción de un solo registro. Por ejemplo:

Table 1: D_1

nombre	edad	salario
Ataúlfo	43	100
Sigerico	**15**	**300**
Walia	50	200
Teodoredo	51	100
Turismundo	16	500
Tedorico II	40	300
Eurico	14	700
Alarico II	49	300
Gesaleico	69	500
Amalarico	17	200

En este caso, D_2 es igual a D_1 con la excepción de que le falta un registro, el correspondiente a *Sigerico*.

Cuando la diferencia entre dos bases de datos D_1 y D_2 es de un solo registro, lo denotamos

Table 2: D_2

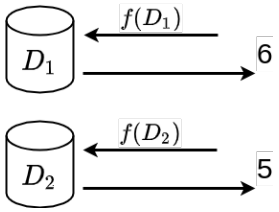
nombre	edad	salario
Ataúlfo	43	100
Walia	50	200
Teodoredo	51	100
Turismundo	16	500
Tedorico II	40	300
Eurico	14	700
Alarico II	49	300
Gesaleico	69	500
Amalarico	17	200

como $d(D_1, D_2) = 1$.

¿Que pasa si hacemos la consulta anterior en D_1 y D_2?

- $f(D_1) = 6$
- $f(D_2) = 6$

¿Hay ahora algún problema de privacidad? El lector avispado tenderá a pensar que sí (¿porqué si no estamos haciendo todo esto?).



Supongamos que quien hace las dos consultas (en adelante *el atacante*) sabe que D_2 tiene un registro menos que D_1. Puede incluso saber que dicho registro corresponde a un usuario concreto, ya que este ha solicitado borrar su información de la base de datos. Al no conocer los registros de la base de datos, el atacante no sabe la edad ni el salario de dicho usuario, pero ahora sí **sabe que no es mayor de edad**. Es decir, ha obtenido información nueva que no tenia sobre un usuario concreto de la base de datos.

3 Entra en juego la privacidad diferencial

La **privacidad diferencial** proporciona una definición de privacidad que se basa en controlar la información que se filtra en el ejemplo anterior. De manera muy zafia podemos decir que:

> **Definición 1** Una consulta a una base de datos, como nuestra consulta f, cumple privacidad diferencial si al hacer la misma consulta en D_1 y D_2 no obtenemos mucha información nueva.

Es decir, si el atacante hace la consulta a las dos bases de datos, no se revela mucha más información de la que ya tenia. O dicha información esta acotada. Aquí hay varios puntos a destacar:

- ¿Qué queremos decir con que se revele *mucha* o *poca* información? Este es un punto clave y difícil de precisar que de momento no abordaremos de forma directa. Veremos más adelante como se mide o acota esta información en privacidad diferencial.

- Esto se tiene que cumplir para cualquier par de versiones de la base de datos que difieran en un solo registro (sea cual sea ese registro). Es decir, cualquier D_i y D_j tal que $d(D_i, D_j) = 1$.

- Hablamos de que la información que se obtiene está acotada y no de que no se obtiene ninguna información. Esto último seria ideal desde el punto de vista de la privacidad, pero si fuese así, la consulta f tendría que, por ejemplo, retornar un valor constante independientemente del registro que falte. Cosa que no resultaría nada útil (no serviría de nada realizar dicha consulta).

¿Como podemos conseguir esto? La manera más inmediata es intentar sustituir nuestra consulta f por otra versión más segura que llamaremos K_f. Esta nueva función debe cumplir con dos requisitos importantes:

1. Debe ser *útil*. Quien hace la consulta quiere obtener información sobre el numero de registros con personas mayores de edad. Esto tiene que seguir siendo así en la medida de lo posible. No podemos retornar un valor constante o muy alejado del real.

2. Deber ser segura o conseguir un nivel de privacidad aceptable. La información que un atacante pueda obtener haciendo la consulta a diferentes versiones de la base de datos tienes que ser muy poca.

¿Como podemos implementar K_f?

4 Privacidad diferencial con ruido

La manera más sencilla de pensar en esta función segura K_f es considerarla como la función original pero con cierto ruido añadido: $K_f(D) = f(D) + r$, donde r es un valor de ruido elegido al azar. Es decir, cuando hacemos la consulta f obtendremos el número de personas mayores de edad con un ruido o error añadido.

¿Porque es esto interesante?

1. El resultado puede ser útil en la medida en que el ruido sea relativamente pequeño. Al hacer la consulta, no obtenemos el número exacto de personas mayores de edad, pero sí que obtenemos una aproximación que en muchos casos puede ser suficiente.

2. Se consigue un cierto nivel de privacidad debido a la **incertidumbre** que hay ahora sobre el resultado. Cuando el atacante hace dos consultas en dos versiones de la base de datos la diferencia entre los valores que obtiene puede venir determinada por el ruido de las dos consultas y no necesariamente o directamente por la diferencia que pueda haber entre los valores reales.

Todo esto que planteamos de forma un poco confusa, se concreta en la conocida definición de privacidad diferencial, que de forma un poco más formal detallamos a continuación.

> **Definición 2** Una función K_f para una consulta f, proporciona ϵ-**privacidad diferencial** si para todas las bases de datos D_1 y D_2 que difieren en un solo elemento, y para todo $S \subseteq Rango(K_f)$:
>
> $$P[K_f(D_1) \in S] \leq e^{\epsilon} \cdot P[K_f(D_2) \in S] \tag{1}$$

Pero, ¿que quiere decir todo esto?. Vayamos por partes:

- $Range(K_f)$: el rango ($Range$) de una función es el conjunto de todos los valores que puede retornar. Es decir, $Range(K_f)$ es el conjunto de todos los valores que puede retornar la función K_f. En nuestro caso podriamos considerar que el rango de K_f es el conjunto de todos los numero enteros iguales o mayores que 0 (la función f no puede retornar un número negativo).

- $S \subseteq Range(K_f)$: S es un subconjunto de $Range(K_f)$. Es decir cualquier subconjunto de numeros enteros mayores o iguales a 0. Algun ejemplo podria ser $S = \{0, 1, 2, 3, 4, 5\}$, $S = \{3, 5, 11\}$, $S = \{100, 235, 24\}$, ...

- $P[K_f(D_1) \in S]$: se refiere a la probabilidad de que el resultado de aplicar K_f a D_1 pertenezca al subconjunto S.

Estamos mirando entonces la diferencia que hay entre la probabilidad de que la consulta a D_1 y D_2 pertenezca al mismo subconjunto S, sea cual sea $S \subseteq Range(K_f)$. O dicho de otra manera, que la diferencia entre la consulta $K_f(D_1)$ y $K_f(D_2)$ sea lo menos perceptible posible. Donde este *menos perceptible posible* queda acotado de forma más precisa diciendo que tiene que ser menor que e^{ϵ}.

Si la función K_f cumple la condición anterior decimos que cumple privacidad diferencial para ϵ, o que cumple ϵ -*differential privacy.*

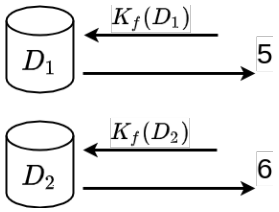
Esta condición a veces se expresa en forma de conciente:

$$\frac{P[K_f(D_1) \in S]}{P[K_f(D_2) \in S]} \leq e^{\epsilon} \tag{2}$$

Por ejemplo, supongamos que definimos el ruido r como un valor elegido de forma aleatoria y uniforme del conjunto de números $\{-2, -1, 0, 1, 2\}$. Siguiendo con el ejmplo de la función K_f, supongamos que hacemos la consulta a D_1 y D_2 y en cada caso obtenemos los valores de ruido $r_1 = -1$, $r_2 = 0$. Es decir:

- $K_f(D_1) = 6 + r_1 = 6 - 1 = 5$
- $K_f(D_2) = 6 + r_2 = 6 + 0 = 6$

Es importante remarcar que los valores r_1 y r_2 los elije la BD y obviamente se mantienen en secreto. Quien recibe el resultado de la función K_f desconoce el ruido que ha sido sumado.



Ahora existe diferencia entre las dos respuestas pero resulta más difícil poder obtener información sobre el valor del registro que ha sido eliminado en D_2. No podemos saber con certeza si corresponde a una persona mayor o menor de edad, ya que no sabemos el ruido que ha sido añadido en cada caso.

Sí que sabemos que dicho ruido puede ser un valor del conjunto $\{-2, -1, 0, 1, 2\}$. Como cada uno de estos valores tiene la misma probabilidad de ser elejido por la función K_f, podemos intentar ver cual seria el valor original.

En el caso de D_1 tenemos que $K_f(D_1) = 5$, y sabemos que $K_f(D_1) = f(D_1) + r_1$. El atacante, no conoce el valor $f(D_1)$, ni r_1. ¿Puede llegar a saber el valor de $f(D_1)$? Inicialmente no, pero puede hacer una estimación. Dependiendo del ruido que se haya utilizado tendremos los valores que se muestran en al siguiente tabla:

Es decir, los posibles valores de $f(D_1)$ son $\{7, 6, 5, 4, 3\}$. De la misma manera, los posibles valores de $f(D_2)$ serian $\{8, 7, 6, 5, 4\}$.

A modo ilustrativo, podemos ver en la siguiente tabla la probabilidad de que $K_f(D_1)$ o $K_f(D_2)$ pertenzca a algun conjunto S de ejemplo.

Podemos decir p.e. que nuestra función K_f no satisface privacidad diferencial para $\epsilon = 1$.

Table 3: Posibles valores de $f(D_1)$ sabiendo que $K_f(D_1) = 5$.

r_1	$f(D_1)$
-2	7
-1	6
0	5
+1	4
+2	3

Table 4: Probabilidades para algunos valores de S

S	$P[K_f(D_1) \in S]$	$P[K_f(D_2) \in S]$
$\{3\}$	$1/5$	0
$\{6, 5, 4\}$	$3/5$	$3/5$
$\{8, 7\}$	$2/5$	$1/5$

Para mostrarlo, podemos dar un contraejemplo ya que es fácil ver que no se cumple la siguiente condición:

$$P[K_f(D_1) \in \{3\}] \leq eP[K_f(D_2) \in \{3\}]$$

Como veremos más adelante, en privacidad diferencial se utilizan valores de ruido obtenidos a partir de distribuciones de probabilidades concretas. Siendo la más popular la distribución de Laplace.

En la terminología de privacidad diferencial se utilizan los siguiente términos:

***mechanism* (mecanismo)** función o algoritmo que recibe una base de datos como parámetro y retorna un resultado. En nuestro caso seria la función f.

***randomized mechanism* (mecanismo aleatorio)** mecanismo al que se añade aleatoriedad con el fin de proporcionar privacidad a la respuesta. En nuestro caso, la función K_f.

5 La privacidad o revelación de información es acumulativa

La privacidad diferencial permite establecer de manera precisa la noción de acumulación de privacidad o, como se suele decir, **presupuesto de privacidad (*privacy budget*)**. La idea es que la propiedad de privacidad diferencial se puede ver como un presupuesto de privacidad que se va *gastado* cada vez que se usa.

Por ejemplo, tomamos el mecanismo aleatorio anterior K_f que retorna la media de edad mas un ruido elegido de forma uniforme entre el conjunto $\{-2, -1, 0, 1, 2\}$.

Dada una base de datos D_1 como la anterior podemos obtener el resultado $K_f(D_1) = 5$. Con este resultado un atacante no sabe el valor original $f(D_1)$ (que recordamos es 6), pero puede saber que será un número del conjunto $\{7, 6, 5, 4, 3\}$. La probabilidad de adivinar el valor original es entonces de 1/5.

Pero ¿que pasa, si ahora volvemos a repetir la misma consulta con la misma base de datos? Al repetirla podriamos obtener, por ejemplo, que $K_f(D_1) = 7$ de donde el atacante aprende que los posibles valores de $f(D_1)$ en este caso serian $\{9, 8, 7, 6, 5\}$. En la tabla siguiente vemos lo que aprende el atacante en cada caso:

Consulta	$K_f(D_1)$	¿ $f(D_1)$?
1	5	$7, 6, 5, 4, 3$
2	7	$9, 8, 7, 6, 5$

El avispado atacante, una vez realiza las dos consultas, puede concluir que el valor que busca ($f(D_1)$) estará en el conjunto de posibles valores de las dos consultas. Es decir, es su intersección: $\{7, 6, 5\}$. Como vemos, ahora el atacante tiene una probabilidad de 1/3 de adivinar el valor $f(D_1)$. Es decir ha reducido la incertidumbre respecto al valor $f(D_1)$ y por tanto se ha reducido la privacidad que conseguimos con la función K_f.

Si un atacante puede realizar la misma consulta varias veces a la misma base de datos, la incertidumbre que proporciona el mecanismo aleatorio se reduce. En general, cuantas más veces pueda repetir la consulta, mayor certeza tendrá sobre el valor original y mayor será la probabilidad de acierto.

Este ejemplo ilustra la idea de presupuesto de privacidad que se va gastando (o reduciendo) cada vez que se hace una consulta. Tendremos entonces que imponer restricciones o límites en el numero de consultas que se pueden hacer. Estas restricciones quedan formuladas en privacidad diferencial con lo que se conoce como *teorema de composición*.

6 Continuará...

¿Cómo podemos diseñar mecanismos más realistas? ¿Cómo se calcula ϵ en el caso general? ¿Cómo establecemos el presupuesto de privacidad y cómo lo calculamos?

7 References

[1] M. Guevara, "Google Developers Blog: Enabling developers and organizations to use differential privacy." 2019. Accessed: Feb. 10, 2021. [Online]. Available: `https://developers.googleblog.com/2019/09/enabling-developers-and-organizations.html`

[2] Apple Inc., "Differential Privacy Overview," 2016. Accessed: Feb. 10, 2021. [Online]. Available: `https://www.apple.com/privacy/docs/Differential_Privacy_Overview.`

pdf