k-Anonimato Probabilístico

Jordi Soria-Comas*, Josep Domingo-Ferrer*, David Rebollo-Monedero†

*Universitat Rovira i Virgili
Dept. d'Enginyeria Informàtica i Matemàtiques
Cátedra UNESCO de Privacidad de Datos
Av. Països Catalans 26, E-43007 Tarragona
E-mail {jordi.soria,josep.domingo}@urv.cat

†Universitat Politècnica de Catalunya
Dept. d'Enginyeria Telemàtica
Campus Nord, Mòdul C5, Despatx S102A, E-08034 Barcelona
E-mail david.rebollo.monedero@upc.edu

***Resumen*—El k-anonimato es una propiedad usada para limitar el riesgo de revelación en ficheros de microdatos. Un fichero de microdatos k-anónimo está compuesto por grupos de k registros indistinguibles entre ellos en base a los cuasi-identificadores. Introducimos la noción de k-anonimato probabilístico, que relaja los requisitos de indistinguibilidad pero mantiene el mismo nivel de probabilidad de reidentificación de los registros que ofrece el k-anonimato. Se proponen dos métodos para obtener k-anonimato probabilístico, ambos basados en agrupación e intercambio. Se proporcionan resultados experimentales que comparan la calidad de los datos obtenidos aplicando k-anonimato y k-anonimato probabilístico.**

***Index Terms*—Anonymization; clustering; microaggregation; swapping; k-anonymity.**

I. Introducción

Los ficheros de microdatos se obtienen como resultado de un proceso de recogida de datos y contienen información específica de cada uno de los individuos cuyos datos se han recogido. Para cada individuo hay un registro que contiene información sobre un conjunto de atributos. Los ficheros de microdatos son un recurso muy valioso para analistas e investigadores, pero también suponen un importante riesgo para la privacidad de los individuos sobre los que informan. Antes de publicar un fichero de microdatos, éste debe pasar por un proceso de anonimización, en el que se disocie la identidad de cada individuo de su registro correspondiente.

Dos aspectos fundamentales de todo método de anonimización son: el nivel de protección contra la revelación que ofrece y la pérdida de información que produce. La literatura sobre métodos de anonimización para ficheros de microdatos es muy extensa: [1], [2], [3] ofrecen una buena visión de conjunto de la literatura relacionada.

De entre las posibles opciones para anonimizar ficheros de microdatos, nosotros nos centramos en el k-anonimato. Más que un método para anonimizar, el k-anonimato es una propiedad que establece los requisitos que debe satisfacer un fichero de microdatos. Si se considera que el k-anonimato es una garantía suficiente para la privacidad de los individuos que hay en el fichero de microdatos, el siguiente paso es aplicar algún método para transformar el fichero original en uno k-anónimo.

El k-anonimato garantiza que cada registro es indistinguible en relación a los cuasi-identificadores dentro de un conjunto de k registros; es decir, todos los registros que pertenecen al mismo grupo comparten los mismos valores para los cuasi-identificadores. Esto supone una reducción en la cantidad de información que aportan los cuasi-identificadores. Ello puede ser un problema, especialmente si hay un gran número de cuasi-identificadores [4].

Nuestro objetivo es obtener el mismo nivel de protección contra la revelación que proporciona el k-anonimato, a la vez que reducimos la pérdida de información que produce. Nuestra propuesta consiste en relajar el estricto requerimiento de indistinguibilidad para los cuasi-identificadores, requiriendo únicamente que la probabilidad de reidentificación sea $1/k$. Esto amplía el repertorio de métodos que puede usarse, y por tanto lleva a una reducción potencial de la pérdida de información.

II. Antecedentes

Modelamos un fichero de microdatos como una tabla, donde cada fila corresponde a un individuo diferente y cada columna a un atributo. Usamos la notación $T(A_1, \ldots, A_n)$ para representar un fichero de microdatos con los atributos $A_1, \ldots, A_n$.

No todos los atributos suponen la misma amenaza para la privacidad de los individuos en el conjunto de microdatos. Clasificamos los atributos de acuerdo al tipo de riesgo que inducen las siguientes categorías no necesariamente disjuntas.

- Identificadores. Se dice que un atributo es un identificador si el valor de este atributo es suficiente para reidentificar el registro sin ambigüedad; esto es, asignarle una identidad. Por ejemplo, el número de la seguridad social, el número del pasaporte, etc. En un fichero de microdatos, la presencia de un identificador hace que sea posible enlazar los registros a las identidades de los individuos, y por tanto descubrir el valor de todos

los atributos para dicho individuo. Para evitarlo, los identificadores son eliminados o cifrados. En el resto del artículo suponemos que el fichero de microdatos $T(A_1, \ldots, A_n)$ no contiene identificadores.

- Cuasi-identificadores. Decimos que un atributo es un cuasi-identificador si por sí solo no es suficiente para reidentificar a ningún individuo, pero en combinación con otros cuasi-identificadores sí puede conducir a una reidentificación. A diferencia de los identificadores, no es posible eliminar los cuasi-identificadores, puesto que provocaria una importante pérdida de utilidad de los datos. Además cualquier atributo es potencialmente un cuasi-identificador y por tanto no es posible eliminarlos todos. Que un atributo sea considerado cuasi-identificador depende de la información de la que dispone el atacante. Si un atacante conoce el valor que toma ese atributo para un individuo específico, el atributo debe ser considerado cuasi-identificador.
- Atributos confidenciales. Contienen la información sensible. El objetivo principal de las técnicas de protección de microdatos es evitar que a partir del fichero de microdatos se pueda descubrir información confidencial sobre individuos concretos. Esto incluye no solamente que un atacante descubra el valor exacto que cierto atributo toma para cierto individuo, sino también que un atacante pueda incrementar su conocimiento respecto a cierto individuo (por ejemplo acotando el rango de valores que cierto atributo puede tomar para un individuo concreto).
- Atributos no confidenciales. Se consideran como no confidenciales los atributos que no pertenecen a ninguna de las categorías anteriores. Como no contienen información confidencial y tampoco son de utilidad para reidentificar individuos, no los consideraremos en nuestra discusión. Supondremos que el fichero de microdatos $T(A_1, \ldots, A_n)$ no contiene atributos no confidenciales.

Según hemos dicho, el objetivo de las técnicas de protección de microdatos es evitar que el fichero de microdatos pueda revelar información confidencial específica a un individuo. Clasificamos el riesgo de revelación en dos categorías [5]:

- Revelación de la identidad. El atacante es capaz de determinar la identidad correspondiente a un registro, y de esta manera puede asignar un valor concreto a todos los atributos confidenciales para ese individuo.
- Revelación de atributos. Aunque el intruso no pueda reidentificar un registro, puede ser posible determinar algún tipo información referente a un individuo a partir del fichero de microdatos. Por ejemplo, si el atacante sabe que un individuo trabaja de contable, y tiene acceso a un fichero de microdatos que contiene el tipo de trabajo y el sueldo, entonces puede acotar el sueldo de ese individuo inferior y superiormente.

Para proteger la privacidad de los individuos cuyos datos están en el fichero de microdatos, no se publica el fichero original $T(A_1, \ldots, A_n)$, sino una versión modificada $T'(A_1, \ldots, A_n)$, donde los valores de los cuasi-identificadores y/o atributos confidenciales han sido enmascarados.

Una posibilidad para limitar el riesgo de revelación consiste en publicar un fichero de microdatos $T'(A_1, \ldots, A_n)$ que verifique la condición de k-anonimato. Un fichero de microdatos es k-anónimo si cada registro es indistinguible dentro de un conjunto de k registros en lo que respecta a los cuasi-identificadores. De esta manera, la posibilidad de reidentificación se limita a conjuntos de k registros; es decir, dado un individuo que sabemos que está en el fichero $T'(A_1, \ldots, A_n)$, podemos a lo sumo determinar un conjunto de k registros entre los que se encuentra su registro. La propuesta original para construir un fichero de microdatos k-anónimo [6] se basaba en reducir la granularidad de la información contenida en los cuasi-identificadores mediante generalización y supresión. Otra propuesta posterior [7], [8] se basa en la técnica de microagregación.

El nivel de protección que ofrece el k-anonimato contra la revelación de atributos es, en general, bastante limitado. Otras propuestas intentan mejorar este déficit del k-anonimato: la l-diversidad [9] requiere la presencia de l valores diferentes en cada uno de los atributos confidenciales para cada grupo de registros que comparten el valor de los cuasi-identificadores; la t-proximidad [10] requiere que la distribución de los atributos confidenciales dentro de cada grupo sea similar a la distribución global.

La aplicación del k-anonimato requiere que sea posible determinar qué atributos son cuasi-identificadores; es decir, qué atributos están disponibles en un conjunto externo y no anónimo de datos. En el resto del artículo trabajaremos en dos escenarios diferentes:

- Intruso no informado. El intruso no conoce el contenido de ningún atributo confidencial para ningún individuo.
- Intruso informado. El intruso puede conocer el valor que alguno de los atributos confidenciales toma para alguno de los individuos. Esto puede suceder, por ejemplo, si el intruso conoce a alguno de los individuos del fichero de microdatos.

El caso de un intruso no informado es el más sencillo: los atributos pueden ser cuasi-identificadores o confidenciales, pero no ambas cosas. En el caso de un intruso informado, la intersección entre los conjuntos de cuasi-identificadores y de atributos confidenciales es no nula. Los atributos confidenciales de los cuales el intruso pueda conocer información deben considerarse también como cuasi-identificadores, pues de otra manera el intruso podría usar la información de que dispone para realizar una reidentificación más precisa. En el escenario con intruso informado cabe la presencia de múltiples intrusos, cada uno con acceso a una determinada información externa. Nosotros supondremos que hay tantos intrusos como atributos confidenciales, y que cada intruso conoce el contenido de todos los atributos confidenciales excepto uno, del cual no tiene ningún conocimiento. También supondremos que no hay ninguna confabulación entre los intrusos, pues la confabulación de dos de dichos intrusos permitiría determinar la información confidencial de todos los usuarios sin necesidad del conjunto de microdatos. Consideramos que este escenario

es suficientemente estricto, pues plantea un conjunto de intrusos con mucho conocimiento, aunque entendemos que aún es posible dotar a los intrusos de un mayor conocimiento.

III. k-Anonimato probabilístico

Hemos visto que el k-anonimato requiere que cada registro sea indistinguible respecto de los cuasi-identificadores dentro de un conjunto de al menos k registros. De esta manera, no es posible reidentificar un individuo concreto; a lo sumo, podemos determinar un conjunto de k registros dentro del cual se encuentra. Esto quiere decir que la probabilidad de realizar una reidentificación es como mucho $1/k$.

El k-anonimato probabilístico relaja los requisitos de indistinguibilidad impuestos por la noción de k-anonimato, buscando únicamente garantizar la misma probabilidad de reidentificación que ofrece el k-anonimato, esto es $1/k$. Igual el k-anonimato estándar, el k-anonimato probabilístico se obtiene enmascarando el valor que toman los cuasi-identificadores. Esto no significa que los valores de los atributos confidenciales queden intactos, pues dependiendo de los intrusos que consideremos puede haber atributos que son cuasi-identificadores y confidenciales a la vez.

Un concepto similar al de k-anonimato probabilístico se presentó en [11], donde el conjunto de registros se particiona en grupos de cardinalidad k, y después se aplica una permutación sobre cada uno de estos grupos. Esta es la misma estrategia que aplicaremos para obtener k-anonimato probabilístico en la sección IV. Sin embargo, la noción de k-anonimato probabilístico es más general que la presentada en [11], pues permite el uso de cualquier técnica, mientras se consiga la probabilidad de reidentificación deseada. Proponemos el uso de permutaciones pues éstas simplifican los cálculos de probabilidades. Además, [11] considera únicamente un atributo confidencial, mientras que nosotros consideramos el caso de múltiples atributos confidenciales que pueden ser a la vez cuasi-identificadores.

Definición 1 (k-Anonimato probabilístico)**.** Sea $T(A_1, \ldots, A_n)$ un fichero de microdatos y $T'(A_1, \ldots, A_n)$ el fichero resultante de aplicar un mecanismo de anonimización M sobre $T(A_1, \ldots, A_n)$. Sea $\mathcal{I}$ el conjunto de intrusos para los que queremos obtener protección. Para cada $I \in \mathcal{I}$, sea E_I el fichero de datos externo y no anónimo disponible para el intruso I. Decimos que $T'(A_1, \ldots, A_n)$ satisface k-anonimato probabilístico para el conjunto de intrusos $\mathcal{I}$ si, para cada intruso $I \in \mathcal{I}$, con conocimiento de T',M y E_I, la probabilidad de re-identificar un registro correctamente es a lo sumo $1/k$.

Si un método de anonimización conduce al k-anonimato entonces también conduce al k-anonimato probabilístico. En este sentido, se puede pensar que el k-anonimato ofrece una garantía de privacidad más fuerte. Sin embargo, ambos métodos ofrecen la misma probabilidad de que un intruso pueda reidentificar un registro.

La ventaja que ofrece el k-anonimato probabilístico en comparación con el k-anonimato estándar es que, relajando el requisito de indistinguibilidad, el repertorio de métodos que podemos utilizar se amplía y, por lo tanto, podemos buscar métodos con una perdida de información menor.

Como el k-anonimato probabilístico se expresa en términos de probabilidades, es natural pensar en el fichero de microdatos publicados $T'(A_1, \ldots, A_n)$ como una perturbación del fichero original $T(A_1, \ldots, A_n)$. Usaremos la notación de la Figura 1. Hemos separado los registros x_i de T en dos partes: los cuasi-identificadores qi_i, y los atributos confidenciales c_i. Los registros en T' se obtienen del los correspondientes registros de T aplicando una perturbación: $x'_i = X(x_i)$. El intruso tiene acceso al fichero de datos no anónimos E, que utiliza para asignar una identidad a los registros de T'. Conviene notar que el intruso solamente puede determinar la identidad de los usuarios contenidos en E, y por tanto la relación se establece entre registros en T' e identidades en E. La función Rid asigna a cada identidad en E el registro en T' correspondiente a la reidentificación hecha por el intruso. La función Id asigna a cada identidad en E el registro en T' correspondiente a esa identidad. Las funciones Rid y Id retornan $\emptyset$ si el intruso no ha asignado dicha identidad a ningún registro de T', o si no hay ningún registro en T' que se corresponda con dicha identidad, respectivamente.

El objetivo del k-anonimato probabilístico es limitar la probabilidad de enlazar los registros en T' con sus respectivas identidades en E. Con las notaciones definidas anteriormente, esto se puede expresar como: para todo $e_i \in E$ y para toda función de re-identificación Rid

$$P(Rid(e_i) = Id(e_i)) \leq \frac{1}{k}$$

Esta fórmula dice que, cualquiera que sea la función que el intruso usa para re-identificar los registros, la probabilidad de que la re-identificación de un registro sea correcta es a lo sumo $1/k$. La simplicidad de la fórmula proviene de que consideramos posible cualquier función de re-identificación. Esto oculta el comportamiento que tendría un intruso con un comportamiento racional. Dado un $e_i \in E$, un intruso racional toma como $Rid(e_i)$ el registro de T' que tiene mayor probabilidad, dado su conocimiento de T',E y M. Los siguientes ejemplos clarifican el comportamiento de un intruso racional. En ambos ejemplos se considera que E contiene la identidad de todos los registros en T, que es el mayor nivel de conocimiento que se puede esperar de un intruso.

Ejemplo 2. Suponemos que T, T' y E contienen la información que muestra el Cuadro I. Desde el punto de vista del intruso, la identidad del registro x'_1 debe ser una de las e_i. Un intruso racional seleccionaría la identidad que tenga mayor probabilidad, dado su conocimiento de T', E y M.

La probabilidad de que x'_1 se corresponda con e_i, , $P(x'_1 \sim e_i)$, es la probabilidad de obtener qi'_1 a partir de qi_i^E, dividida por la probabilidad total de obtener qi'_1 a partir de cualquier registro de E.

$$P(x'_1 \sim e_i) = \frac{P(X'(qi_i^E) = qi'_1|M)}{\sum_{(qi_j^E, id_j) \in E} P(X'(qi_j^E) = qi'_1|M)}$$

$T(A_1,\ldots,A_n)$	$\xrightarrow{X'}$	$T'(A_1,\ldots,A_n)$	$\xleftarrow{Id,\ Rid}$	E
$x_1 = (qi_1, c_1)$		$x'_1 = (qi'_1, c_1)$		$e_1 = (qi_1, id_1)$
$\vdots$		$\vdots$		$\vdots$
$x_{\|T\|} = (qi_{\|T\|}, c_{\|T\|})$		$x'_{\|T'\|} = (qi'_{\|T'\|}, c_{\|T\|})$		$e_{\|E\|} = (qi_{\|E\|}, id_{\|E\|})$

Figura 1. Notaciones para k-anonimidad probabilística

Cuadro I
FICHEROS DE DATOS EN EL EJEMPLO 2

T	T'	E
$x_1 = (qi_1, c_1)$	$x'_1 = (qi'_1, c_1)$	$e_1 = (qi_1^E, id_1)$
$\vdots$		$\vdots$
$x_N = (qi_N, c_N)$		$e_N = (qi_N^E, id_N)$

En el ejemplo anterior hemos visto que el enlace entre registros de T' e identidades en E busca maximizar la probabilidad de acierto. Para que el máximo de las probabilidades de que un registro en T' y una identidad en E se correspondan sea menor que $1/k$, es necesario que todas las probabilidades de correspondencia sean menores que $1/k$. Es decir, para tener k-anonimato probabilístico debemos tener para todo $e \in E$ y para todo $x' \in T$

$$P(x' \sim e) \leq \frac{1}{k} \tag{1}$$

El cálculo de la probabilidad $P(x' \sim e)$ puede ser complejo. En el Ejemplo 2 hemos visto como calcular esta probabilidad en un caso particular pero el cálculo de esta probabilidad en general puede ser complejo. En la siguiente sección proponemos usar un mecanismo M basado en permutaciones, cosa que simplifica estos cálculos.

IV. k-ANONIMATO PROBABILÍSTICO MEDIANTE AGRUPACIÓN E INTERCAMBIO

El método que proponemos consta de dos fases: (i) particionar el conjunto de registros de T en grupos de k elementos y (ii) aplicar una permutación sobre los elementos de cada uno de los grupos. El primer punto es el más importante, y admite muchas variaciones en función de como se haga la agrupación de los registros de T.

Conviene observar que como la misma permutación se aplica sobre todos los cuasi-identificadores, no estamos ocultando la identidad del individuo. Sin embargo, estamos disociando la identidad de los valores de los atributos confidenciales, y de esta manera limitando a $1/k$ la probabilidad de que un intruso pueda adivinar el valor de los atributos confidenciales. Si queremos evitar que se mantenga la relación entre los cuasi-identificadores, alguno de ellos debe ser considerado como un atributo confidencial.

Primero introduciremos el método que proporciona protección contra intrusos no informados, y después lo extenderemos para intrusos informados.

IV-A. Intrusos no informados

Consideramos que T, T' y E contienen los registros que se muestran en el Cuadro II. Es fácil comprobar que el método de particionado e intercambio descrito anteriormente cumple la desigualdad (1) requerida por el k-anonimato probabilístico.

$$P(x' \sim e_i) = \begin{cases} 1/k & \text{si } x' \in G(id(e_i)) \\ 0 & \text{si no} \end{cases}$$

donde $G(id(e_i))$ es el grupo de registros de T que contiene el registro correspondiente a la identidad e_i.

Cuadro II
FICHEROS DE DATOS EN EL ESCENARIO CON INTRUSOS NO INFORMADOS

T	T'	E
$x_1 = (qi_1, c_1)$	$x'_1 = (qi'_1, c_1)$	$e_1 = (qi_1^E, id_1)$
$\vdots$	$\vdots$	$\vdots$
$x_N = (qi_N, c_N)$	$x'_N = (qi'_N, c_N)$	$e_N = (qi_N^E, id_N)$

Para minimizar la perdida de información conviene que el particionado dé como resultado unos grupos de registros lo más homogéneos posible. Nuestras simulaciones se han basado en el algoritmo MDAV [7], [8] sobre el conjunto de cuasi-identificadores.

IV-B. Intrusos informados

El escenario con intrusos informados presentado en la Sección II consideraba un número igual de intrusos al de atributos confidenciales, de manera que cada intruso conocía todo el contenido de todos los atributos confidenciales excepto de uno, del cual no conocía nada. Para fijar ideas, los atributos serán $A_0, A_1, \ldots, A_n$, siendo A_0 un cuasi-identificador no confidencial, y $A_1, \ldots, A_n$ cuasi-identificadores confidenciales. El intruso I_i conoce el valor de todos los atributos excepto de A_i.

El método para intrusos informados consiste en aplicar repetidas veces el método para intrusos no informados, una vez para cada intruso. Cada vez que aplicamos el método para un intruso I_i estamos disociando el contenido del atributo confidencial A_i del resto de atributos (cuasi-identificadores). El Cuadro III resume el número de veces que hay que aplicar el método y qué atributos hay que considerar como confidenciales y cuáles como cuasi-identificadores.

La aplicación del método para intrusos no informados repetidas veces tiene un problema: hay que aplicar diferentes permutaciones sobre grupos diferentes pero no disjuntos de atributos (los cuasi-identificadores correspondientes a cada intruso). Para evitar este problema, en vez de aplicar una

Cuadro III
CUASI-IDENTIFICADORES Y ATRIBUTOS CONFIDENCIALES PARA CADA INTRUSO INFORMADO

Intruso	Atributos cuasi-identificadores	Atributos confidenciales
I_1	$A_0, A_2, \ldots, A_n$	A_1
I_2	$A_0, A_1, A_3 \ldots, A_n$	A_2
$\vdots$	$\vdots$	$\vdots$
I_n	$A_0, A_1, \ldots, A_{n-1}$	A_n

permutación σ sobre el conjunto de cuasi-identificadores, aplicamos la permutación inversa σ^{-1} sobre el atributo confidencial. De esta manera cada permutación se aplica sobre un atributo diferente.

La observación anterior acerca de la aplicación de la permutación inversa sobre el atributo confidencial desconocido lleva a usar microagregación por ordenación individual. En este caso, hacer la agrupación teniendo en cuenta únicamente este atributo conduce a una reducción importante de la pérdida de información, pues el intercambio de valores para este atributo se produce entre grupos de máxima homogeneidad (ver [12] acerca de la pérdida de información derivada de la microagregación por ordenación individual).

Se puede argumentar que, aplicando este tipo de agrupación sobre los registros, el problema de la revelación de atributos en el fichero de microdatos resultante se ve incrementado respecto al caso de agrupación por cuasi-identificadores. Sin embargo, es posible mitigar este problema incrementado la k. De esta manera podemos obtener un nivel similar de protección contra la revelación de atributos, a la vez que incrementamos la garantía proporcionada por el k-anonimato probabilístico.

Un beneficio adicional de agrupar por ordenación individual es que, como se trabaja sobre cada atributo independientemente, se evitan los problemas de variación de escala entre atributos diferentes, y por tanto no es necesario aplicar ningún tipo de normalización a los atributos antes de agrupar.

V. RESULTADOS EXPERIMENTALES

Hemos implementado los tres métodos siguientes:

- *MDAV-ID*. Se aplica el algoritmo de microagregación MDAV sobre los cuasi-identificadores para particionar el fichero de datos en grupos de k elementos. En cada uno de estos grupos, los valores de los cuasi-identificadores se remplazan por el centroide, de manera que todos los registros de un grupo tengan el mismo valor para los cuasi-identificadores. Este procedimiento fue propuesto en [8] y con él se obtiene la noción estándar de k-anonimato como fue propuesta en [6].
- *MDAV-SWAP*. Se aplica el método descrito en la Sección IV-B utilizando MDAV para agrupar sobre los cuasi-identificadores y permutando el contenido del atributo confidencial dentro de cada uno de los grupos.
- IR-SWAP. Se aplica el método descrito en la Sección IV-B agrupando por el atributo confidencial y permutando el contenido de dicho atributo dentro de cada uno de los grupos.

Las simulaciones se han hecho sobre los conjuntos de datos de referencia "Census" y "EIA", propuestos en el proyecto CASC [13]. Los datos en el fichero "Census" están más esparcidos, mientras que en el fichero "EIA" están más agrupados. El fichero de datos "Census" contiene 1080 registros con 13 atributos continuos. Para realizar las simulaciones consideramos los seis primeros atributos como cuasi-identificadores no confidenciales y los siete últimos como confidenciales.

Para evaluar la calidad de de los datos, comparamos las correlaciones entre los atributos entre el fichero de microdatos original y el fichero resultante de aplicar los algoritmos MDAV-ID, MDAV-SWAP y IR-SWAP. Como MDAV-SWAP y IR-SWAP mantienen constantes los atributos no confidenciales, la correlación entre estos será la misma que en el fichero original. En la comparación solo consideraremos las correlaciones entre atributos confidenciales.

Otra característica que podemos destacar de MDAV-SWAP y IR-SWAP es que mantienen la media y la varianza (tanto la global como la de cada grupo), pues se mantienen los valores originales de cada atributo, aunque permutados.

Para obtener resultados con una cierta significación estadística, hemos ejecutado los algoritmos MDAV-ID, MDAV-SWAP y IR-SWAP 100 veces, y hemos tomado la media de las observaciones. Los Cuadros IV y V muestran la media y la desviación estándar del valor absoluto de la diferencia entre las correlaciones de los atributos confidenciales en el fichero original y las correlaciones entre atributos confidenciales en cada uno de los ficheros anonimizados resultantes. Valores más cercanos a cero en el cuadro indican una mayor calidad de datos. Un valor cercano a uno para la media significa que la mayor parte de la dependencia entre los atributos se ha perdido.

Los resultados de la simulación muestran que, en efecto, MDAV-SWAP y IR-SWAP ofrecen una calidad de datos mayor a la proporcionada por MDAV-ID. Si bien la diferencia entre MDAV-ID y MADV-SWAP es pequeña, la mejora proporcionada por IR-SWAP es bastante significativa. Por ejemplo, para el conjunto de datos "Census" se han obtenido valores similares en el Cuadro IV usando MDAV-ID con $k = 11$, MDAV-SWAP con $k = 25$ y IR-SWAP con $k = 300$.

Cuadro IV
MEDIA Y DESVIACIÓN ESTÁNDAR DEL VALOR ABSOLUTO DE LA DIFERENCIA EN LA CORRELACIÓN ENTRE ATRIBUTOS CONFIDENCIALES ENTRE EL FICHERO "CENSUS" ORIGINAL Y EL ANONIMIZADO

	MDAV-ID		MDAV-SWAP		IR-SWAP	
k	mean	st.dev.	mean	st.dev.	mean	st.dev.
5	.055	.064	.037	.045	.0021	.0041
7	.062	.071	.048	.056	.0022	.0039
9	.069	.078	.055	.064	.0028	.0049
11	.078	.085	.061	.070	.0038	.0068
25	.11	.11	.091	.093	.0061	.012
50	.14	.13	.13	.12	.010	.020
100	.17	.15	.19	.17	.020	.030
200	.29	.27	.31	.28	.044	.047
300	.38	.39	.37	.34	.087	.071

Cuadro V
MEDIA Y DESVIACIÓN ESTÁNDAR DEL VALOR ABSOLUTO DE LA DIFERENCIA EN LA CORRELACIÓN ENTRE ATRIBUTOS CONFIDENCIALES ENTRE EL FICHERO "EIA" ORIGINAL Y EL ANONIMIZADO

	MDAV-ID		MDAV-SWAP		IR-SWAP	
k	mean	st.dev.	mean	st.dev.	mean	st.dev.
5	.018	.017	.017	.035	.00064	.00075
7	.02	.017	.024	.05	.0012	.0018
9	.034	.031	.028	.053	.0015	.0018
11	.039	.036	.029	.052	.0019	.0023
25	.085	.078	.043	.081	.0063	.0072
50	.13	.12	.053	.089	.011	.011
100	.15	.14	.058	.092	.029	.037
200	.19	.18	.09	.11	.093	.074
300	.2	.18	.12	.13	.14	.091

VI. CONCLUSIONES

El k-anonimato es una propiedad ampliamente utilizada en la protección de microdatos contra revelación. En un fichero de datos k-anónimo, los registros están agrupados en grupos de k elementos, de manera que todos los registros de un grupo tienen los mismos valores para los cuasi-identificadores; esto es, no es posible diferenciar un registro de otro del mismo grupo utilizando los cuasi-identificadores. Por tanto, el k-anonimato implica una pérdida de variabilidad en los valores de los cuasi-identificadores, y en consecuencia una pérdida de calidad de los datos. Esta pérdida tiene especial relevancia en presencia de intrusos informados, pues los atributos que conoce el intruso deben considerarse como cuasi-identificadores.

Para intentar mitigar la pérdida de información, introducimos el concepto de k-anonimato probabilístico. Dicho concepto reduce los requisitos de indistinguibilidad de los registros (pudiendo de esta manera mantener la variabilidad dentro de cada grupo) a la vez que garantiza que la probabilidad de reidentificar un registro sea a lo sumo de $1/k$, la misma que proporciona el k-anonimato estándar. Con esta reducción en los requisitos, ampliamos el repertorio de métodos aplicables y, por tanto, podemos buscar métodos que tengan una perdida de información menor.

Hemos propuesto dos métodos para obtener k-anonimato probabilístico: MDAV-SWAP y IR-SWAP. Ambos se basan en agrupación e intercambio, pues aplicando estas técnicas es sencillo garantizar que se cumplen los requisitos del k-anonimato probabilístico, que de otra manera requerirían cálculos de probabilidades que pueden ser complejos. Los resultados experimentales muestran que los métodos propuestos para k-anonimato probabilístico ofrecen mejor calidad de datos que los métodos para k-anonimato estándar.

Un problema que no está resuelto en k-anonimato probabilístico es el de revelación de atributos. Un punto interesante sería combinar k-anonimato probabilístico con otras propiedades como l-diversidad o t-proximidad.

AGRADECIMIENTOS

Los dos primeros autores están encuadrados en la Cátedra UNESCO de Privacidad de Datos, pero los puntos de vista expresados en este artículo no necesariamente reflejan la posición de UNESCO. Este trabajo ha recibido apoyo del Gobierno de España mediante los proyectos TSI2007-65406-C03-01 "E-AEGIS", TIN2011-27076-C03-01 "CO-PRIVACY", CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES" y TEC2010-20572-C02-02 "Consequence", de la Generalitat de Catalunya mediante las ayudas 2009 SGR 1135 y 2009 SGR 1362, y de la Comisión Europea mediante el proyecto del 7PM "DwB". Prof. Domingo-Ferrer está parcialmente financiado como investigador ICREA Acadèmia por la Generalitat de Catalunya. D. Rebollo-Monedero disfruta de una beca posdoctoral Juan de la Cierva, JCI-2009-05259, otorgada por el Ministerio Español de Ciencia e Innovación.

REFERENCIAS

[1] J. Domingo-Ferrer, "A Survey of Inference Control Methods for Privacy-Preserving Data Mining", en *P*rivacy-Preserving Data Mining, ser. *A*dvances in Database Systems. Springer, vol. 34, pp. 53–80, 2008.

[2] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments", en *A*CM Computing Surveys, vol. 42, no. 4, pp. 14:1–14:53, 2010.

[3] V. Ciriani, S. D. C. di Vimercati, S. Foresti, and P. Samarati, "Microdata protection", en *S*ecure Data Management in Decentralized Systems, ser. *A*dvances in Information Security, Springer, vol. 33, pp. 291–321, 2007.

[4] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality", en *P*roceedings of the 31st international conference on Very large data bases, pp. 901–909, 2005.

[5] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. Schulte-Nordholt, G. Seri, and P.-P. DeWolf, "Handbook on Statistical Disclosure Control (version 1.2)", ESSNET SDC Project, 2010.

[6] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression", Tech. rep. SRI-CSL-98-04, SRI Computer Science Laboratory, Palo Alto, CA, 1998.

[7] J. Domingo-Ferrer and J. M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control", en *I*EEE Transactions on Knowledge and Data Engineering, vol. 4, no. 1, pp. 189–201, 2002.

[8] J. Domingo-Ferrer and V. Torra, "Ordinal, continuous and heterogenerous k-anonymity through microaggregation", en *D*ata Mining and Knowledge Discovery, vol. 11, no. 2, pp. 195–212, 2005.

[9] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-Diversity: Privacy beyond k-anonymity", en *2*2nd IEEE International Conference on Data Engineering, 2006.

[10] N. Li and T. Li,, "t-Closeness: Privacy beyond k-anonymity and ldiversity", en *P*roceedings of IEEE 23rd IntŠl Conf. on Data Engineering, 2007.

[11] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu, "Aggregate query answering on anonymized tables", en *I*nternational Conference on Data Engineering, pp. 116–125, 2007.

[12] J. Domingo-Ferrer and V. Torra, "A quantitative comparison of disclosure control methods for microdata", en *C*onfidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, Eds., pp. 111–134, 2001.

[13] R. Brand, J. Domingo-Ferrer, and J. M. Mateo-Sanz, "Reference data sets to test and compare SDC methods for protection of numerical microdata", European FP5 Project IST-2000-25069 CASC, 2002.