¿CÓMO MEDIR LA PRIVACIDAD?

David Rebollo-Monedero y Jordi Forné

Grupo de Seguridad de la Información, Departamento de Ingeniería Telemática
Universitat Politècnica de Catalunya (UPC)
Campus Nord, Módulo C5, Despacho S102A
C. Jordi Girona 1-3, E-08034, Barcelona, Spain

Tel.: +34 93 401 7027, e-Mail: {david.rebollo, jordi.forne}@entel.upc.edu

Resumen — En el presente estudio revisamos el estado del arte sobre métricas de privacidad en métodos con perturbación para el control estadístico de revelación. Aunque el artículo se enfoca en microagregación de datos, dichos métodos también son aplicables a una gran variedad de escenarios alternativos, tales como la ofuscación en servicios basados en la localización. Concretamente, examinamos el criterio de k-anonimato y alguna de las propuestas para mejorarlo. Motivados por la vulnerabilidad de estos criterios frente a ataques de similitud y sesgo, comparamos tres recientes métricas de privacidad, basadas en conceptos de teoría de la información, que pretenden resolver dichas vulnerabilidades.

Palabras Clave — Privacidad de la información, control estadístico de revelación, anonimización de microdatos, teoría de la información, k-anonimato, l-diversidad, t-cercanía, δ-revelación.

I. INTRODUCCIÓN

El derecho a la privacidad fue reconocido ya en 1948 por las Naciones Unidas en la Declaración Universal de los Derechos Humanos. Con el crecimiento exponencialmente acelerado de las tecnologías de la información, y la tendencia hacia la adquisición de una presencia virtual en La Red por parte de prácticamente cualquier persona, objeto o entidad, la privacidad innegablemente posee una importancia cada día más decisiva. Motivados por ello, deseamos poder diseñar servicios que protejan la privacidad y poder evaluar su vulnerabilidad frente a ataques, de manera objetiva, sistemática, científica. Pero para convertir una realidad en ciencia, debemos cruzar el puente que conecta lo calificable con lo cuantificable. Así pues, la cuestión es inevitable: ¿cómo medir la privacidad?

Precisamente, el objetivo de este estudio es el de revisar el estado del arte sobre métricas de privacidad en métodos con perturbación para el control estadístico de revelación. Dichos métodos consisten en perturbar la información de los individuos de manera óptima para maximizar la privacidad, manteniendo un cierto grado de utilidad de los datos. Para ello, se aplican potentes conceptos y técnicas derivados de la estadística y la teoría de la información, entre otras doctrinas.

Aunque este artículo se enfoca en microagregación de datos, los métodos con perturbación para la privacidad también son aplicables a una gran variedad de escenarios alternativos, tales como la ofuscación en servicios basados en la localización, la búsqueda en Internet y las redes P2P. Especialmente, examinamos el criterio de k-anonimato y alguna de las propuestas para mejorarlo. Motivados por la vulnerabilidad de estos criterios frente a ataques de similitud y sesgo, comparamos tres recientes métricas de privacidad, basadas en conceptos de teoría de la información, que pretenden resolver dichas vulnerabilidades. En

particular, comparamos el riesgo promedio de privacidad propuesto en [11], t-cercanía [8] y δ-revelación [1].

Ya hemos apuntado que existe un compromiso entre privacidad y utilidad de los datos inherente a cualquier método de privacidad con perturbación. Naturalmente, una especificación completa del problema de optimización que contemple dicho compromiso requiere así mismo una especificación de una métrica de utilidad de los datos. Por otro lado, la solución del problema de optimización puede estar lejos de ser trivial. En cualquier caso, para no extendernos excesivamente perdiendo el enfoque, reduciremos el alcance de nuestro estudio a las métricas de privacidad.

El resto del presente estudio está organizado de la manera siguiente. La Sección II describe dos escenarios de aplicación. La Sección III revisa el estado del arte sobre métricas de privacidad para CER. La Sección IV proporciona un análisis en mayor profundidad sobre tres de los criterios basados en teoría de la información para medir la privacidad. Las conclusiones se exponen en la Sección V.

II. ESCENARIOS DE APLICACIÓN

Esta sección motiva la importancia del control de revelación de información en lo que respecta a la privacidad, introduciendo dos problemas relacionados: anonimización de microdatos y la recuperación privada de información basada en la localización.

A. Anonimización de Microdatos

Un *conjunto de microdatos* es una base de datos cuyos registros proporcionan información concerniente a individuos encuestados, ya sea personas físicas o empresas. Este conjunto habitualmente contiene *atributos clave* o *cuasi identificadores*, es decir, atributos que, conjuntamente, pueden enlazarse con información externa para reidentificar los individuos a los que los registros del conjunto de microdatos hacen referencia. Ejemplos incluyen ocupación, dirección, edad, sexo, altura y peso. Además, el conjunto de datos contiene *atributos confidenciales* con información aún más delicada sobre el individuo, como salario, religión, afiliación política o estado de salud. La clasificación de atributos como clave o confidenciales puede depender en última instancia de la aplicación específica y de los requisitos de privacidad para los cuales el conjunto de microdatos está pensado.

Intuitivamente, la perturbación de atributos clave nos permite preservar la *privacidad* hasta un cierto punto, a costa de perder *utilidad de los datos* con respecto a la versión sin perturbar. k-Anonimato es el requisito de que cada tupla de atributos clave sea compartida por al menos k registros en el conjunto de datos. Esto puede conseguirse mediante la alternativa de microagregación ilustrada en el ejemplo de la Fig. 1, donde altura y peso se consideran atributos clave, y la concentración de colesterol (lipoproteínico de baja densidad) en la sangre, un atributo confidencial. En lugar de proporcionarse la tabla original, se publica una versión k-anónima conteniendo registros agregados, en el sentido de que los atributos clave de cada grupo son remplazados por una tupla común representativa. A pesar del hecho de que el k-anonimato como medida de privacidad no está libre de inconvenientes, su simplicidad lo convierte en un criterio ampliamente popular en la literatura de *control estadístico de revelación* (CER).

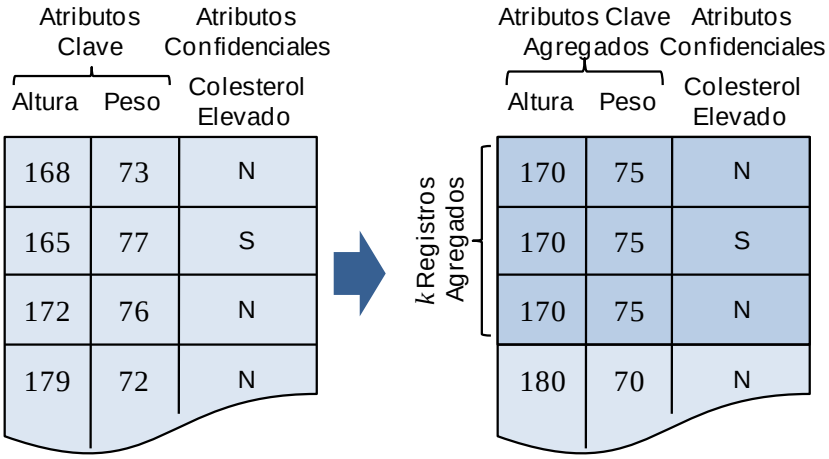

Fig. 1. Microagregación de valores de atributos claves para alcanzar *k*-anonimato.

B. Privacidad en Servicios Basados en la Localización

El problema de anonimización de microdatos que hemos motivado surge, al menos conceptualmente, en un amplio rango de diversas aplicaciones. Un ejemplo de particular relevancia es el de *servicios basados en la localización* (SBL). En la forma más simple de interacción entre un usuario y un proveedor de SBL interviene un mensaje directo del primero al segundo conteniendo una consulta y la ubicación a la que la consulta se refiere. Un ejemplo sería la consulta "¿Dónde está el banco más cercano a mi casa?", acompañada por las coordenadas geográficas de la residencia del usuario, o simplemente de su dirección. Bajo la suposición de que el sistema de comunicación utilizado permite al proveedor de SBL reconocer el ID del usuario, existe un riesgo patente de privacidad. En concreto, el proveedor podría establecer perfiles de usuario de acuerdo con sus ubicaciones, los contenidos de sus consultas y su actividad.

Esencialmente, un método con perturbación análogo a la microagregación de datos puede usarse para hacer frente a este riesgo de privacidad, como se representa en la Fig. 2. En general, los usuarios pueden contactar un proveedor de SBL en el que no confían plenamente, de manera directa, perturbando la información de su ubicación para frenar los esfuerzos del proveedor para vulnerar la privacidad del usuario en términos de localización, aunque no en términos de contenido de la consulta o actividad. Esta alternativa, llamada a veces *ofuscación*, presenta el compromiso entre utilidad de los datos y privacidad inherente a cualquier método de privacidad con perturbación. El paralelismo con la anonimización de microdatos puede ahora dibujarse simplemente al identificar ID de usuario e información de localización con atributos confidenciales y clave, respectivamente.

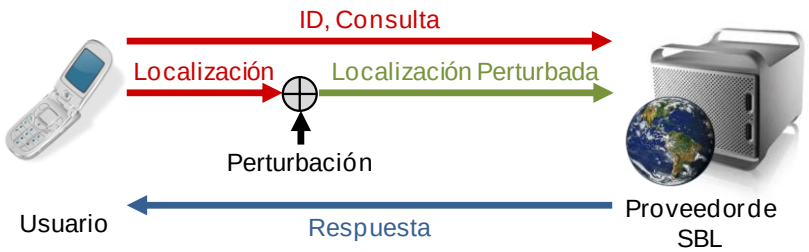

Fig. 2. Los usuarios pueden contactar un proveedor de SBL en el que no confían, directamente, perturbando su información de localización para ayudar a proteger su privacidad.

III. EL *k*-ANONIMATO Y SUS VARIANTES COMO MÉTRICAS DE PRIVACIDAD EN EL CONTROL ESTADÍSTICO DE REVELACIÓN

Ya mencionamos en la Sección II.A que una porción específica de datos sobre un grupo particular de individuos satisface el requisito de *k*-anonimato (para un entero positivo *k*) si el origen de cualquiera de sus componentes no puede determinarse más allá de un subgrupo de al menos *k* individuos. También apuntamos que el concepto de *k*-anonimato, originalmente propuesto por la comunidad de CER [13], es un criterio de privacidad muy popular, en parte debido a su manejabilidad matemática.

La formulación original de este criterio de privacidad, basada en la generalización y el registro de atributos clave, fue transformada en la propuesta basada en microagregación que ya hemos comentado, e ilustrado en la Fig. 1, en [3]. Ambas formulaciones pueden entenderse como casos particulares de una más general que utiliza una medida de distorsión abstracta entre los datos perturbados y los datos sin perturbar, posiblemente tomando valores en alfabetos significativamente distintos.

Se ha demostrado que la microagregación multivariante es NP-difícil. Se han propuesto una serie de métodos heurísticos, que se pueden clasificar en métodos de tamaño fijo y de tamaño variable, dependiendo de si todos los grupos excepto uno tienen exactamente *k* elementos con un valor común de atributo clave. El algoritmo de máxima distancia (MD) y una variación menos exigente computacionalmente, llamada algoritmo de máxima distancia al vector medio (MDAV) [6], son algoritmos de tamaño fijo que funcionan particularmente bien en términos de la distorsión que introducen, para un gran número de distribuciones de datos. El algoritmo de Lloyd de probabilidades restringidas (PCL) [12] es una reciente propuesta que extiende el algoritmo de Lloyd-Max, un famoso algoritmo de compresión de datos que frecuentemente produce agrupaciones óptimas.

Desafortunadamente, aunque el *k*-anonimato impide la revelación de la identidad, todavía puede fracasar en la protección contra la revelación del atributo. Más concretamente, la definición de este criterio establece que la reidentificación completa no es factible dentro de un grupo de registros que compartan la misma tupla de valores de atributos claves perturbados. Sin embargo, si los registros en el grupo también comparten un valor común de algún atributo confidencial, la asociación entre un individuo enlazable al grupo de atributos clave perturbados y los atributos confidenciales correspondientes permanece revelada, como ilustra el ejemplo de la Fig. 3. Con mayor generalidad, el problema con el *k*-anonimato como criterio de privacidad es su vulnerabilidad frente a la explotación de la diferencia entre la distribución *a priori* de los datos confidenciales de toda la población, y su distribución condicional o *a posteriori* en un grupo, dados los atributos clave perturbados observados. Por ejemplo, imaginemos que en la Fig. 1 la proporción de individuos encuestados con el colesterol elevado es mucho mayor que en los datos generales de la población. Esto se conoce como *ataque de sesgo*.

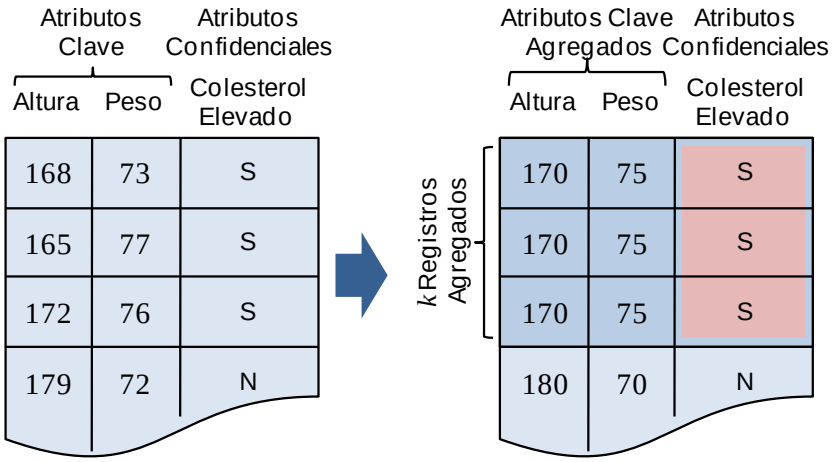

Fig. 3. El k-anonimato de los atributos clave no siempre garantiza la confidencialidad.

Dichas vulnerabilidades han motivado la propuesta de mejoras en criterios de privacidad, alguna de las cuales procedemos a esbozar brevemente, junto con modificaciones de algoritmos. Una restricción del k-anonimato llamada k-anonimato p-sensible fue presentada en [15]. Además del requisito de k-anonimato, se requiere que haya al menos p valores diferentes para cada atributo confidencial dentro del grupo de registros que comparten la misma tupla de valores de atributo clave perturbado. Evidentemente, grandes valores de p pueden perfectamente conducir a una pérdida prohibitiva en utilidad de los datos. Una generalización menor llamada l-diversidad [10] ha sido definida con el mismo propósito de mejorar el k-anonimato. La diferencia con respecto a p-sensibilidad consiste en que el grupo de registros debe contener al menos l valores "muy representados" para cada atributo confidencial. Dependiendo de la definición de "muy representado", l-diversidad puede convertirse en k-anonimato p-sensible o ser más restrictiva. Nos gustaría subrayar que ninguna de esas mejoras logra eliminar completamente la vulnerabilidad del k-anonimato frente a ataques de sesgo. Es más, ambas siguen siendo susceptibles a *ataques de similitud*, en el sentido de que mientras que los valores de los atributos confidenciales dentro de un grupo de registros agregados puede ser p-sensible o l-diversa, pueden también ser semánticamente similares, por ejemplo enfermedades o salarios parecidos.

Un criterio de privacidad encaminado a superar los ataques de similitud y sesgo es t-cercanía [8]. Un conjunto de microdatos perturbados satisface la propiedad de t-cercanía si por cada grupo con valores de atributos claves comunes, cierta distancia entre la distribución a posteriori de los atributos confidenciales en el grupo y la distribución a priori de la población en general no excede un umbral t. En la medida en que la distribución intragrupo de atributos confidenciales se asemeja a la distribución de esos atributos en el conjunto entero de datos, los ataques de sesgo se verán frustrados. Por otro lado, como la distribución intragrupo de atributos confidenciales mimetiza la distribución de esos atributos en el conjunto entero, no puede ocurrir ninguna similitud semántica que no apareciese ya en los datos globales.

La principal limitación del trabajo [8] sobre t-cercanía es la falta de especificación de un procedimiento computacional para alcanzarla. Un criterio de privacidad basado en teoría de la información, inspirado en la t-cercanía, fue propuesto en [11]. En este último trabajo, el riesgo de privacidad se define como una medida de teoría de la información sobre la discrepancia entre las distribuciones a priori y a posteriori. Conceptualmente, el riesgo de

privacidad definido puede considerarse como una versión promediada del requisito de t-cercanía, sobre todos los grupos de registros agregados. Es importante recalcar igualmente que el criterio de privacidad de [11], a pesar de su conveniente manejabilidad matemática, como cualquier criterio basado en promedios, puede no adecuarse a todas las aplicaciones. Un criterio relacionado aunque más conservador, llamado δ-revelación, se ha propuesto en [1], y mide la diferencia máxima entre las distribuciones a priori y a posteriori. El riesgo de privacidad promedio de [11], t-cercanía y δ-revelación se estudian con mayor detalle en la Sección IV.

En cuanto al paralelismo con SBL establecido en la Sección II.B nos gustaría recalcar que se ha propuesto una amplia variedad de métodos con perturbación para la recuperación privada de información basada en la localización [7]. No es sorprendente que algunos empleen el criterio de k-anonimato como medida de privacidad. Un ejemplo ilustrativo se presenta en [5]. Fundamentalmente, k usuarios añaden ruido aleatorio de media nula a sus coordenadas y comparten el resultado para la computación del promedio, que constituye una localización común perturbada enviada al proveedor de SBL. Desafortunadamente, algunos de esos usuarios pueden utilizar cancelación de ruido para intentar revelar la ubicación de un usuario que se desplace lentamente. En [12] se propone un anonimizador de localizaciones que agrupa ubicaciones exactas para proporcionar k-anonimato en SBL mediante el algoritmo PCL.

IV. MÉTRICAS DE PRIVACIDAD BASADAS EN TEORÍA DE LA INFORMACIÓN

La siguiente es una discusión en mayor profundidad sobre algunas de las métricas de privacidad propuestas más recientemente, basadas en conceptos de teoría de la información, que intentan contrarrestar las vulnerabilidades del k-anonimato y de sus mejoras. A pesar de que las métricas son novedosas, así como las formulaciones matemáticas correspondientes del problema de anonimización de microdatos que las usan, junto con sus soluciones, veremos que las métricas en sí están estrechamente relacionadas con conceptos ya propuestos por Shannon en los años cincuenta.

Nuestro análisis será exclusivamente conceptual. Por lo tanto, es suficiente recordar que la *entropía* de una variable estadística es una medida de su incertidumbre, que la *información mutua* entre dos variables estadísticas es una medida de la información que una contiene sobre la otra, y que la *divergencia* de Kullback-Leibler (KL) es una medida de discrepancia entre distribuciones de probabilidad. Se aconseja a aquellos lectores interesados en las definiciones matemáticas de estas medidas de teoría de la información que consulten [2].

A. Riesgo de Privacidad, Equivocación de Shannon y Ganancia de Información

En el problema de anonimización de microdatos introducidos en la Sección II.A, modelamos (tuplas de) atributos confidenciales mediante una variable estadística W, con una distribución de probabilidad correspondiente. (Tuplas de) atributos clave se representan por una variable X, y son perturbadas para producir (tuplas de) datos $\hat{X}$ ligeramente modificadas. En lugar de publicar la tabla, o con mayor generalidad, la distribución de

probabilidad conteniendo X y W, se publica la versión *saneada* con $\hat{X}$ y W. Dado que una tabla puede considerarse como la definición de una distribución de probabilidad empírica, nuestro modelo es ligeramente más general. Recordemos que nuestro objetivo es el de frenar los esfuerzos de los atacantes para enlazar la identidad de los encuestados con los datos confidenciales.

Consideremos ahora, por un lado, la distribución a priori de los atributos confidenciales W, y por el otro, la distribución a posteriori o condicional de W dados los atributos perturbados $\hat{X}$. Siempre que la distribución a posteriori difiera de la distribución a priori, habremos de hecho ganado información sobre los individuos estadísticamente enlazados a los atributos clave perturbados, en contraste con las estadísticas de la población general. En términos del ejemplo ilustrado en la Fig. 1, la probabilidad de colesterol elevado en la población puede ser, digamos del 25%, mientras que la probabilidad de colesterol elevado dentro del grupo correspondiente a una altura y peso quantificados de 170 cm y 75 kg respectivamente, es aproximadamente de 33%. Intuitivamente, un individuo de altura y peso conocidos que cuadre en dicha categoría tiene mayor probabilidad de padecer problemas de colesterol de lo que se podría haber inferido exclusivamente a partir de la distribución de la población. Esta situación debe reconocerse como un riesgo de privacidad, aunque no sea tan grave como el ilustrado en la Fig. 3.

Para cuantificar la intuición anterior, en primer lugar recordemos el concepto de *equivocación* introducido por Shannon en 1949 [14], es decir, la entropía condicional de un mensaje privado dado un criptograma observado. La aplicación del principio de equivocación de Shannon a privacidad no es en absoluto novedosa. Por ejemplo, en [4], el grado de anonimato observable por un atacante se mide como la entropía de la distribución de probabilidad de los posibles remitentes de un mensaje determinado. Conceptualmente, y con un grado ligeramente mayor de generalidad, consideraremos la equivocación de Shannon como la entropía de la información privada, no observada, dada la información pública, observada.

En términos de nuestra formulación, concordantemente comparamos la entropía de W asociada con la distribución a priori de los atributos confidenciales, con la equivocación, es decir, la entropía W dado $\hat{X}$, asociada con la distribución a posteriori dados los atributos clave perturbados que se observan. La reducción en incertidumbre, es decir, la diferencia de entropías, se toma directamente como medida de riesgo de privacidad en nuestro propio trabajo [11]. Es más, dicho trabajo demuestra que esta reducción de entropía coincide precisamente con la información mutua entre W y $\hat{X}$, que a su vez coincide con la divergencia condicional de KL entre las distribuciones a posteriori y a priori.

Recordemos que la divergencia condicional es una divergencia entre distribuciones condicionales de la variable condicionada, promediada sobre la variable condicionante. En el caso más simple de *microagregación determinística*, donde un valor de X se asigna a un único valor de $\hat{X}$, conceptualmente hablando, el riesgo de privacidad definido es un promedio entre las discrepancias de las distribuciones a posteriori de cada grupo de registros con un valor de $\hat{X}$ común, con respecto a la distribución a priori.

De acuerdo con las propiedades de información mutua y de divergencia de KL, el riesgo de privacidad definido es no negativo, y se anula si y sólo si W y $\hat{X}$ son estadísticamente independientes, o equivalentemente, si las distribuciones a priori y a posteriori coinciden.

Por supuesto, en este caso extremo, la utilidad de los datos publicados se vería gravemente comprometida. En el otro extremo, dejar los datos originales sin distorsión alguna en general perjudicará la privacidad, puesto que en general las distribuciones a priori y a posteriori difieren.

Podemos también remontar a los años cincuenta la interpretación desde el punto de vista de teoría de la información de la divergencia entre las distribuciones a priori y a posteriori, llamada ganancia de información (promedio) en ciertos campos de la estadística [9]. Además de la obra ya citada, existen otros ejemplos donde se usa la entropía de Shannon como una medida de pérdida de información, señalando limitaciones que afectan aplicaciones concretas. Nos gustaría hacer hincapié en que nosotros hemos introducido una divergencia de KL como medida de revelación de información (en lugar de pérdida), en consonancia con la equivalencia entre el caso en el que las distribuciones a priori y a posteriori son iguales, y la ausencia total de riesgo para la privacidad.

Tal vez la característica más interesante del criterio de privacidad de [11] es que conduce a una formulación matemática del compromiso entre privacidad y utilidad que generaliza un problema de teoría de la información bien establecido y exhaustivamente estudiado. A saber, el problema de compresión con pérdidas con un criterio de distorsión, originalmente propuesto por Shannon en 1959 [2].

B. t-Cercanía y δ-Revelación

Hemos mencionado en la Sección A que el criterio de privacidad de [11] es una divergencia condicional, donde la variable condicionada es W y la condicionante, $\hat{X}$. En el caso de microagregación determinística, una divergencia condicional es un promedio intergrupo de divergencias intragrupo, donde los grupos comparten un valor común de $\hat{X}$. En virtud de la definición de divergencia de KL, resulta que las divergencias intragrupo son en sí mismas promedios de ratios logarítmicos de probabilidades.

Dicha medida de privacidad está estrechamente relacionada con la medida de t-cercanía de [8]. En términos de la formulación introducida en la Sección A y para el simple caso de distribuciones discretas y agrupaciones determinísticas, t-cercanía puede definirse como el máximo intergrupo de las divergencias intragrupo, en sí promedios de ratios logarítmicos de probabilidades. Un criterio relacionado aunque más conservador, llamado δ-revelación, se ha propuesto en [1], y mide la diferencia máxima entre las distribuciones a priori y a posteriori para cada grupo con un $\hat{X}$ común.

En pocas palabras, la medida de riesgo de privacidad de [11], comentada en la Section A, es un promedio intergrupo de promedios intragrupo (es decir, un promedio), t-cercanía es un máximo intergrupo de promedios intragrupo, y δ-revelación es un máximo intergrupo de máximos intragrupo (es decir, un máximo). Por lo tanto, dichas medidas van desde el modelado del caso promedio al peor de los casos.

V. CONCLUSIONES

En conclusión, hemos motivado la importancia de los métodos con perturbación para privacidad en la anonimización de microdatos y también ofuscación para SBL. En cuanto a los criterios de privacidad revisados, nos gustaría hacer hincapié en que a pesar de las deficiencias del k-anonimato y de sus mejoras, todavía es un criterio ampliamente popular

para CER, principalmente debido a su simplicidad y manejabilidad matemática. Sin embargo, dada la vulnerabilidad del k-anonimato y de sus mejoras frente a los ataques de similitud y sesgo, recientemente se han propuesto métricas de privacidad basadas en conceptos de teoría de la información.

Concordantemente, hemos examinado tres medidas de privacidad basadas en teoría de la información, a saber: el riesgo promedio de privacidad de [11], t-cercanía y δ-revelación. En primer lugar, es justo destacar que una optimización para el caso promedio puede no abordar los peores casos apropiadamente. En otras palabras, reconocemos que el criterio promediado de privacidad, como cualquier criterio basado en promedios, puede no adecuarse a todas las aplicaciones. Sin embargo, el precio a pagar de una optimización para el peor caso es, en general, un peor promedio, *ceteris paribus*. Por otra parte, el trabajo citado demuestra que las principales ventajas del criterio promediado de privacidad de [11] son su manejabilidad matemática, y el hecho de que generaliza el problema de compresión con pérdidas con una medida de distorsión originalmente propuesto por Shannon en 1959 [2].

En términos más generales, reconocemos que la formulación de cualquier problema de privacidad y utilidad se basa en a adecuación de los criterios de optimización, que a su vez dependen de la aplicación entre manos, del grado de utilidad de los datos que estamos dispuestos a sacrificar, y del modelo de adversario y de los mecanismos contra la privacidad contemplados. Ningún criterio de privacidad parece presentarse como ganador absoluto en la anonimización de bases de datos [1].

AGRADECIMIENTOS

Este trabajo ha sido financiado parcialmente por el Gobierno Español a través de los proyectos CONSOLIDER INGENIO 2010 CSD2007-00004 “ARES”, TSI2007-65393-C02-02 “ITACA” y TSI2007-65406-C03-01 “E-AEGIS", y por la Generalitat de Catalunya bajo la beca 2009 SGR 1362.

REFERENCIAS

[1] J. Brickell y V. Shmatikov, “The cost of privacy: Destruction of data-mining utility in anonymized data publishing,” en *Proc. ACM SIGKDD Int. Conf. Knowl. Disc., Data Min. (KDD)*, Las Vegas, EE.UU., Ago. 2008.

[2] T. M. Cover y J. A. Thomas, *Elements of Information Theory*, 2ª ed. Nueva York: Wiley, 2006.

[3] D. Defays y P. Nanopoulos, “Panels of enterprises and confidentiality: The small aggregates method,” en *Proc. Symp. Design, Anal. Longitudinal Surveys, Stat. Canada,* Ottawa, Canadá, 1993, págs. 195-204.

[4] C. Díaz, S. Seys, J. Claessens y B. Preneel, “Towards measuring anonymity,” en *Proc. Workshop Privacy Enhanc. Technol. (PET)*, ser. Lecture Notes Comput. Sci. (LNCS), Springer-Verlag, vol. 2482, Abril 2002.

[5] J. Domingo-Ferrer, “Microaggregation for database and location privacy,” en *Proc. Int. Workshop Next-Gen. Inform. Technol., Syst. (NGITS)*, ser. Lecture Notes Comput. Sci. (LNCS), Springer-Verlag, vol. 4032, Israel, Jul. 2006, págs. 106–116.

[6] J. Domingo-Ferrer, A. Martínez-Ballesté, J. M. Mateo-Sanz y F. Sebé, “Efficient multivariate data-oriented microaggregation,” *VLDB J.*, vol. 15, núm. 4, págs. 355-369, 2006.

[7] M. Duckham, K. Mason, J. Stell y M. Worboys, "A formal approach to imperfection in geographic information," *Comput., Environ., Urban Syst.*, vol. 25, núm. 1, págs. 89–103, 2001.

[8] N. Li, T. Li y S. Venkatasubramanian, "*t*-Closeness: Privacy beyond *k*-anonymity and *l*-diversity," en Proc. *IEEE Int. Conf. Data Eng. (ICDE)*, Estambul, Turquía, Abril 2007, págs. 106-115.

[9] D. V. Lindley, "On a measure of the information provided by an experiment," *Annals Math. Stat.*, vol. 27, núm. 4, págs. 986-1005, 1956.

[10] A. Machanavajjhala, J. Gehrke, D. Kiefer y M. Venkitasubramanian, "*l*-Diversity: Privacy beyond *k*-anonymity," en *Proc. IEEE Int. Conf. Data Eng. (ICDE)*, Atlanta, EE.UU., Abril 2006, p. 24.

[11] D. Rebollo-Monedero, J. Forné y J. Domingo-Ferrer, "From *t*-closeness-like privacy to postrandomization via information theory," *IEEE Trans. Knowl. Data Eng.*, 2009. En línea: http://doi.ieeecomputersociety.org/10.1109/TKDE.2009.190.

[12] D. Rebollo-Monedero, J. Forné y M. Soriano, "Private location-based information retrieval via *k*-anonymous clustering," en *Proc. CNIT Tyrrhenian Int. Workshop Digital Commun.*, Pula, Cerdeña, Italia, Sep. 2-4, 2009.

[13] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE Trans. Knowl. Data Eng.*, vol. 13, núm. 6, págs. 1010-1027, 2001.

[14] C. E. Shannon, "Communication theory of secrecy systems," *Bell Syst., Tech. J.*, 1949.

[15] T. M. Truta y B. Vinay, "Privacy protection: *p*-sensitive *k*-anonymity property," en *Proc. Int. Workshop Privacy Data Manage.* (PDM), Atlanta, EE.UU., 2006, p. 94.