Microagregación con Privacidad Diferencial: Ruido Adaptativo a Nivel de Clúster

Jordi Soria-Comas
y David Megías
Universitat Oberta de Catalunya,
Est. d'Inf, Mult i Telecom,
UOC-TECH Research Centre,
CYBERCAT,
Rambla del Poblenou, 154
08018, Barcelona, Spain
{jordi_sc,dmegias}@uoc.edu

Jordi Castellà-Roca
y Carles Anglés-Tafalla
Universitat Rovira i Virgili,
Dept. d'Eng Inf i Mat,
CYBERCAT,
Av. Països Catalans 26,
E-43007 Tarragona, Spain
{jordi.castella, carles.angles}@urv.cat

Guillermo Navarro-Arribas
y Julián Salas
Universitat Autonoma de Barcelona,
Dept. d'Eng. de la Inf i les Com,
CYBERCAT,
08193 Bellaterra, Catalonia, Spain
{guillermo.navarro, julian.salas}@uab.cat

***Resumen*—La microagregación es una técnica establecida para la anonimización de datos que reemplaza los registros originales por centroides de clústeres, preservando la utilidad estadística, pero sin ofrecer garantías teóricas de privacidad. Para mejorar las garantías, se puede combinar con Privacidad Diferencial (DP), añadiendo ruido a los centroides. Sin embargo, los enfoques actuales aplican una cantidad uniforme de ruido a todos los centroides. Esto tiene un impacto en la utilidad de los datos en regiones densas del dominio, donde los clústeres son inherentemente más compactos y menos susceptibles a la reidentificación. En este trabajo, proponemos una nueva metodología que integra la microagregación con la Privacidad Diferencial Individual (iDP), con el objetivo de optimizar el equilibrio entre privacidad y utilidad. El núcleo de nuestra propuesta reside en ajustar dinámicamente la cantidad de ruido introducido basándose en la sensibilidad local de cada clúster individual.**

***Palabras clave*—Microagregación, privacidad diferencial, utilidad**

I. Introducción

La microagregación es una técnica de anonimización de datos utilizada para proteger la privacidad de los individuos, a la vez que se mantiene la utilidad de los datos para el análisis. Consiste en agrupar puntos de datos individuales en pequeños clústeres y luego reemplazar los datos de cada individuo dentro de ese grupo por un valor central (conocido como centroide) que es representativo de todo el clúster. Es habitual calcular el centroide como la media de los valores del clúster. El objetivo es suprimir la información sensible que podría identificar a los individuos, a la vez mantener la utilidad estadística de los datos.

Aunque la microagregación se utiliza ampliamente para anonimizar datos, no proporciona garantías de privacidad teóricas como da la Privacidad Diferencial (DP). En ciertas situaciones, los datos microaggregados pueden revelar información personal:

- Revelaciones por tamaño de grupo insuficiente o falta de diversidad estadística. Cuando los grupos son demasiado pequeños o tienen poca variabilidad interna (por ejemplo, todos los individuos tienen valores muy similares para los atributos agregados), la media u otros valores agregados no introducen suficiente ambigüedad o ruido para proteger la privacidad individual.
- Revelaciones por conocimiento externo. Si un atacante tiene acceso a conjuntos de datos externos con información sobre los individuos (como características demográficas o registros públicos), podría vincular esos datos con centroides concretos. Al hacerlo, el atacante reduce efectivamente el tamaño del grupo.
- Revelaciones debidas a valores atípicos (outliers). Si hay valores atípicos (por ejemplo, un individuo con un valor significativamente diferente a los demás individuos del clúster, como un ingreso muy alto en un grupo), el centroide se verá fuertemente influenciado por ese valor, comprometiendo la anonimidad.

Una forma efectiva de abordar los problemas de privacidad inherentes a la microagregación es combinarla con la privacidad diferencial [1]. El objetivo es mejorar la protección de la privacidad sin que esto comporte una gran perdida en la utilidad de los datos. En términos de privacidad, al añadir ruido DP a los valores agregados después de la microagregación, reducimos significativamente el riesgo de revelar los datos de cualquier individuo. Incluso si un atacante solo desconoce un registro en un grupo, el ruido añadido por DP asegura que el valor del registro restante no pueda inferirse del valor del centroide. En términos de utilidad, dado que la microagregación ya ofusca a los individuos al transformar los datos individuales en un valor representativo del grupo, la cantidad de ruido necesaria para la privacidad diferencial es menor que si añadiéramos ruido DP a nivel de datos individuales.

La combinación de microagregación y DP ha sido analizada en varios trabajos [6], [7], [4]. Sin embargo, estos enfoques se enfrentan a un desafío significativo que impacta negativamente en la utilidad de los datos: aplican la misma cantidad de ruido a todos los centroides, independientemente de las características del clúster. La razón es que la protección debe garantizarse con independencia del conjunto de datos.

Para minimizar el impacto del ruido, sería conveniente que el ruido se ajustara en función de la extensión del clúster, de modo que las regiones con mayor densidad de datos (valores más comunes) reciban menos ruido, mientras que las regiones más dispersas (valores poco frecuentes), que son más propensas a la reidentificación, reciban más ruido. Esto ayudaría a equilibrar la protección de la privacidad con una

mejor utilidad de los datos.

Para abordar este problema, recurrimos a la Privacidad Diferencial Individual (iDP), que, al igual que la DP, tiene como objetivo hacer que la contribución de cualquier individuo sea indistinguible (la probabilidad de obtener la misma respuesta a una consulta debe ser similar independientemente de la presencia o ausencia de ese individuo en el conjunto de datos). Sin embargo, a diferencia de la DP, la iDP solo busca ofrecer protección para el conjunto de datos real (no sobre cualquier conjunto de datos posible). Para ello basta con exigir indistinguibilidad entre el conjunto de datos real y aquellos que difieren de él en un único registro. La idea es que a un individuo no le importa la protección que obtendría si se consultara un conjunto de datos arbitrario, solo le importa la protección sobre el conjunto de datos real.

El resto del artículo se organiza de la siguiente manera. La Sección II recuerda los conceptos centrales de Privacidad Diferencial (DP) y microagregación involucrados en nuestra propuesta. La Sección III describe el uso de la Privacidad Diferencial Individual (iDP) para enmascarar los centroides obtenidos mediante microagregación. La Sección IV presenta los resultados empíricos en una variedad de escenarios de datos para demostrar las mejoras de nuestra propuesta e identificar los contextos en los que se obtiene el mayor beneficio. La Sección final ofrece las conclusiones y enumera las líneas de investigación futura.

II. Marco Teórico

II-A. Microagregación Univariada

La microagregación es una técnica utilizada en el contexto de la privacidad de datos para proteger la información sensible a la vez que se preservan sus propiedades estadísticas. Consta de dos pasos principales:

- Agrupación (*Clustering*) de los datos: los datos se dividen en grupos o clústeres no superpuestos, cada uno de los cuales contiene k o más registros.
- Agregación de valores dentro de cada grupo: cada clúster de registros se reemplaza por un valor representativo, el centroide (por ejemplo, la media).

Para minimizar la pérdida de información en la microagregación, los clústeres deben ser lo más homogéneos posible, lo que significa que los valores dentro de cada grupo deben ser lo más similares entre sí.

En este artículo, nos centramos en la microagregación univariada de datos numéricos. Los siguientes pasos resumen el proceso de microagregación en dichos conjuntos de datos:

1. **Entrada:**
   - Un conjunto de datos numéricos $D = \{x_1, x_2, \ldots, x_n\}$
   - Un tamaño de grupo especificado k, donde k es el número de puntos de datos en cada grupo.
2. **Clasificación de los datos:**
   - Se ordena el conjunto de datos D en orden no decreciente de los valores.
   - El conjunto de datos ordenado pasa a ser $D_{\text{sorted}} = \text{sort}(D)$.
3. **División de los datos en grupos:**
   - Se divide el conjunto de datos ordenado D_{sorted} en grupos no superpuestos de tamaño k.
   - Si el tamaño del conjunto de datos n no es divisible por k, los registros restantes se distribuyen en los clústeres.
4. **Cálculo del valor central para cada grupo:**
   - Para cada grupo, se calcula la media de los valores dentro del grupo:
$$\text{media}_i = \frac{\sum_{j=1}^{k} \text{grupo}_i[j]}{k}.$$
5. **Sustitución de los valores de cada grupo por el valor central:**
   - Todos los valores dentro de cada grupo se reemplazan por el valor central calculado.

II-B. Privacidad Diferencial

La DP asume la presencia de un tercero de confianza que: (i) posee el conjunto de datos, (ii) recibe las consultas enviadas por los usuarios de los datos, y (iii) responde a ellas de una manera que garantiza la privacidad. La noción estándar de DP se formaliza en términos de indistinguibilidad de la respuesta a las consultas entre conjuntos de datos que difieren en un registro: la probabilidad de obtener una misma respuesta debe ser similar en ambos casos.

Definición 1 (ε-Privacidad Diferencial)**.** Una función aleatorizada κ proporciona ε-Privacidad Diferencial (o ε-DP) si, para todos los conjuntos de datos D_1 y D_2 que difieren en un registro (también conocidos como conjuntos de datos vecinos), y para todo $S \subset \text{Rango}(\kappa)$, se tiene

$$\Pr(\kappa(D_1) \in S) \leq \exp(\varepsilon) \cdot \Pr(\kappa(D_2) \in S).$$

Dada una consulta f, el objetivo de la DP es encontrar una κ_f que satisfaga los requisitos de privacidad establecidos en la Definición 1 y se aproxime a f lo más posible. El valor de $\kappa_f(D)$ se devuelve entonces como el resultado de la consulta, en lugar del valor real $f(D)$. El valor de ε nos indica el nivel de privacidad perdida, cuanto más grande sea menos garantías de privacidad.

Una propiedad interesante de la privacidad diferencial es que podemos hacer un seguimiento de la privacidad perdida debido a la realización de varias consultas; esto se conoce como composición. En este trabajo solo tratamos con composición secuencial [2].

Teorema 2 (Composición secuencial)**.** *Sea κ_1 una función aleatorizada que proporciona ε_1-DP y κ_2 una función aleatorizada que proporciona ε_2-DP. Entonces, cualquier función determinista de (κ_1, κ_2) proporciona $(\varepsilon_1 + \varepsilon_2)$-DP.*

Una forma común de lograr DP para consultas numéricas es añadir un ruido cuya variabilidad se ajusta al cambio máximo entre conjuntos de datos vecinos (también conocida como sensibilidad global).

Definición 3 (Sensibilidad global)**.** La sensibilidad global de una función $f : \mathcal{D} \to \mathbb{R}^k$ es

$$\Delta f = \max_{D_1, D_2 \text{ vecinos}} \|f(D_1) - f(D_2)\|_1 .$$

Un ruido distribuido por Laplace es una alternativa común para lograr DP.

Proposición 4. *Sea f una función de consulta con valores en $\mathbb{R}^k$. El mecanismo $\kappa_f(D) = f(D) + (N_1, \ldots, N_k)$, donde N_i son ruidos aleatorios independientes e idénticamente distribuidos extraídos de una distribución* $\mathrm{Laplace}(0, \varepsilon/\Delta f)$*, proporciona ε-DP.*

II-C. Microagregación con DP

Sea f la función que evaluada en un conjunto de datos, retorna los clústeres de la microagregación. Para obtener DP, tenemos que enmascararlos añadiendo ruido.

En el caso de microaggregación univariada, el cambio más grande en los centroides se da cuando un registro que estava en un extremo del dominio pasa al otro extremo [4]. En este caso, si ΔD es la dimensión del dominio, la suma de los cambios en los centroides es $\Delta D/k$, donde el cociente proviene del hecho que al hacer la media el cambio en cada centroide queda dividido por k. Por tanto, la función $\kappa_f(D) = (c_1(D), \ldots, c_{n/k}(D)) + (L_1, \ldots, L_{n/k})$, donde $L_i \sim L(0, {}^{\Delta D}/_{k\varepsilon})$ independientes, da ε-DP.

II-D. Privacidad Diferencial Individual

Antes de formalizar la DP, en el artículo [1] se propone la siguiente intuición:

> *Cualquier revelación dada será, dentro de un pequeño factor multiplicativo, igual de probable participe o no el individuo en la base de datos. Como consecuencia, hay un riesgo marginalmente mayor para el individuo que participa y solo una ganancia marginal para el individuo que oculta o tergiversa sus datos.*

Sin embargo, la formalización de la DP (véase la Definición 1) es más rigurosa de lo que requiere la intuición anterior. De acuerdo con la intuición previa, es suficiente decirle al individuo que las respuestas a las consultas basadas en el conjunto de datos no permitirán a nadie aprender algo que no fuera ya aprendible sin su presencia. En otras palabras, si D es el conjunto de datos recopilado, las garantías de privacidad objetivo pueden lograrse simplemente exigiendo la indistinguibilidad de las respuestas entre D y sus conjuntos de datos vecinos. No es necesario pedir la indistinguibilidad entre cualquier par de conjuntos de datos vecinos.

Definición 5 (ε-Privacidad Diferencial Individual [5])**.** Dado un conjunto de datos D, un mecanismo de respuesta $\kappa(\cdot)$ satisface ε-Privacidad Diferencial Individual (o ε-iDP) si, para cualquier conjunto de datos D' que sea vecino de D, y cualquier $S \subset \mathrm{Rango}(\kappa)$ se tiene

$$\exp(-\varepsilon) \cdot \Pr(\kappa(D') \in S) \leq \Pr(\kappa(D) \in S)$$
$$\Pr(\kappa(D) \in S) \leq \exp(\varepsilon) \cdot \Pr(\kappa(D') \in S).$$

Como el papel de los conjuntos de datos D y D' no es intercambiable, necesitamos solicitar ambas desigualdades.

Los teoremas de composición para la DP descritos en la Sección II-B también son trivialmente satisfechos por la privacidad diferencial individual. El mecanismo de adición de ruido Laplace descrito en la Sección II-B también conduce a ε-iDP. Sin embargo, podemos ajustar el ruido a la sensibilidad local, en lugar de a la sensibilidad global.

Definición 6 (Sensibilidad local)**.** Sea f una función que se evalúa en conjuntos de datos y devuelve valores en $\mathbb{R}^k$. La sensibilidad local de f en D es

$$\mathrm{LS}_f(D) = \max_{D', D \text{ vecinos}} \|f(D') - f(D)\|_1 .$$

Proposición 7. *Sea f una función de consulta que toma valores en $\mathbb{R}^k$. El mecanismo $\kappa(x) = f(x) + (N_1, \ldots, N_k)$, donde N_i son ruidos aleatorios independientes e idénticamente distribuidos* $\mathrm{Laplace}(0, LS_f(D)/\varepsilon)$*, proporciona ε-iDP.*

III. Microagregación con iDP

Al microagregar un conjunto de datos en clústeres de tamaño k, la densidad de los registros en cada región del dominio de datos tiene un impacto significativo en la medida de los clústeres. Los clústeres formados en áreas más densas del dominio contienen registros que están más próximos entre sí y reemplazarlos por el centroide provoca una pérdida de información menor. Por el contrario, los clústeres formados en áreas más dispersas contienen registros que están distantes entre sí y reemplazarlos por el centroide conduce a una mayor pérdida de información. Esto es coherente con los enfoques tradicionales de privacidad de datos, donde la cantidad de enmascaramiento que se realiza a los registros depende de cuán único sea un registro.

Cuando aplicamos DP sobre los centroides, como explicamos en la Sección II-C, cada uno de ellos se enmascara con la misma cantidad de ruido: un ruido laplaciano con escala $\Delta D/k\varepsilon$. Así, el impacto es más notable en los clústeres en regiones más densas. En las regiones dispersas, el centroide era una aproximación burda y, por lo tanto, proporcionalmente el nuevo ruido es menor.

El objetivo es ajustar la cantidad de ruido añadido a la densidad de la región en la que se encuentra cada clúster. Al hacerlo, podremos ofrecer resultados más precisos en regiones más densas (que probablemente representan un mayor porcentaje del total de registros) a expensas de introducir más ruido en regiones más dispersas. En DP, la sensibilidad global de cada centroide es $\Delta D/k$, ya que necesitamos considerar cualquier par posible de conjuntos de datos vecinos, incluso aquellos en los que uno de los registros del clúster cambia de un extremo del dominio de datos al otro. Por el contrario, si empleamos iDP como noción de privacidad, es posible calcular una sensibilidad mucho más precisa para cada uno de los centroides.

Antes de calcular la sensibilidad de cada centroide, fijemos algunas notaciones. Sea $D = \{x_1, x_2, \ldots, x_n\}$ el conjunto de datos real. Sin pérdida de generalidad, asumimos que $x_1 \leq x_2 \leq \ldots \leq x_n$. Sea $D' = \{x'_1, x'_2, \ldots, x'_n\}$ un conjunto de datos vecino. Al igual que antes, asumimos que los registros en D' están ordenados de forma ascendente. Sea k el tamaño del clúster y, por simplicidad, asumiremos que n es múltiplo de k. Es decir, que podemos generar n/k clústeres de tamaño k sin que quede ningún registro.

Como se describe en la Sección II-A, los clústeres de microagregación en D y D' contienen cada uno k registros contiguos. Sea $C_D^i = \{x_{ik+1}, \ldots, x_{(i+1)k}\}$ el i-ésimo clúster sobre el conjunto de datos D. De manera similar, $C_{D'}^i$ es el i-ésimo clúster sobre el conjunto de datos D'. Como D y D' difieren como máximo en un registro, los clústeres C_D^i y $C_{D'}^i$ también difieren como máximo en un registro. En particular,

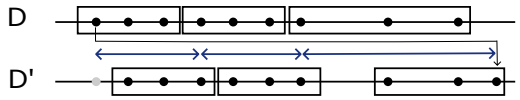


Figura 1. Ilustración de cómo el cambio de un único registro afecta a los clústeres en conjuntos de datos ordenados D y D'

si los conjuntos de datos D y D' difieren en los valores de los registros x_r y x'_s, entonces los clústeres que contienen los registros entre x_r y x'_s difieren en un registro, mientras que todos los demás clústeres permanecen sin modificar (véase la Figura 1).

Así, la sensibilidad local de un centroide es el máximo cambio en el centroide debido al cambio de uno de los registros del clúster. El cambio máximo ocurre cuando:

- Se elimina el registro más pequeño del clúster y se añade el registro siguiente al clúster.
- Se elimina el registro más grande del clúster y se añade el registro anterior al clúster.

Cuando se trata del primer y el último clúster, el nuevo registro puede ir hasta el límite inferior o superior del dominio de datos (si el dominio está acotado). Para tener esto en cuenta, añadimos dos registros adicionales:

- x_0, que se establece en el límite inferior del dominio de datos.
- x_{n+1}, que se establece en el límite superior del dominio de datos.

Estos registros no participan en el cálculo del centroide (ya que no son registros propiamente dichos), pero permiten evitar tener que lidiar con diferentes casos al calcular la sensibilidad local.

Por lo tanto, la sensibilidad del centroide c_i es

$$\Delta c_i = \text{máx}\left\{\frac{x_{(i+1)k+1} - x_{ik+1}}{k}, \frac{x_{(i+1)k} - x_{ik}}{k}\right\}.$$

Por composición secuencial, si consultamos cada centroide y obtenemos una respuesta ε_i-iDP, el conjunto de centroides es $\sum \varepsilon_i$-iDP. Para obtener la misma protección para cada centroide, hacemos que todos los ε_i sean iguales. Tomamos $\varepsilon_i = \varepsilon/(n/k)$. Por lo tanto, enmascaramos cada centroide añadiendo un ruido Laplace:

$$c_i + \text{Laplace}\left(0, \Delta c_i \times (n/k)/\varepsilon\right).$$

Como el ruido que se introduce es proporcional a la longitud del clúster, estamos añadiendo, efectivamente, menos ruido a los clústeres en áreas densamente pobladas, mientras que aumentamos el ruido añadido a las áreas escasamente pobladas. Para ilustrar esto, analicemos el caso de datos distribuidos uniformemente en todo el dominio de datos. En tal caso, es probable que cada clúster abarque la misma proporción del dominio de datos y se espera que la sensibilidad del centroide sea:

$$\Delta c_i = {}^{\Delta D/(n/k)}/_{k}.$$

Para esta sensibilidad, el ruido que tenemos que añadir es $\text{Laplace}(0, \Delta D/(k\varepsilon))$, el mismo ruido que añadimos al usar DP simple. A partir de esta base, es fácil ver que si la distribución no es uniforme, el ruido añadido en áreas densamente pobladas se vuelve menor (ya que el Δc_i en esas áreas se reduce), y el ruido en áreas escasamente pobladas se vuelve mayor (ya que el Δc_i en esas áreas aumenta).

IV. Resultados experimentales

Esta sección presenta la evaluación empírica sobre conjuntos de datos sintéticos. Los conjuntos de datos generados contienen 1.000 registros con un único atributo numérico, cuyo valor sigue una distribución $N(0,1)$ truncada en el intervalo $[-5,5]$.

Para evaluar la pérdida de información del conjunto de datos (es decir, la diferencia entre los datos originales y los enmascarados), utilizamos el Error Absoluto Medio (MAE) y el Error Cuadrático Medio (MSE), que son medidas estándar de pérdida de información en microagregación. Estas métricas contabilizan la distancia euclidiana promedio y la distancia euclidiana al cuadrado entre los registros originales y los enmascarados:

$$\text{MAE} = \frac{1}{|D|}\sum_{r\in D}\|r - r'\|_2,$$

$$\text{MSE} = \frac{1}{|D|}\sum_{r\in D}\|r - r'\|_2^2,$$

donde r son los registros del conjunto de datos, y r' son las versiones enmascaradas de cada r.

Si bien las medidas globales de pérdida de información nos dan una visión general, no son suficientes para evaluar nuestra propuesta, ya que esperamos que el ruido no esté distribuido de manera uniforme. Por esta razón, también evaluaremos el MAE y el MSE sobre cada uno de los clústeres.

Para la privacidad diferencial (DP), utilizamos $\varepsilon = 1$, que es un valor común. Para la microagregación, hemos utilizado tres tamaños de clústeres diferentes $k \in \{10, 50, 100\}$, lo que abarca un amplio rango de tamaños de clústeres.

Para dar contexto a nuestra propuesta, la comparamos con la microagregación simple (que muestra el límite inferior de la pérdida de información) y la microagregación combinada con DP, *microagregación+DP*. Para cada método y parametrización, se han realizado 10 rondas y se proporciona el promedio de todas ellas. Para hacer la comparación más consistente, todos los valores detallados se calcularon sobre los mismos 10 conjuntos de datos.

En cuanto a las medidas globales de MAE y MSE, no observamos diferencias significativas entre *microagregación+DP* y *microagregación+iDP*. En algunos casos, uno es ligeramente mejor que el otro, y viceversa. Por supuesto, ambos muestran una pérdida de información significativamente mayor que la microagregación simple, lo que proviene de las fuertes garantías de privacidad que incorporan. En la Tabla I se muestran los valores de MAE i MSE globales para diferentes niveles de mmicroagregación.

La información que obtenemos al analizar la pérdida de información a nivel de clúster confirma la redistribución del ruido entre clústers. En la Figura 2, observamos que, independientemente del tamaño del clúster, *microagregación+iDP* produce mejores resultados en una gran parte de los clústeres. Aproximadamente el 95 % de los clústeres presentan mejores resultados con *microagregación+iDP*.

Para analizar la distribución del ruido debido a iDP entre las regiones densas y dispersas, se representa el MSE de

Tabla I
MAE I MSE PARA LOS MÉTODOS EVALUADOS Y DIFERENTES NIVELES DE MICROAGREGACIÓN (K).

	$k = 10$		$k = 50$		$k = 100$	
	MAE	MSE	MAE	MSE	MAE	MSE
Microagregación	0,0174	0,00279	0,0688	0,0152	0,126	0,041
Microagregación + DP	1,0	1,9	0,225	0,096	0,179	0,0635
Microagregación + iDP	0,562	1,25	0,242	0,0423	0,167	0,0881

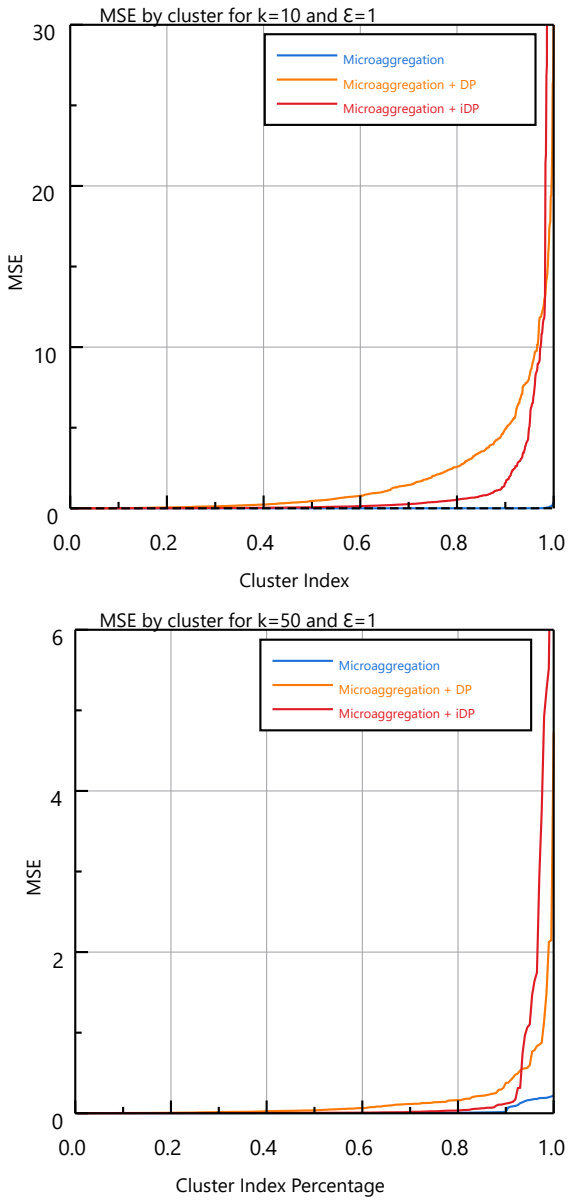


Figura 2. Distribución de MSE por clúster ordenada de menor a mayor MSE, para $k = 10$ y $k = 50$.

los clústeres sin reordenarlos. De esta manera, los valores al principio y al final del eje x corresponden a los clústeres en las regiones cercanas a –5 y 5 (que son dispersas), mientras que los valores en el centro del eje x corresponden a los clústeres cercanos a 0 (que son más densos). Hay que tener en cuenta que el eje x representa el índice del clúster y no la distribución real de los valores en el rango $[-5, 5]$ (que sigue una distribución $N(0, 1)$). En la Figura 3, se representa el $\log_{10}$ del MSE de los clústeres para los tres métodos analizados. Utilizamos escala logarítmica para separar los puntos que corresponden a diferentes métodos de forma más clara. Esto nos permite observar mejor las tendencias en el MSE. Los puntos correspondientes a la microagregación tienen un MSE más bajo y es más pequeño en el centro (región densa) que en los extremos (regiones dispersas). Por el contrario, la *microagregación+DP* muestra un MSE que es bastante uniforme en todo el dominio de los datos. La razón es que el ruido debido a DP es uniforme en todos los clústeres y domina sobre el ruido debido a la microagregación. Finalmente, los puntos relacionados con la *microagregación+iDP*, aunque muestran valores de MSE más grandes, podemos observar que siguen una tendencia similar a la de la microagregación (con valores más grandes en los extremos y más pequeños en el centro).

V. CONCLUSIONES

Si bien la microagregación es una técnica comúnmente utilizada para el control de divulgación estadística, carece de las garantías de privacidad que ofrece la DP. Por ejemplo, al vincular los registros de datos con clústeres particulares, los intrusos pueden reducir el nivel de protección. Para evitar esto, se puede aplicar DP a los centroides de los clústeres. Dado que cada centroide depende de un grupo de k o más registros, la sensibilidad se reduce por un factor k, lo que reduce la cantidad de ruido que es necesario agregar.

Sin embargo, en la DP estándar, cada centroide se enmascara con la misma cantidad de ruido. Agregar la misma cantidad de ruido a cada clúster tiene un impacto mayor en los clústeres de las regiones más densas, ya que son más pequeños. Para los clústeres en áreas más dispersas, como el centroide ya es una estimación aproximada, el ruido adicional es menos notorio. Este tratamiento uniforme de los clústeres tiene un profundo impacto en la pérdida de información.

Para resolver este problema, proponemos el uso de iDP, que nos permite adaptar la cantidad de ruido a cada clúster. A los clústeres en áreas más densas se les agrega menos ruido, mientras que a los clústeres en áreas más dispersas se les agrega más ruido. Esta adaptación nos permite reducir significativamente la pérdida de información en las áreas más densas, a costa de agregar significativamente más ruido en las áreas más dispersas. Sin embargo, dado que la precisión de los centroides en las áreas dispersas ya era baja, el impacto de agregar más ruido es menos notorio.

Hemos evaluado nuestra propuesta en un conjunto de datos sintético generado a partir de una distribución normal. Los resultados muestran que, aunque las medidas globales de pérdida de información tienen valores similares entre DP e iDP, al evaluar a nivel de clúster iDP preserva mejor la utilidad de los clústers en regiones densas.

En este trabajo, la microagregación se usa únicamente para reducir la cantidad de ruido introducido por la DP.

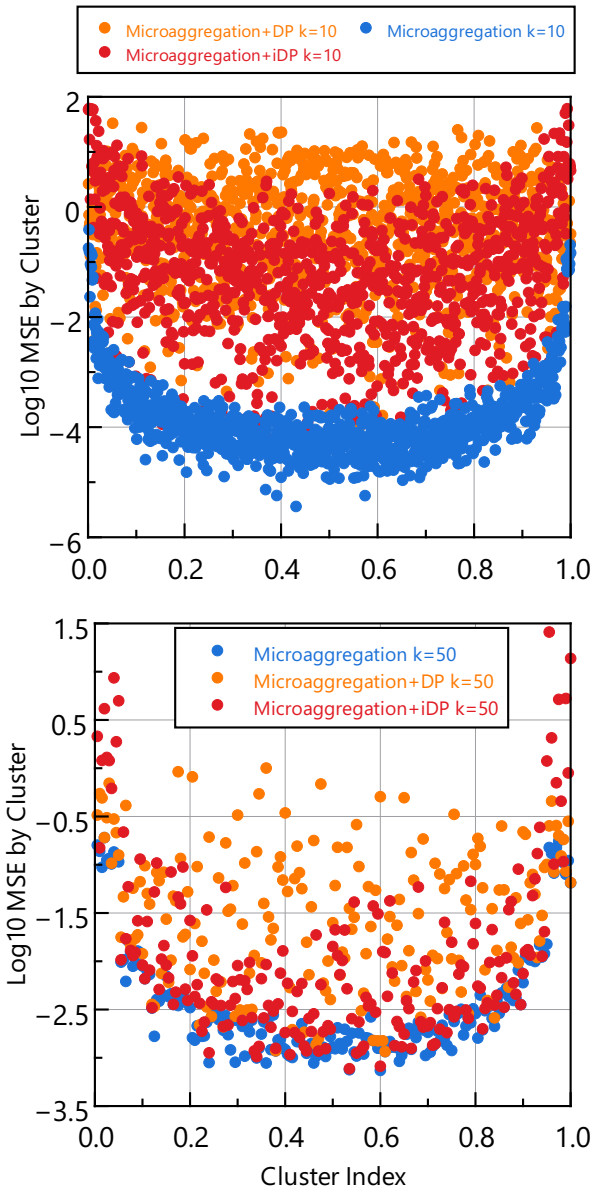


Figura 3. Log10 del MSE de los clústeres sin reordenar para $k = 10$ y $k = 50$.

Como trabajo futuro queremos determinar que medida tiene que tener cada clúster para que se minimice la perdida de información total que resulta de agregar la procedente del ruido de DP a la de la microagregación, teniendo en cuenta que el nivel óptimo de microagregación puede variar en función de la densidad del área.

Agradecimientos

Este trabajo ha sido apoyado por los proyectos estratégicos en ciberseguridad HERMES y DANGER (C062-23), y la cátedra en ciberseguridad ARTEMISA (C057/23) financiados por INCIBE a través de la Unión Europea – NextGenerationEU y el Plan de Recuperación, Transformación y Resiliencia; por los proyectos PROVTOPIA (PID2023-150098OB-I00), SECURING/CYBER (PID2021-125962OB-C31), SECURING/DATA (PID2021-125962OB-C32), SECURING/NET (PID2021-125962OB-C33), SAFE/DATA (PID2024-156914OB-C42), SAFE/BLOCKCHAIN (PID2024-156914OB-C43), SAFE/CYBER (PID2024-156914OB-C41) del Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, y *European Regional Development Fund* (ERDF). Los proyectos SGR2021-00643, 2021SGR 00115 de la Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) de Catalunya.

Referencias

[1] C. Dwork. Differential privacy. *Automata, languages and programming*, page 1–12, 2006.

[2] Frank D. McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data (SIGMOD '09)*. Association for Computing Machinery, New York, NY, USA, 19–30.

[3] J. Domingo-Ferrer and J. M. Mateo-Sanz, Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189–201, Feb 2002.

[4] J. Soria-Comas, D. Sánchez, J. Domingo-Ferrer, S. Martínez, L. Del Vasto-Terrientes. Conciliating privacy and utility in data releases via individual differential privacy and microaggregation. *Transactions on Data Privacy*, 18:29–50, Jan 2025.

[5] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and D. Megias. Individual differential privacy: A utility-preserving formulation of differential privacy guarantees. *IEEE Transactions on Information Forensics and Security*, 12:1418–1429, Jun 2017.

[6] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez and S. Martínez. Enhancing data utility in differential privacy via microaggregation-based k-anonymity. *The VLDB Journal*, 23(5):771–794, Sep 2014.

[7] X. Ye, Y. Zhu, M. Zhang and H. Deng, Differential Privacy Data Release Scheme Using Microaggregation With Conditional Feature Selection. *IEEE Internet of Things Journal,* 10(20):18302–18314, Oct 2023,