# Modelo de Protección

## en contra de la Inferencia Lógica en las Bases de Datos Estadísticos


## Resumen

A pesar de la implementación de gran variedad de los mecanismos de seguridad – mecanismos de control de acceso, autenticación, control de flujos, etc.– ninguno de estos sistemas de seguridad puede tratar eficazmente con el problema de la inferencia lógica o estadística – deducción de los datos secretos (contenido de algunos registros particulares) por medio de aplicación de los métodos matemáticos a los datos estadísticos disponibles.

En este articulo se considera un sistema de protección de base de datos estadísticos en contra de la inferencia lógica, que se basa en el censor lógico. Al utilizar la clasificación apropiada de las fórmulas características de las consultas, se demuestra el teorema que el censor lógico propuesto tiene las propiedades de necesidad y suficiencia para detectar cualquier intento de la inferencia lógica.


*Palabras y frases clave*: seguridad de las bases de datos, seguridad de datos, bases de datos estadísticos, tracker.

## Introducción

Actualmente las bases de datos estadísticos pueden considerarse como las fuentes únicas, de los datos objetivos sobre el estado actual y tendencias de cambios de los diferentes índices del desarrollo de la sociedad – economía, medicina, crimen, ecología, recursos humanos, etc.– La importancia de estas fuentes de datos recientemente es tan grande, que sólo la existencia de estas bases de datos determinan la existencia de gran variedad de las organizaciones gubernamentales y empresariales. Es casi obvio que la eficacia de la utilización de los datos almacenados en estas bases depende directamente de su disponibilidad para la comunidad de investigadores. En realidad, en la mayoría de los casos prácticos, el acceso a las bases de datos estadísticos más importantes (economía, censo, medicina, crimen, etc.) es restringido a los grupos limitados de investigadores, y las razones principales no sólo son subjetivas (tales como el egoísmo de propietario) sino también por las razones objetivas: la seguridad.

A pesar de la implementación de gran variedad de mecanismos de seguridad –mecanismos de control de acceso, autenticación, control de flujos, etc.– ninguno de estos sistemas de seguridad puede tratar eficazmente con el problema de la inferencia lógica o estadística. Deducción de los datos confidenciales (un caso común, el contenido de algunos registros particulares en una relación) por medio de aplicación de los métodos matemáticos a los datos estadísticos disponibles.

Es natural aceptar que el espacio de interés para especialistas, los cuales utilizan las bases de datos estadísticos, incluye toda la variedad de las estadísticas razonables, hasta las estadísticas de los grupos pequeños. También es razonable aceptar que estos especialistas (según definición) tienen experiencia y conocimientos completos sobre la estructura de los datos – cantidad de los atributos, conjuntos de valores de los atributos, restricciones para los conjuntos característicos permisibles para una consulta estadística, etc.– Efectivamente, en caso contrario, la base estadística es una caja negra para el usuario y la eficacia de su utilización es nula.

Por otro lado, en varias publicaciones, por ejemplo [2], [3], se muestra la dependencia crítica entre la posibilidad de guardar secretos en una base estadística y el nivel de conocimientos de sus usuarios. Aún más, bajo las preposiciones razonables sobre el conocimiento y experiencia de los usuarios la tarea de deducción de cualquier registro particular o bien, cualquier conjunto pequeño de los registros particulares, por

medio de un método conocido de manipulación con las consultas completamente legales es una tarea bastante fácil para un usuario curioso.

Los medios de protección en contra de la inferencia lógica tales como, por ejemplo, valores grandes k y N-k de los límites de los conjuntos característicos ([2]), o bien partición horizontal de la tabla relacional de la base estadística ( [6] ), son bastante eficaces sólo para usuarios que desconocen la estructura de datos de la base estadística.

Así pues, existe una contradicción grave entre la seguridad de las bases de datos estadísticos y la eficacia de utilización de estas fuentes de datos importantes para toda la comunidad.

## 2. Modelo de la base de datos estadística

Una base de datos estadística (aquí se consideran solo las bases de datos relacionales) incluye una tabla relacional W con N registros. Cada registro consiste de **m** atributos $v_1$, $v_2$,... $v_m$ (columnas) donde el atributo $v_j$: $\{v_{1j}, v_{2j},..., v_{rj}\}$ puede tomar $\mathbf{r_i}$ valores de un conjunto de valores $V_j$. De tal manera, la tabla relacional W, que se expone en una base de datos estadística, en el caso general, puede considerarse como un subconjunto del producto cartesiano:

$W \subset V_1 \times V_2 \times ... \times V_m$. Mas rigurosamente, W no puede coincidir con el conjunto del producto cartesiano porque conceptualmente una base de datos representa un modelo abstracto de una parte o un objeto particular del mundo real, y este objeto en el mundo real se identifica entre otros por las relaciones ciertas entre los atributos pertinentes. Por lo tanto, principalmente solo un subconjunto característico de las combinaciones de valores de los atributos puede pertenecer a la tabla relacional de una base de datos.

Si suponemos, por ejemplo, que las cardinalidades de todos conjuntos de los valores de atributos $V_1$, $V_2$,...$V_m$ son iguales, $|V_1| = |V_2| = ... |V_m| = r$, entonces la cantidad total de las tuplas distinguidas es $\mathbf{m^r} > N$. Por otra parte, sabemos exactamente que la base de datos original incluye N tuplas diferentes que son tuplas *legítimas* para el modelo del mundo real subyacente, entonces efectivamente sólo las igualdades $m^{r^*} = N$, $r^* < r$, o $m^{*r} = N$, $m^* < m$ son *legítimas* (dependiente de la interpretación de las relaciones entre atributos de un objeto modelado por la base estadística dada).

Por ejemplo, sea m=10, r= 10, N=100 000, entonces la cantidad de los registros distinguidos es igual a $10^{10}$, pero para la tabla relacional dada, son legítimos sólo igualdades $m^{r^*}=N$ o $m^{*r}=N$. En el primer caso el valor del exponente $r^*=5$ puede tratarse como la cantidad promedio de los valores de atributos *independientes necesarios* para la descripción del objeto modelado, así pues r - r* valores de atributos restantes son excesivos (dependientes).

En segundo caso podemos tratar $m^* \sim 3$ como el valor de atributos *independientes necesarios* para la descripción del objeto modelado, así pues m–m* atributos restantes son atributos excesivos (dependientes).

Hay dos diferencias principales entre la base de datos estadística y la base de datos convencional:

1. La tabla de datos relacional W que se expone en la base de datos estadística efectivamente es solo un fragmento de la partición vertical de la relación original R. Por las razones de seguridad toda la información privada (los atributos correspondientes) la cual pudiera identificar únicamente los registros particulares, (dirección postal, clave primaria, número de teléfono, número de seguro social, etc.) se elimina de la relación original y se guarda aparte. Por lo tanto, aun cuando el conjunto de todos los atributos en la relación original R hubieran sido una superclave, (esta propiedad se define exclusivamente por la naturaleza de los datos), un subconjunto de los atributos no es superclave según definición [5]. De tal manera, la tabla de datos W en la base de datos estadística no satisface a las restricciones de la clave primaria o superclave, y por lo tanto puede contener la cantidad arbitraria de los duplicados, es decir, la base de datos estadística W formalmente no es el conjunto de registros sino un multiconjunto.

Formalmente este significa que el axioma de elección:

$f: \{V_1 \times V_2 \times ... \times V_m \rightarrow R\} \sim (f(V1), f(V2), ..., f(V_m)) = v^*_1, v^*_2, ... v^*_m$ que siempre existe para una familia de conjuntos $V_1,...,V_m$ no vacíos y define únicamente un elemento particular $v^*_1, v^*_2, ... v^*_m$, es cierta para una relación R de la base original, es decir, siempre existe la función de elección f que regresa un registro único. Pero el axioma de selección:

$f': \{V_1 \times V_2 \times ... \times V_{m-h} \rightarrow W\}$ para la tabla relacional W de la base estadística correspondiente, no es cier-

ta, es decir, *en el caso general* no existe la función de elección que identifica únicamente un registro particular en W, porque W no es conjunto, sino multiconjunto.

En el marco del ejemplo anterior esto significa que para la tabla de datos expuesta en la base estadística tenemos: $N=\tilde{N}(m\text{-}h)^{r*}$ donde **h** denota la cantidad de los atributos eliminados de la relación R original y $\tilde{N}=(m/(m\text{-}h))^{r*}$ puede tratarse como la cantidad promedio de repetición de cada registro en una base estadística. Si seguimos al ejemplo anterior y suponemos que h= 2, entonces cada registro en W debe de repetirse en promedio 3 veces.

2. En contraste con la base de datos convencional, las consultas a la base de datos estadística no pueden regresar los conjuntos característicos de registros –la base de datos estadística acepta sólo consultas agregadas – tales que incluyen en el operador SELECT sólo funciones de agregación: $SUM(v_{ij})$, $AVG(v_{ij})$, $COUNT(X_z)=n_z$ y asi mismo, las funciones de agregación MIN, MAX y sobre todo, a veces, las formas más generales para la función SUM, por ejemplo, $SUM(v_{ij}^k)$ para la evaluación de los momentos altos de distribución.

Denotamos una consulta a la base de datos estadística como $p(Z)|agr(v_i)$, $i \in m$, donde agr() denota una función de agregación y Z se llama la fórmula característica (la expresión lógica) de la consulta:

$Z : \{z_i, i \in m\}$,

$z_i : \{ \cup_{i \in r}(v_i \; \mathbf{op} \; \alpha_j)\}$, donde ***op*** - cualquier operación del conjunto de operaciones de comparación, y $\alpha_j \in V_j$.

**Observación 2.1:** según la descripción, cada consulta puede tratarse más exactamente como el predicado $P(Z: z_i, i \in m)$, de dos valores (la función lógica no uniforme) y **m** variables, donde los valores de las variables ( $z_1$, ..., $z_m$) se escogen de los conjuntos correspondientes de los valores $V_1$, $V_2$,...$V_m$. En el contexto de consideración, **P** se trata como un conjunto característico de elementos del conjunto general W para los cuales la formula característica **Z** del predicado **P(Z)** es verdadera. Cuando el conjunto general W no se cambia, el conjunto característico **P** está bien definido por la fórmula característica Z correspondiente, es decir, el conjunto característico se define unívocamente por la fórmula característica correspondiente.

Para la exposición siguiente se necesita considerar más detalladamente las relaciones entre las fórmulas características de las consultas. Para este propósito consideramos la clasificación de todas las consultas según el criterio "cantidad de las variables". Entonces tenemos **m** tipos de las consultas: $\mathbf{P_1}$,$\mathbf{P_2}$, ..., $\mathbf{P_m}$ y cada tipo **Pi** incluye $C^m_i$ de las fórmulas características diferentes. De tal manera, la clasificación aceptada define el conjunto ÿ de $2^m$ subconjuntos de las clases de las fórmulas características.

Ahora consideramos dos pares de los tipos particulares de las consultas:

$P^*_1(Z: z_i)$, $P^*_2(Z: z_i,z_j)$, j = i, i,j = 1,...,m; y $P_1(Z:z_i)$, $P_2(Z: z_n,z_j)$, n,j, i, i,n,j=1,...,m;

Dentro del primer par $P^*_1$, $P^*_2$ de los conjuntos característicos existe la relación de inclusión $P^*_1 \subset P^*_2 \subset N$, porque si $v_i \in Vi$, entonces $v_i \in V_i \times V_j$. Pero al considerar otro par $P_1$, $P_2$, no podemos decir nada sobre cualquier relación entre estos tipos diferentes de las consultas. De tal manera, al considerar por analogía de todo par de los tipos particulares de las consultas, tenemos: $P^*_1 \subset P^*_2 \subset ... \subset P^*_m \subset W$. A veces, un par de conjuntos conectados por la operación de inclusión se llaman conjuntos comparables.

**Definición 2.1**. Cualquier par de los conjuntos característicos de las consultas diferentes son *conjuntos comparables* si las formulas características de estas consultas contienen por lo menos un atributo común.

De esta consideración sigue que la relación de inclusión parte el conjunto ÿDÿ de todos los pares de las clases de consultas en dos subclases: subclase $\mathbb{C}1 \subset \mathbb{C}$ de todos pares de las *consultas comparables*, y subclase $\mathbb{C}2 \subset \mathbb{C}$ de pares de las *consultas no comparables*.

Al seguir adelante con la clasificación de las fórmulas características de las consultas, consideramos la familia de las consultas simples con un atributo arbitrario $v_{j\bullet}$:

$P_1(Zi: z = v_{ji}) | \; COUNT_i(), \; i = 1,\ldots,r.$ (1)

Si N/r >k entonces todas estas consultas son legales y sobre todo, para cualquier par de la familia de los conjuntos característicos correspondientes $P_{Zi}$, $P_{Zk}$ se cumple: $P_{Zi} = P_{Zk}$, si i=k, y $P_{Zi} \cap P_{Zk} = \emptyset$, si $i \neq k$. De tal manera la familia particular de las consultas (1) define la clase de equivalencia $\mathbb{C}^1\sim : \cup_{j \in m} P_1,_{Zj}$. De la misma manera se puede ver que la familia de las consultas particulares de dos variables $P_2(Z :z1=v_{ji}, z2=v_{ki})$, j,k ∈ m, i = 1,...r, también define la clase de equiva-

lencia. De esta consideración breve es obvio que existe una gran variedad de subconjuntos de las consultas que definen las clases de equivalencia diferentes. Aquí omitimos la consideración más detallada porque para los objetivos de este trabajo es suficiente mostrar que, entre el conjunto de todas las consultas existe una gran cantidad de subconjuntos de las consultas las cuales pertenecen a clases diferentes de equivalencia.

Las bases de datos estadísticas en caso general incluyen varios atributos sensibles – denotados $v_s$ – en tal sentido que sus valores particulares $v_{si} \in V_s$ (o una función de agregación $agr_P(v_{si})$ para un conjunto pequeño de valores particulares) se clasifican por la política de seguridad como secreto. Entonces, para prevenir la revelación del secreto por medio de una aplicación directa de una función de agregación a un atributo $v_{si}$ particular, en las bases de datos estadísticos el tamaño permisible de un conjunto característico P se restringe por razones de la seguridad tal que $k \leq |P| \leq N–k$, donde k – es el límite inferior para el tamaño de un conjunto característico correspondiente a la fórmula característica Z de una consulta, y N-k es el liímite superior correspondiente a la negación de la formula característica ¬Z de la misma consulta. Para las consultas que satisfacen a las restricciones impartidas por la política de seguridad se utiliza el término "consultas legales" y estas consultas regresan la respuesta al usuario. Para las consultas que no satisfacen a estas restricciones se utiliza el término "consultas ilegales" y la base de datos estadística rechaza estas consultas.

En [2], por ejemplo, se recomienda, para prevenir la posibilidad de la inferencia lógica establecer el valor k muy grande, cerca de N/4.

Para la exposición siguiente tenemos que tomar varias preposiciones.

1.- Se supone que la base de datos estadística se utiliza por una comunidad de especialistas experimentados y bien familiarizados con el contenido de los datos en esta base: cantidad de los atributos y bandas de valores para cada atributo, y sobre todo, los valores de restricción **k** y **N-k** para la cantidad de elementos en el conjunto característico legal. Esta preposición es muy grave para la seguridad y mientras es más natural, porque en caso contrario la base de datos estadística es una caja negra para cualquier usuario y la eficacia de utilización de esta fuente de datos es casi nula.
2. Los usuarios recuerdan todas las respuestas recibidas por las consultas legales.
3. Algunos usuarios pueden recibir del ambiente externo una información sobre valores de varios campos para varios registros particulares de la base de datos estadísticos dada, pero no pueden saber el contenido completo para algún registro particular. Esto significa que un usuario es capaz de construir una consulta ilegal $X_Z$, pero no sabe el valor de secreto correspondiente.

## 3. Métodos de la inferencia lógica.

De las publicaciones dedicadas a la descripción de los métodos de la inferencia lógica tres métodos básicos son más conocidos – *tracker individual*, *tracker general* y *tracker doble*, donde el término tracker se utiliza para las consultas legales que se utilizan en el proceso de evaluación (deducción) de valores particulares (o grupos pequeños de los valores particulares) de los valores sensibles (que se tratan como valores secretos).

**a. Tracker individual**. En [1] se mostró que una base de datos estadística puede estar comprometida a pesar de las restricciones para el límite bajo y el límite alto de la cantidad de elementos en el conjuntos característicos de las consultas legales.

Este método consiste en lo siguiente:

**S**ea $P(Z^*|agr(v_{si}))$ - una consulta ilegal con la fórmula característica $Z^*$ tal que el tamaño de su conjunto característico $|X_z| < k$ (de aquí en delante por razones de brevedad no consideramos fórmulas de inferencia para negación de las fórmulas características). Entonces para evaluar el valor $agr_Z(v_{si})$ en este caso la fórmula característica $Z^*$ se descompone en dos fórmulas características legales:

$$Z^* = Z1 \cap Z2 \qquad (2)$$

de tal manera que:

$$k < COUNT(Z1 \cap \neg Z2) < COUNT(Z1) < N-k \qquad (3)$$

Si la condición (3) se cumple, entonces la fórmula característica $Z1 \cap \neg Z2 = T$ es legal y se llama el *tracker individual*. La evaluación del secreto $agr_Z(v_{si})$ correspondiente a la consulta ilegal en este caso se realiza como sigue:

$$agr_Z(v_{si}) = avg_{Z1}(v_{si}) - avg_T(v_{si}); \qquad (4)$$

Esto significa que para el usuario es suficiente evaluar dos consultas legales

$P(Z1)|agr(v_{si})$ y $P(T)|agr(v_{si})$ y después evaluar la diferencia entre los valores agregados recibidos. El Diagrama de Euler para este método se ilustra en la figura 1.

**Observación 3.1a**: en este método (así como y en la discusión más baja) se utiliza la preposición (2.3) que el usuario sabe la función $Z^*$ de selección para un registro individual (o para el conjunto pequeño de los registros) y para poder deducir la respuesta para la consulta ilegal se utilizan dos consultas legales $P(Z_1)$ y P(T), tales que sus conjuntos características $P_T$ y $P_{Z1}$ son conjuntos comparables: $P_T \subset P_{Z1}$,

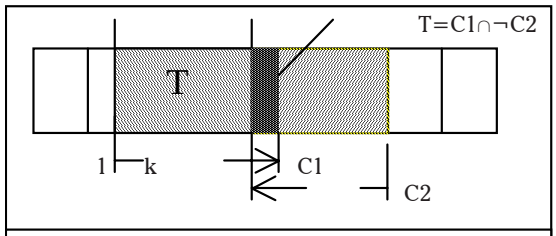


Figura 1. Diagrama de Euler para el tracker individual. Aquí N-cardinalidad de la base estadistica, k, N-k-límite bajo y límite alto para las consultas

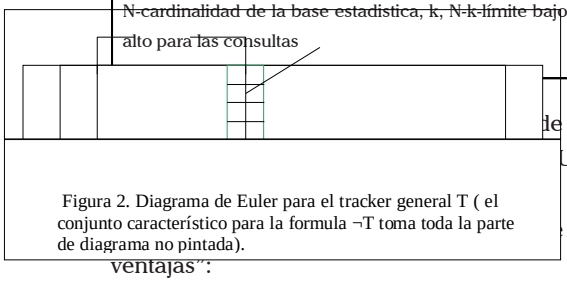

Figura 2. Diagrama de Euler para el tracker general T ( el conjunto característico para la formula ¬T toma toda la parte de diagrama no pintada).

...de estos con-

...UNT($P_{Z1}$) –

... dos "des-ventajas":

- Obviamente, que para cada nueva consulta ilegal hay que buscar la descomposición (2) de nuevo, pero encontrar esta descomposición no siempre es fácil sin conocimiento de las funciones de distribución para los valores de cada atributo.
- Según el modelo aceptado, la mayoría de los registros en la base de datos estadísticos tienen la cantidad aleatoria de duplicados, entonces casi para cada evaluación de la consulta ilegal con el valor $COUNT(P_Z) > 1$ el usuario no puede estar seguro si el valor evaluado es el valor secreto único, o bien, es el resultado promedio de varios valores secretos agregados.

**b. Tracker general.** En [2] fue mostrado que una base de datos estadística puede estar comprometida por medio de un tracker general T, eso es, cualquier consulta T que satisface a las restricciones $2k < COUNT(T) < N-2k$, puede ser utilizada para evaluar a cualquiera consulta ilegal $P(Z^*)$. Para este propósito en [2] ha sido propuesta la fórmula para evaluación:

**s**ea T – un tracker general y $P(Z^*)|agr(v_{si})$ es la consulta ilegal, entonces el valor secreto $-agr(v_{si})$,i<k- se evalúa de la manera siguiente:

$$agr_Z(v_{si}) = agr_T(v_{si}) + agr_{\neg T}(v_{si}) - agr_{T+Z}(v_{si}) - agr_{\neg T+Z}(v_{si}); \quad (5)$$

El Diagrama de Euler correspondiente se muestra en la figura 2.

$P_T$ $X_C$

1 k
N-k N

del diagrama no pintada).

**Observación 3.1b**: Los límites dobles 2k y N-2k para el tracker T en este método se utilizan para garantizar la posibilidad de evaluación de las consultas ilegales cuando su tamaño del conjunto característico está cerca de k. Obviamente que cuando una consulta ilegal caracteriza un registro particular, entonces cualquiera consulta legal puede servir como el *tracker general* T.

**Observación 3.2b**: En la figura 2 se muestre el caso general, cuando una parte del conjunto característico $X_Z$ de la consulta ilegal $P(Z^*)$ se incluye en T y otra parte de $X_Z$ en ¬T, y para evaluar cada parte se utiliza un par de las consultas comparables $P(T+Z^*) \subset q(T)$ o $q(\neg T+Z^*) \subset q(\neg T)$ que satisfacen a las condiciones:

$$COUNT(X_{T+Z}) - COUNT(X_T) < k$$
$$COUNT((X_{\neg T+Z}) - COUNT(X_{\neg T}) < k$$

Obviamente que si la consulta ilegal identifica un registro particular, entonces de estos dos pares necesito evaluar sólo un par de las consultas apropiadas, así pues, la fórmula (5) se reduce a la fórmula (4) del método de *tracker individual*.

Este método es mas fácil en la práctica que el método de tracker individual, porque en este caso no necesita buscar la descomposición apropiada para la consulta ilegal, y sobre todo, también no necesita ejecutar el algoritmo para la búsqueda del tracker general

– [1], porque bajo la preposición (1) y observación (**3.1b)** cada usuario tiene a su disposición la gran variedad de las consultas legales y cada una de ellas potencialmente puede ser utilizada como el tracker general.

La "desventaja" principal de este método, así como en el método anterior, se condiciona por la existencia de duplicados: después de cada evaluación de la consulta ilegal con el valor $COUNT(X_Z) > 1$ el usuario no puede estar seguro si el valor evaluado es el valor secreto único, o bien, es el resultado promedio de varios valores secretos agregados.

**c. Tracker doble**: en [2] un *tracker doble* se define como un par de las consultas comparables U y T tales que:

$T \subset U$,

$kRCOUNT(X_T)RN-2k$, y $2kRCOUNT(X_U)RN-k$.

La fórmula para evaluación el valor secreto correspondiente a una consulta ilegal $P(Z^*)|\ agr(v_{si})$ es como sigue ([1]):

$$agr_Z(v_{si})=agr_U(v_{si})+agr_{Z+T}(v_{si})-agr_T(v_{si})-agr_{\neg(Z\bullet T)\bullet U}(v_{si}) \qquad (6)$$

Diagrama de Euler para este método se muestra en la figura 3.

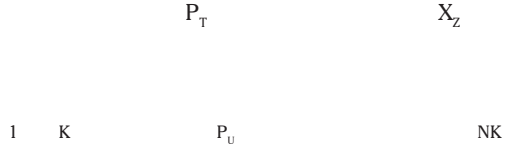


Figura 3. Diagrama de Euler para el tracker doble (U,T):T⊂U

**Observación 3.1c**: Los límites: (2k, N-k) para $Z_T$ y (k, N-2k) para $Z_U$ en este método se utilizan para garantizar la posibilidad de evaluación de las consultas ilegales cuando su tamaño del conjunto característico es cerca de k. Obviamente, que si una consulta ilegal caracteriza sólo varios registros particulares, entonces casi cualquier par de las consultas legales comparables pueden servir como el tracker doble.

**Observación 3.2c**: En la figura 3 se muestra el caso general, cuando el conjunto característico $X_Z$ de la consulta ilegal se descompone por tres partes: una parte $X_Z^1$ se incluye en el conjunto característico U, $X_Z^1 \subset U$; otra parte $X_Z^2$ se incluye en el conjunto característico $U \cap T$, $X_Z^2 \subset U \cap T$; y la tercera parte $X_Z^3$ se incluye en el conjunto característico $\neg U$, $X_Z^3 \subset \neg U$. Obviamente que para casos particulares esta fórmula se reduce: por ejemplo, si el conjunto característico $X_Z$, de la consulta ilegal incluye sólo un registro, entonces la fórmula (6) se reduce a la fórmula (3) con un par de consultas apropiadas. Generalmente, en cualquier caso particular: $X_Z \subset X_T$, o $X_Z \subset X_U$, o $X_Z \subset X_{\neg U}$ la fórmula general (6) para el tracker doble se simplifica correspondientemente.

**Observación 3.3c:** En este método, así como en los métodos anteriores, para deducir el secreto por medio de evaluación de una consulta ilegal se necesita construir varias consultas de tal manera que sus conjuntos característicos serian comparables, es decir: $X_{C+T} \subset X_T$ y $X_U \subset X_{\neg(CnT)nU}$, y sobre todo, $COUNT(X_{T+C})$– $COUNT(X_T) < k$,

$COUNT(X_U) - COUNT(X_{\neg(C\bullet T)\bullet U}) < k$, (según definición, $X_U \subset X_T$).

La "desventaja" principal de este método, así como el método anterior, se condiciona por la existencia de duplicados: después de cada evaluación de la consulta ilegal con el valor $COUNT(X_Z) > 1$ el usuario no puede estar seguro si el valor deducido es el valor secreto único, o bien, es el valor promedio de varios valores secretos agregados.

Ahora bien, la tecnología de inferencia lógica utilizada en todos los métodos considerados es bastante comprensible, y al tener en cuenta la proposición (1) sobre el conocimiento y experiencia de los usuarios, se ve como sigue:

Sea la fórmula característica Z que identifica un conjunto pequeño de registros $|X_Z| < k$ o un registro particular, entonces para deducir el valor secreto $agr_Z(v_s)$ se necesita juntar esta fórmula lógica Z con cualquiera fórmula característica de una consulta legal $P(B)|agr(v_{si})$ tal que $A=Z \cup B$, evaluar la consulta $P(A)|agr(v_{si})$ y recibir el valor del secreto deseado:

$agr_Z(v_{si})=agr_A(v_{si})-agr_B(v_{si})$– según la preposición (2) cada usuario recuerda todas las consultas evaluadas.

Esta última observación significa que la deducción de secretos por medio de las consultas legales específicas es equivalente a la solución de una ecuación para conjuntos del tipo:

$$Z \cup B = A$$

Efectivamente, la resolución formal para esta ecuación es como sigue (omitimos las evaluaciones intermedias):

$$(\neg A \cap X) \cup ((A+B) \cap \neg X) = \emptyset$$

de donde surge:

$$\neg A \cap X = \emptyset \text{ y } (A+B) \cap \neg X = \emptyset$$

De tal manera, la solución existe si y sólo si $X \subset A$ y $X \subset A+B$ y según la transitividad: $A+B \subset A$ o $B \subset A$. La solución es cualquier conjunto X que satisface a la condición de inclusión: $A+B \subset X \subset A$, o $A \backslash B \subset X \subset A$; en el caso $B \subset A$, el diagrama de Euler correspondiente a esta solución se muestre en la figura 4.

Podemos notar que todos los métodos de la inferencia lógica observados se basan en la resolución de una ecuación para conjuntos y nombramos estos métodos como la inferencia lógica algebráica.

**Observación 3.1**. De lo dicho anteriormente y sin tener en cuenta la preposición (2.1) del modelo de la base de datos estadística, los valores **k** y **N-k** son capaces de proteger la base de datos estadística en contra del acceso directo a los valores de campos en un registro particular, pero no tienen que ver con la protección en contra de métodos de la inferencia lógica.

De toda la discusión anterior sobre los métodos de la inferencia lógica tenemos, que cualquier intento de la inferencia lógica en una base de datos estadística puede ser detectado por medio de la revisión de la propiedad de inclusión entre cada par de las consultas comparables de la secuencia arbitraria Q de las consultas. Más rigurosamente:

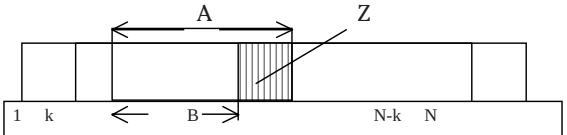


Figura 4. Diagrama de Euler para equacion $Z \cup B = A$ para el caso $A \backslash B \subset A$. Z incluye necesariamente el dominio llenado y sobre todo, puede incluir cualquier subconjunto del conjunto B.

**Teorema 1:** Sea $Q:\{q_1(Z_1), q_2(Z_2), ..., q_i(Z_i), ...\}$ la secuencia arbitraria de las consultas legales y $X_{Z1}, X_{Z2}, ..., X_{Zi}, ...$ los conjuntos características correspondientes, entonces el censor

D= D(D1,D2,Q), donde:

$D1:\{(X_{zi} \subset X_{zj}) \cup (X_{zj} \subset X_{zi})\}$, i=1,2 ...; j=1,2,..., i-1};

$D2 :\{|COUNT(X_{zj}) - COUNT(X_{zj})| < k\}$;

es el censor necesario y suficiente para detectar cualquier intento de inferencia lógica algebraica entre la secuencia arbitraria Q de las consultas legales.

Para demostrar la propiedad de necesidad para el censor D es suficiente demostrar que la solución de una ecuación para conjuntos existe sólo en clase ...1 de pares de las consultas comparables. De tal manera, la demostración sigue directamente del hecho que cualquier ecuación para conjuntos con la parte derecha respecto del conjunto X puede ser presentado en la forma:

$$(X \cap M) \cup (\neg X \cap N) = \emptyset \qquad (7)$$

donde M y N son conjuntos arbitrarios que no incluyen X. Dado que la unión de dos conjuntos es vacio sólo cuando cada uno de ellos es vacio, entonces tenemos el sistema de dos ecuaciones: $X \cap M = \emptyset$, y $\neg X \cap N = \emptyset$. Este sistema tiene sentido sólo cuando $X \subset \neg M$ y $N \subset X$, y por consecuencia: $N \subset \neg M$ (según transitividad de la relación de inclusión). De tal manera el conjunto de solución X existe solo bajo la condición de inclusión $N \subset X \subset \neg M$. Entonces, la parte D1 del censor, que lleva a cabo la revisión de inclusión para cada par $P_{Zi}$, $P_{Zj}$ , j i, i,j=1,2..., de los conjuntos característicos correspondientes a todos los pares de las consultas de $Q^pQ$, es necesario y suficiente para detectar cualquier intento de resolver una ecuación para conjuntos. Pero la parte D1 del censor no es suficiente para separar de todos los pares de las consultas comparables, tales que pudieran violar la política de seguridad. De tal manera, que la condición de suficiencia para el censor debe de seguir de la política de seguridad y esta condición se imparte por medio de la parte D2 del censor D.

**Observación 3.2:** si los conjuntos $P= P_1 \cap P_2 ...$ , $A=A_1 \cap A_2 ...$ y $B=B_1 \cap B_2 ...$ son compuestos, entonces las condiciones de existencia de solución para la ecuación (6) no se cambian.

**Observación 3.3:** De la descripción del modelo de la base estadística sigue que las operaciones de comparación ($P_{Zj} \subset P_{Zi}$) entre un par de conjuntos característicos $P_{Zi}$ y $P_{Zj}$ son equivalentes a las operaciones de comparación entre las fórmulas $Z_i$, $Z_j$ lógicas de las consultas subyacentes durante una sesión entre las actualizaciones de la base estadística.

Desdichadamente, los métodos algebraicos de la inferencia lógica no agotan todas las posibilidades de deducir secretos por medio de la manipulación con las consultas legales. En particular, existe otro grupo de métodos de la inferencia aun más poderosas, las cuales se basan en la utilización de las clases de equivalencia de las fórmulas características de las consultas. Aquí nombramos este grupo de métodos de inferencia lógica como métodos de inferencia lógica por equivalencia.

En la descripción del modelo de la base de datos estadísticos, se ha mostrado que la familia de las consultas (1) imparte la clase de equivalencia $P_{1j}\sim = \cup_{j\in m}$ Pzj, es decir el conjunto de los subconjuntos disjuntos y tales que $P_{Z1}\cup P_{Z2}\cup...\cup P_{Zm}=W$.

Ahora bien, si todas las consultas del tipo (1), para un atributo secreto -$P_1(Zi:z=v_{si})|$ $COUNT_i()$, i=1,…,r -ya han sido ejecutadas, entonces para deducir el valor secreto $v_{si}$ para cualquier consulta ilegal X($Z^*$) que define un registro particular, es suficiente repetir una por una las consultas extendidas $P(Zi:((z=v_{si})\cup Z^*))|COUNT^*_i()$, i = 1,…, r,

al comparar los valores $COUNT_i$ y $COUNT_i^*$, entonces la igualdad $COUNT_i() = COUNT^*_i()$ indicare que $Z^*\subset P_{Zi}$ y esto significa que el valor secreto $v_{si}$ corresponde a la consulta ilegal $Z^*$.

**Observación 3.4:** de este ejemplo se tiene que valores de restricción k y N-k, como antes, no tienen que ver con la protección en contra de inferencia lógica por el método de equivalencia.

La "ventaja" principal de este método es su propiedad de deducir exactamente un valor secreto a pesar de existencia de los registros duplicados.

De la descripción del censor es claro, que en caso de inferencia lógica por equivalencia, el censor D cuando $COUNT_i()$ $COUNT^*_i()$ también rechaza todas estas consultas extendidas según los resultados de las comparaciones con las consultas correspondientes previas, no dependiente del resultado de comparación de los valores COUNT. Pero el evento COUNT = COUNT* tiene una especificación sutil y este evento debe de ser considerado más concienzudamente.

Efectivamente, cuando los conjuntos característicos de dos consultas arbitrarias coinciden esto puede significar que se realiza una prueba de inferencia o bien, alguna consulta simplemente se repite. Si se rechazan las consultas repetibles, entonces cualquier consulta que ya ha sido ejecutada por un usuario no puede ser ejecutada después por otro usuario. Para evitar esta situación tediosa y específica para las pruebas de inferencia con el uso de las clases de equivalencia, se necesitó aumentar el censor D por la parte D2, que pudiera analizar los casos de coincidencia aparte.

**Definición 3.1**: un par de consultas, los conjuntos característicos de las cuales coinciden, son repetibles si y sólo si ambas pertenecen a una clase en 1.

**Definición 3.2**: En un par de consultas, los conjuntos característicos de las cuales coinciden, son consultas de equivalencia si las clases de ellas en 1 son diferentes.

Entonces, para procesar correctamente eventos de coincidencia de los conjuntos característicos de las consultas comparables, según definiciones 1 y 2, es suficiente empotrar en el censor D adicionalmente la comparación de las clases de las consultas en el caso de un evento $|COUNT(X_{Zj}) - COUNT(X_{Zj})| = 0$.

Así pues, si denotamos la clase de una consulta con la fórmula característica Z como $P_Z\sim$, $P_Z\sim\subset 1$, entonces el censor D* (D1,D2,D3,D4) extendido:

$$D1 : \{(X_{Zi}\subset X_{Zj})\},\ i = 1,2 ...;\ j = 1,2,..., i-1\},D2: \{(X_{Zj}\ ?\ X_{Zi})\}, \quad (8)$$
$$D1\cup D2 \to D3: \{|COUNT(X_{Zj}) - COUNT(X_{Zj})| < k\},$$
$$D1\cap D2 \to D4 : \{P_{Zi}\sim = P_{Zj}\sim\}$$

es necesario y suficiente para detectar cualquier prueba de la inferencia lógica o bien por el método algebraico, o bien por el método de equivalencia.

**Observación 3.5:** del ejemplo anterior se tiene que el método de inferencia lógica basado en la utilización de las clases de equivalencia también puede realizarse sólo en la clase 1 de las consultas comparables. Entonces, según el teorema 1, el censor (8) extendido guarda las propiedades de necesidad, y sobre todo, desde que la extensión (8) no viola la política de seguridad, entonces la propiedad de suficiencia se guarda también.

## 4. Sistema de protección en contra de la inferencia lógica.

Como sistema de protección en la base de datos estadística aquí se utiliza el concepto que ha sido propuesto en [3]: "para los secretos desconocidos rechazo lo cual es mejor que mentir", esto significa que la base de datos estadística con el sistema de protección en contra de inferencia lógica regresa la respuesta exacta para las consultas legales y rechaza (no responde) a las consultas ilegales.

De la descripción de los métodos de la inferencia lógica sigue:

**Definición 4.1**: las consultas de dos tipos: $p\{Z: (z_i\ \boldsymbol{op}\ v_i),\ v_i\ \ v_s,\ i=1,2,..\}|agr(v_{si})$ o bien $p_i\{Z:(z_i\ \boldsymbol{op}\ v_i),\ v_i=v_s,\ i=1,2,..\}|agr()$ son consultas "sospechosas" – las cuales denotamos como p*(Zi).

El sistema de protección S puede ser representado por el cuádruplo:

S(Q, B, D, ÿ1), donde

Q: {$p_i$(Zi)|agr(), i =1,2,....}, como antes, la serie infinita de las consultas, B: {$\dot{E}_{j\supset J}$ $Z_j$} – base de conocimientos reflexiva (en tal sentido que " la base sabe que el usuario sabe...") la cual guarda todas las fórmulas características $Z_j$ de las consultas sospechosas de la serie Q no rechazadas por el censor y enviados al usuario;

D – censor (7)

ℂ1 - el conjunto de todas las clases de las consultas comparables.

En la definición 4.1 se asume que no todos los atributos en una base de datos estadística se definen por la política de seguridad como secretos, en el caso contrario obviamente que todas las consultas son sospechosas. Efectivamente, no hay mucho sentido marcar como secretos tales atributos en una base de datos si sus valores son disponibles (según preposición (3)) del ambiente externo.

La expresión (8) define el censor extendido D* (D1,D2,D3,D4) solo como una colección de las operaciones necesarios y suficientes para detectar cualquier prueba de inferencia lógica. Si se tratan las operaciones D1,D2,D3,D4 como enunciados entonces la descripción general del funcionamiento del censor en el álgebra de preposiciones es como sigue:

(p*(Zi)→D1); (p*(Zi) →D2);

((D1∩D2) → D4) n(D4 → ¬p*()Zi)); (9)

((D1∩D2) →D3) n (D3 → ¬p*(Zi));

En esta descripción formal la sustitución ¬p*(Zi) significa rechazo de la consulta sospechosa si el valor D4 o D3 es verdadero (por deducción esto significa que los valores D1 o D2, o bien ambos, son verdaderos). Utilizamos la forma abreviada D(p*(Zi))→¬p*(Zi) para la expresión (7).

Los cambios de los estados de la base reflexiva B se describen recursivamente como sigue:

$Bt_0$ = Ø;

Bt:{$\cup_{j\in J}$ $Z_j$, J – la cantidad de las fórmulas características guardadas};

$B_{t+1}$: { $B_t$, si D(p*(Zi)) };

$B_{t+1}$: { $B_t \cup$ Zi, J= J+1, si ¬D(p*(Zi))};

El esquema generalizado del sistema de protección en contra de la inferencia lógica se muestra en la figura 5.

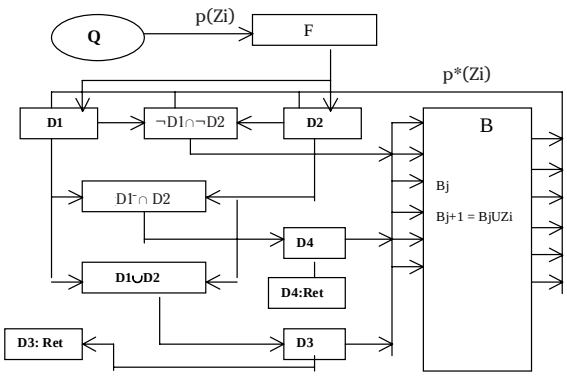


Figura 6. Esquema generalizado del sistema de protección en contra de la inferencia lógica en las bases de datos estadísticas (vease en el texto). (aquí F – denota un filtro hipotético que elige del flujo de las consultas sólo las consultas sospechosas).

El contenido de la base reflexiva en el momento inicial naturalmente es nulo. Después, de la secuencia de las consultas Q se elige cada consulta sospechosa p*($Z_i$) y se revisa por el censor. Obviamente que la primera consulta se inscribe en la base de datos reflexiva y la respuesta se envía al usuario porque la comparación con nulo resulta en negación ¬D. Cada consulta sospechosa siguiente se compara en el censor con cada una de las J consultas guardadas en la base reflexiva. Si la comparación para cualquier par - consulta corriente, consulta guardada - resulta en D, entonces la consulta corriente p*$_i$($Z_i$) se destruye y el contenido de la base reflexiva no se cambia, la base de datos estadística no responde a esta consulta y el sistema de seguridad espera la consulta sospechosa siguiente. Hay sólo una exclusión: una consulta repetible resulta en negación del censor ¬D, entonces el resultado de esta consulta se envía al usuario, pero la fórmula característica de esta consulta no se inscribe repetida a la base reflexiva. Así pues, el conocimiento de la base reflexiva coincide con el conocimiento de toda la comunidad de los usuarios y no incluye "conocimientos peligrosos" – es decir, no incluye cualquier par de las fórmulas características de las consultas comparables, que pudieran ser utilizadas para deducir un campo secreto particular.

**Observación 4.1:** Como se ve de la exposición anterior, el sistema de protección no distingue usuarios particulares al llevar a cabo la protección general en contra de los ataques individuales, o bien en con-

tra de ataques del grupo. Pero para el mantenimiento de una política de seguridad mas sensible puede ser útil partir la base de datos reflexiva en las partes personales para cada usuario.

**Observación 4.2:** Las consultas legales no sospechosas son desviadas y el sistema de protección y la base siempre regresa respuesta al usuario. De tal manera cualquier usuario es libre para deducir el valor particular de un atributo no secreto, pero la probabilidad de éxito en este caso se reduce muy rápido, cuando la cantidad de los atributos secretos crece, porque cuando hay menos atributos disponibles, mas difícil es caracterizar atributos particulares. Entonces, el sistema de protección, al llevar a cabo la protección explícita de los atributos secretos, imparte la protección implícita en contra de la deducción de los valores particulares de los atributos no secretos.

**Observación 4.3:** Como fue mencionado anteriormente, los valores de límites k y N-k para el tamaño de los conjuntos característicos permisibles, bajo la preposición natural sobre los conocimientos y la experiencia de los usuarios, no ejercen ninguna influencia a la protección en contra de los métodos de la inferencia lógica, sino dirigen para prevenir el acceso directo a los campos de los registros particulares. Entonces, desde el punto de vista de disponibilidad, seria racionalmente establecer el valor k mínimo necesario para proveer el intervalo de confianza aceptable para las estadísticas legales ( en este caso puede ser que estos valores deban de ser diferentes para atributos diferentes).

**Observación 4.4:** para cada consulta corriente sospechosa el censor (8) del sistema de protección debe de ejecutar no más de J comparaciones, así pues la cantidad de comparaciones es la función lineal del tiempo y por consiguiente tiene la tendencia de crecimiento discontinuo. Hay varias posibilidades para eliminar esta desventaja, pero para la discusión detallada de estas posibilidades se necesita considerar en detalles varios mecanismos de DBMS, lo que está fuera de este trabajo.

## Referencias


J. Schlorer.
975 "Identification and retrieval of personal records from statistical data bank." *Methods of Info. in Medicine* 14, 1, 7-13.

D. Denning, J. Schlorer.
1980 "A fast procedure for finding a tracker in a statistical database." *ACM Transactions on Database Systems*, vol. 5, No. 1, Pages 88-102.

G. Sicherman, Wiebren de Jonge and R. Van de Riet.
1983 "Answering Queries Without Revealing Secrets." ACM Transactions on Database Systems, Vol.8, No.1, Pages 41-59.

Joachim Biskup.
2000 "For unknown secrecies refusal is better then lying." Elsevier Science, Data&Knoledge Engineering, 33, 1-23.

Raghu Ramacrishnan.
1998 "Database Management Systems". McGraw-Hill Inc., p. 325.

J. C. Fabre, T. Perennou.
1997 "Processing of confidential information in distributed systems by partitioning." Computer Communications. Elsevier Science, 20, 177-188.



***Yury Paramonov***
***Universidad Tecnológica de la Mixteca***