

Seguridad, privacidad y aspectos legales

Álvaro López García

Grupo de Computación Avanzada y e-Ciencia
Instituto de Física de Cantabria (IFCA) - CSIC-UC

Máster universitario en ciencia de datos / Master in Data Science



Parte I

Privacidad y anonimato

Tabla de contenidos

1. Motivación

2. Conceptos Generales

Anonimización

Técnica (o conjunto de técnicas) para preservar el anonimato y proteger la información privada y/o sensible de una serie de datos, preservando (hasta un cierto nivel) la utilidad de dichos datos.

- Eliminar información personal y/o sensible de un dataset, de forma que no sea posible obtener información individual de un sujeto.
- Problema: El dataset tiene que seguir siendo útil.
- Compromiso utilidad vs privacidad.
 - Un dataset perfectamente anonimizado permitirá no extraer información de ningún sujeto.
 - Pero tambien puede no ser útil.
- Un dataset anonimizado puede ser vulnerable (data aggregation, data linkage, cross correlation).
- ¿Es suficiente con eliminar la información que pensemos es sensible?

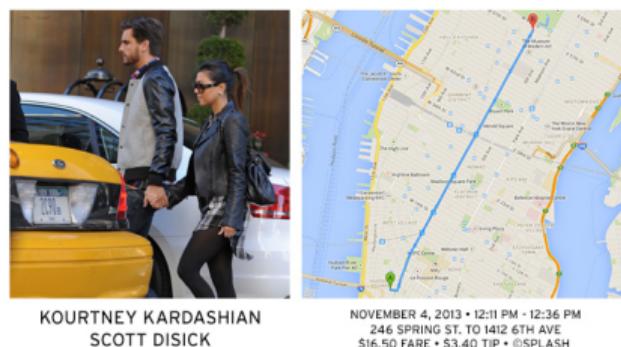
Primera aproximación

Primera y errónea

- Eliminar la información que se piense que es sensible.
- Ejemplo: Eliminar tan solo identificadores únicos (nombre, DNI, etc).
- Problema: La agregación de atributos puede ser suficiente para identificar de forma casi unívoca a una persona.
 - Ejemplo: Sexo, fecha de nacimiento, código postal.
- Es necesario usar métodos matemáticos que aseguren la privacidad.
 - K-anonimato, l-diversity, differential privacy, etc.

Ejemplo: Datos abiertos de la ciudad de Nueva York

- Nueva York publicó los trayectos de taxi¹ (lugar y hora de recogida, número de taxi, lugar y fecha de destino, etc.) como un dataset abierto.
- Ataque 1: Anonimizado a través de MD5 en principio, fue posible desanonymizarlo.
- Ataque 2²: Cruzando este dataset con fotos geolocalizadas (obtenidas a través de Google) de famosos subiendo en taxi se pudo obtener los trayectos que hacen.

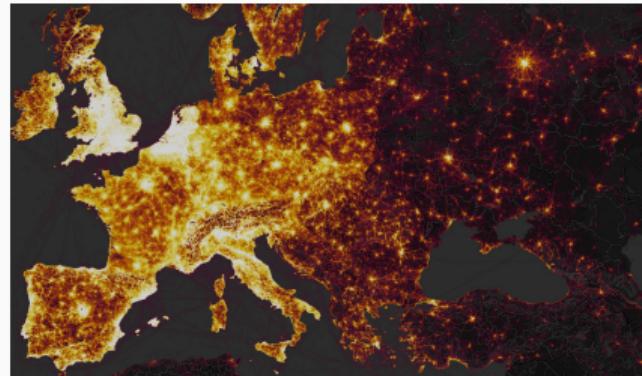


¹http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

²<http://gawker.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546>

Ejemplo: Heatmap de Strava I

- Strava es una aplicación móvil para registrar actividades deportivas (correr, bicicleta, etc.)
- Strava publicó un mapa de calor global con los datos registrados por sus usuarios.



- El heatmap es totalmente anónimo, pero...

Ejemplo: Heatmap de Strava II

Las rutas de running de La Zarzuela



Figura: Rutas de atletismo en el recinto de la Zarzuela.

Ejemplo: Heatmap de Strava III



Figura: Base de la legión en Viator (Almería).

Ejemplo: Heatmap de Strava IV



Figura: Bases de la OTAN en Afganistán.

Base de datos Tabla con N filas y M columnas (atributos).

Cardinalidad Singularidad de los datos contenidos en una columna.

- Cardinalidad alta implica que los datos son poco comunes.
- Cardinalidad baja implica que la columna tiene pocos valores únicos.

Alfabeto (Σ) de una base de datos. Rango de valores que puede tomar una celda en una base de datos determinada.

Divulgación de pertenencia Si un individuo está contenido en un dataset o no.

Divulgación de atributos Obtener información de un individuo, sin identificarlo con una entrada concreta de la base de datos.

Divulgación de identidad Un individuo se puede identificar con una entrada de la base de datos.

Definiciones: Tipos de atributos

Identificadores Permiten identificar únicamente a una persona.

- Ejemplo: DNI, Nombre, Apellidos.

Cuasi-identificador (CI) Conjunto de variables de la tabla que permiten identificar a un individuo (accesibles a un atacante).

- Ejemplo: Código Postal, Edad.

Atributo Sensible (AS) Propiedades de los individuos que no se debe asociar con ellos.

- Ejemplo: Enfermedad, estado civil, etc.

Atributos Insensibles no presentan ningún riesgo.

Definiciones: Tipos de atributos

Identificadores Permiten identificar únicamente a una persona. → **Suprimir**

- Ejemplo: DNI, Nombre, Apellidos.

Cuasi-identificador (CI) Conjunto de variables de la tabla que permiten identificar a un individuo (accesibles a un atacante). → **Generalizar**

- Ejemplo: Código Postal, Edad.

Atributo Sensible (AS) Propiedades de los individuos que no se debe asociar con ellos. → **Dejar intactos o modificar**

- Ejemplo: Enfermedad, estado civil, etc.

Atributos Insensibles no presentan ningún riesgo. → **Dejar intactos**

K-anonimato

K-Anonimato

Una base de datos es k-anónima, si para cada fila r hay al menos $k - 1$ filas idénticas.

K-Anonimato

Una tabla es k-anónima respecto a un cuasi-identificador CI si cada combinación de de atributos de las variables de CI aparece en almenos k tuplas.

- Cada grupo de registros indistinguibles se denomina *Clase de Equivalencia*

- Método propuesto en 1998 por Samarati y Sweeney³.
- **Protege frente a la re-identificación**
- Se eliminan o generalizan atributos de la base de datos (alta cardinalidad).
 - Supresión: se reemplazan los atributos con un *
 - Generalización: se reemplazan los atributos con rangos.
 - El alfabeto de una base de datos k-anónima es $\Sigma \cup \{*\}$
- Objetivo: Conseguir que haya al menos $k - 1$ filas idénticas para evitar la re-identificación.
- Se consigue eliminar, hasta cierto punto, el enlazado de la información.
- Problema: si se sabe que un individuo está en la base de datos, se puede extraer información.
- Problema: la k-anonimización es un problema NP-completo.
 - 3-Anonymity NP-Completo para $|\Sigma| = O(n)$ (Myerson & Williams, 2004)
 - 3-Anonymity NP-Completo para $|\Sigma| = 3$ (Aggarwal, et al, 2005)
 - 3-Anonymity NP-Completo para $|\Sigma| = 2$ (Bonizzoni, et al, 2007)

³Samarati, P., & Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression

Ejemplo

Base de datos hospital

Cuasi Identificadores: (Name, Age, Gender, State of domicile, religion)

Name	Age	Gender	State of domicile	Religion	Disease
Ramsha	29	Female	Tamil Nadu	Hindu	Cancer
Yadu	24	Female	Kerala	Hindu	Viral infection
Salima	28	Female	Tamil Nadu	Muslim	TB
Sunny	27	Male	Karnataka	Parsi	No illness
Joan	24	Female	Kerala	Christian	Heart-related
Bahuksana	23	Male	Karnataka	Buddhist	TB
Rambha	19	Male	Kerala	Hindu	Cancer
Kishor	29	Male	Karnataka	Hindu	Heart-related
Johnson	17	Male	Kerala	Christian	Heart-related
John	19	Male	Kerala	Christian	Viral infection

Ejemplo

Supresión

Name	Age	Gender	State of domicile	Religion	Disease
*	29	Female	Tamil Nadu	*	Cancer
*	24	Female	Kerala	*	Viral infection
*	28	Female	Tamil Nadu	*	TB
*	27	Male	Karnataka	*	No illness
*	24	Female	Kerala	*	Heart-related
*	23	Male	Karnataka	*	TB
*	19	Male	Kerala	*	Cancer
*	29	Male	Karnataka	*	Heart-related
*	17	Male	Kerala	*	Heart-related
*	19	Male	Kerala	*	Viral infection

Ejemplo

Generalización

Name	Age	Gender	State of domicile	Religion	Disease
*	$20 < \text{age} \leq 30$	Female	Tamil Nadu	*	Cancer
*	$20 < \text{age} \leq 30$	Female	Kerala	*	Viral infection
*	$20 < \text{age} \leq 30$	Female	Tamil Nadu	*	TB
*	$20 < \text{age} \leq 30$	Male	Karnataka	*	No illness
*	$20 < \text{age} \leq 30$	Female	Kerala	*	Heart-related
*	$20 < \text{age} \leq 30$	Male	Karnataka	*	TB
*	$\text{age} \leq 20$	Male	Kerala	*	Cancer
*	$20 < \text{age} \leq 30$	Male	Karnataka	*	Heart-related
*	$\text{age} \leq 20$	Male	Kerala	*	Heart-related
*	$\text{age} \leq 20$	Male	Kerala	*	Viral infection

2-anónima respecto a los Cuasi Identificadores: (Age, Gender, State of domicile)

Ataques

- Sensible a ataques utilizando bases de datos complementarias.
 - Ejemplo: Dos tablas publicadas de forma independiente.
 - Ejemplo: Dos tablas publicadas en dos momentos diferentes, con diferente cardinalidad.
- Sensible al conocimiento previo (p.e. si se sabe que una persona está en la base de datos).
- Sensible a la falta de diversidad.

C. Postal	Edad	Enfermedad
390XX	2X	Enfermedad Cardíaca
390XX	2X	Enfermedad Cardíaca
390XX	2X	Enfermedad Cardíaca
3920X	>=40	Gripe
3990X	>=40	Enfermedad Cardíaca
3990X	>=40	Cáncer
390XX	3X	Enfermedad Cardíaca
390XX	3X	Cancer
3900X	3X	Cáncer

- 3-anónima respecto a los CI (Código Postal, Edad).
- Falta de diversidad: Mengano Pérez, 20 años.
- Conocimiento previo: Fulano López, 36 años, código postal 39001.

I-diversidad y otros métodos

L-diversidad

Una base de datos k-anónima es l-diversa con respecto a un AS si para cada clase de equivalencia existen al menos l diferentes valores de AS.

- Objetivo: Proteger la extracción de atributos sensibles.

- Ofrece protección contra ataques de diversidad y de contexto
- Problemas y limitaciones:
 - Más complejo de implementar y de conseguir.
 - Puede haber más de un AS: l-diversidad respecto de un AS no garantiza l-diversidad de forma general
 - Ataques mediante inferencia probabilística.
 - Atributos muy sensibles (alta cardinalidad).
- l-diversidad con entropía (garantizar un mínimo de entropía $\log l$ en la distribución de los AS), l-diversidad recursivo.

t-cercanía

La distribución de los valores del AS en cada clase de equivalencia esté a una distancia no más cercana de t de la distribución del AS en la tabla original.

- Objetivo: Proteger la extracción de atributos sensibles.
- La distancia se calcula a partir de la *Earth Moving Distance* (EMD)⁴

Otros métodos

- δ -Disclosure
- β -Likeness
- δ -Presence
- Privacidad diferencial
 - No es una propiedad del conjunto de datos, sino de las herramientas, sistemas y algoritmos de procesado de datos.
 - Un sistema tiene privacidad diferencial si un atacante no puede saber si la información de un individuo se ha utilizado para realizar un cálculo determinado.
 - Imposible saber con certeza que individuos están en una base de datos.

Preguntas?