Data Privacy

Jordi Casas-Roma[1]

[1]Profesor de los Estudios de Informática, Multimedia y Telecomunicación
Director del Máster Universitario en Ciencia de datos (*Data science*)
Director del Máster en Inteligencia de Negocio y *Big Data*

Universitat Oberta de Catalunya
jcasasr@uoc.edu

UOC Data Day
Madrid, 21 de junio de 2017

índice

Antecedentes y contextualización

Ejemplos iniciales de publicación de datos

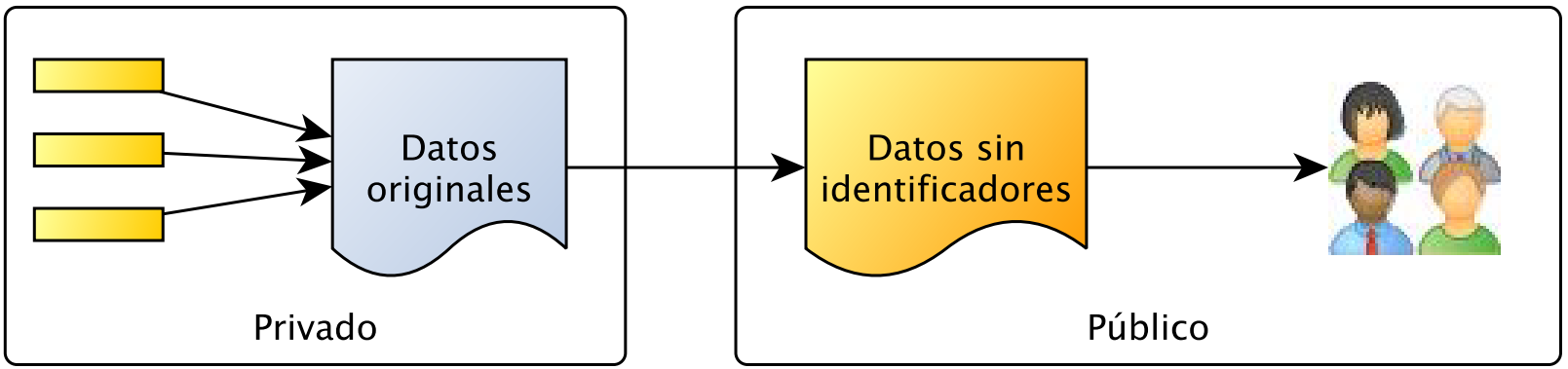
- En junio de 2004, el Comité Consultivo de Tecnologías de la Información (*Information Technology Advisory Committee*) de Estados Unidos publicó un informe titulado "Revolucionando la atención sanitaria a través de las tecnologías de la información". Un punto clave fue establecer un sistema nacional de registros médicos electrónicos que fomentara el intercambio de conocimientos médicos.
- Netflix publicó un conjunto de datos que contiene calificaciones de sus películas de 500.000 suscriptores, en un intento por mejorar la precisión de las recomendaciones de las películas basadas en las preferencias personales.
- AOL publicó un conjunto de registros de consultas, pero rápidamente se vio obligado a retirar los datos debido a la identificación de un usuario en los datos[1].

[1] M. Barbaro and T. Zeller. A face is exposed for AOL searcher no. 4417749. Technical report, New York Times, 08 2006.

Publicación de datos

Escenario básico para la publicación de datos (naïve anonymization)

- Proporcionar datos a terceras partes para realizar análisis.
- Preservar la privacidad de los usuarios que aparecen en los datos.



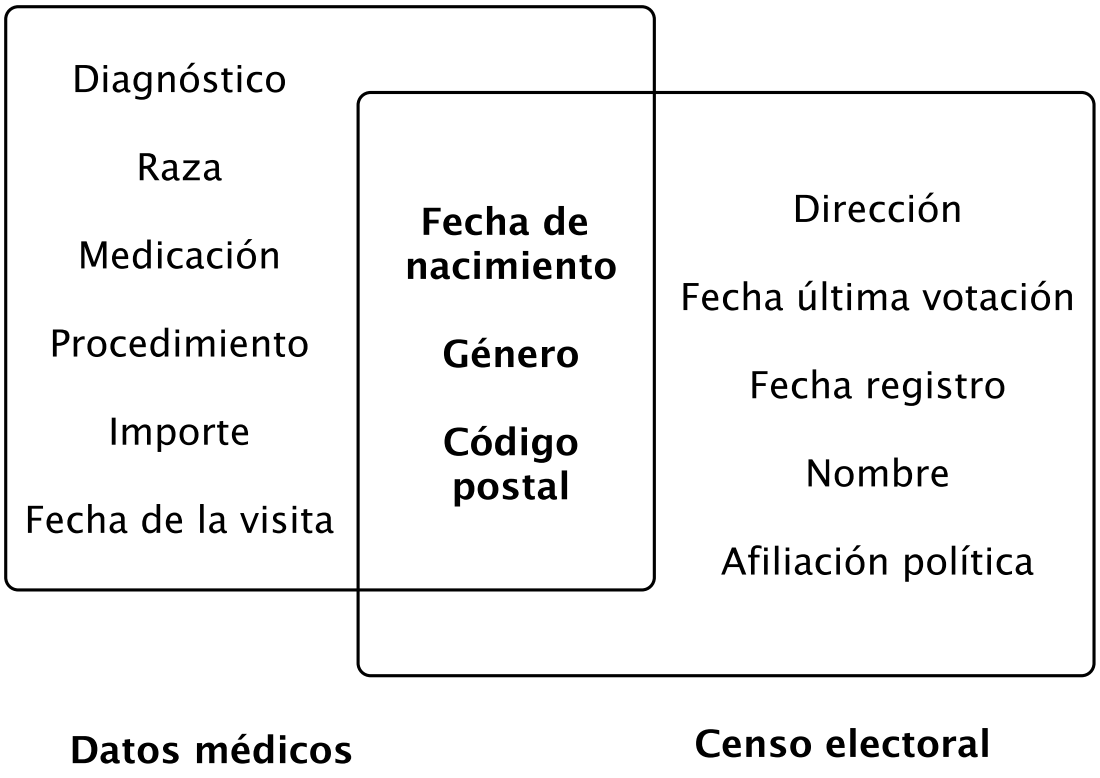
Tipos de datos

Tipología de datos según el tipo de información que contienen

- **Identificadores**: Conjunto de atributos que permiten identificar de forma explícita a un individuo.
- **Casi-identificadores**: Conjunto de atributos que potencialmente podrían identificar a un individuo.
- **Atributos sensibles**: presentan información específica y sensible de un individuo en concreto.
- **Atributos no sensibles**: los atributos que no caben en ninguna de las categorías anteriores.

Ataque de Sweeney[2]

- Re-identificación de un gobernador de Estados Unidos.
- Estudios posteriores elevan la cifra al 87 % de la población de Estados Unidos.

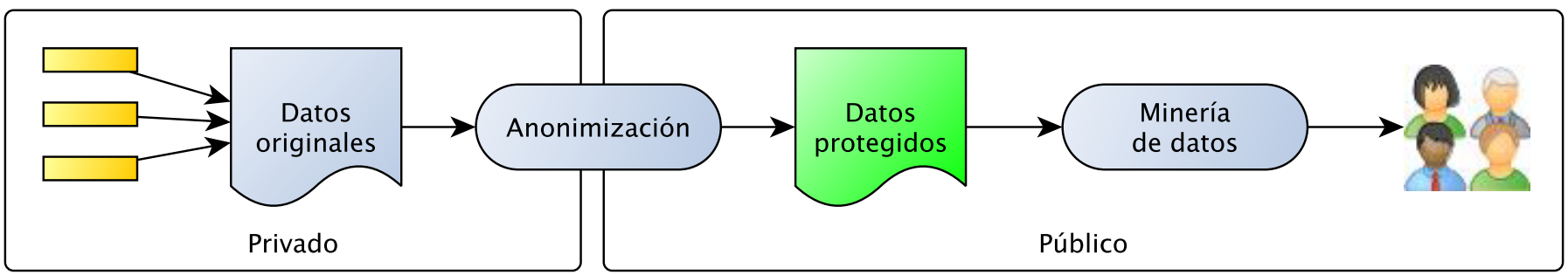


[2]Latanya Sweeney. "Achieving *k*-anonymity privacy protection using generalization and suppression". Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 10(5):571–588, 2002.

Preservación de la privacidad

Escenario básico para la publicación de datos preservando la privacidad

- El objetivo es que un único individuo sea indistinguible respecto a un conjunto de individuos suficientemente grande para proteger su identidad, de tal forma que el atacante sólo puede deducir cierta información con una cierta probabilidad.



Tipología de modelos de protección

Enfoques principales para limitar el riesgo de divulgación en procesos de publicación de datos:

- **Protección no interactiva**, mediante la cual se genera y se libera una versión protegida del conjunto de datos original recopilado de los sujetos de datos.
- **Protección interactiva**, mediante la cual se realiza una consulta de datos con fines analíticos análisis en el conjunto de datos original y, a continuación, se devuelve una versión protegida de los resultados al usuario que ha realizado la consulta.

Aleatorización

Definición

Consiste en introducir ruido en los datos originales, de tal forma que un atacante no pueda saber, a ciencia cierta, si la información que está extrayendo es cierta o ha sido alterada durante este proceso de anonimización aleatoria.

- Balance entre privacidad y utilidad de los datos.
- Método simple y eficiente en grandes conjuntos de datos.
- Dificultad para proteger los valores extremos (*outliers*).

Aleatorización

Ejemplo de perturbación mediante distribución normal

ID	Edad
1	45
2	30
3	74
4	72
5	73
6	27
7	84
8	52
9	62
10	14

ID	Valor perturbado
1	45
2	30
3	78
4	70
5	67
6	27
7	80
8	57
9	64
10	16

k-anonimidad

Definición

La k-anonimidad es una propiedad de los datos que garantiza que un individuo no pueda ser distinguido de otros $k-1$ individuos también representados en esos datos.

- Introducido por L. Sweeney[3] en 2002.
- Uno de los modelos de **protección no interactiva** más ámpliamente investigado y empleado en la publicación de datos.
- Problema NP-Hard.

[3] Latanya Sweeney. "K-anonymity: A model for protecting privacy". Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 10(5):557–570, 2002.

k-anonimidad

Casi-identificadores			**Atributos Sensibles**
CP	Edad	Nacion.	Enferm.
13053	28	Rusa	Arritmia
13068	29	Española	Arritmia
13068	21	Japonesa	Infección
13053	23	Española	Infección
14853	50	India	Cáncer
14853	55	Rusa	Arritmia
14850	47	Española	Infección
14850	49	Española	Infección
13053	31	Española	Cáncer
13053	37	India	Cáncer
13068	36	Japonesa	Cáncer
13068	35	Española	Cáncer

Cuadro 1: Datos originales

Casi-identificadores			**Atributos Sensibles**
CP	Edad	Nacion.	Enferm.
130**	< 30	*	Arritmia
130**	< 30	*	Arritmia
130**	< 30	*	Infección
130**	< 30	*	Infección
1485*	> 40	*	Cáncer
1485*	> 40	*	Arritmia
1485*	> 40	*	Infección
1485*	> 40	*	Infección
130**	< 40	*	Cáncer
130**	< 40	*	Cáncer
130**	< 40	*	Cáncer
130**	< 40	*	Cáncer

Cuadro 2: Datos k-anónimos, con $k = 4$

Privacidad diferencial

Definición

Un mecanismo de privacidad diferencial debe garantizar que la contribución de los datos de un individuo al resultado global de consulta es limitada. Es decir, La definición da garantías de que la presencia o ausencia de un individuo no afectará significativamente el resultado final del algoritmo.

- Introducida por C. Dwork[4] en 2006.
- Modelo de privacidad para la **protección interactiva** en el contexto de las bases de datos estadísticas.
- Pérdida importante de utilidad en los datos protegidos.

[4] Cynthia Dwork. "Differential Privacy". In International Conference on Automata, Languages and Programming, Volume 4052 of Lecture Notes in Computer Science, 2006. Springer-Verlag.

Privacidad diferencial

Formalmente, un algoritmo o función A es ϵ-diferencialmente privado si, y sólo si, para todos los conjuntos de datos D_1 y D_2 que difieren un sólo individuo, se cumple:

$$\frac{Pr[A(D_1) \in S]}{Pr[A(D_2) \in S]} \leq e^{\epsilon} \tag{1}$$

donde ϵ es un número real positivo y $S \subset rango(A)$.

Privacidad diferencial

Supongamos que pedimos a un grupo de personas que respondan a la pregunta "¿Tienes la enfermedad X?"
La respuesta de cada individuo seguirá el siguiente procedimiento:

- Tirar una moneda.
- Si sale "cara", entonces el individuo responderá con honestidad a la pregunta formulada.
- Si sale "cruz", luego se tira la moneda de nuevo y se responde "Sí" si sale "cara", y "No" si sale "cruz".

Así, si p es la proporción verdadera de personas con la enfermedad X, entonces esperamos obtener respuestas positivas de:

$$\frac{1}{4}(1-p)+\frac{3}{4}p=\frac{1}{4}+\frac{p}{2}$$

Por lo tanto es posible estimar p sin comprometer la privacidad de ninguno de los usuarios que responden a la pregunta que les formulamos.

Métodos de enmascaramiento

- **Métodos perturbativos**. El conjunto de datos original es perturbado de algún modo, y el nuevo conjunto de datos puede contener información errónea.
  - Ruido aditivo (*additive noise*)
  - Micro-agregación (*microaggregation*)
  - Intercambio de rango (*rank swapping*)
- **Métodos no perturbativos**. La protección se logra a través de la sustitución del valor original por otro valor que no es incorrecto pero es menos específico, es decir, más general.
  - Generalización (*generalization*)
- **Generadores de datos sintéticos**. En este caso, en lugar de distorsionar los datos originales, se crean nuevos datos artificiales para sustituir los valores originales.

Métodos perturbativos

Ruido aditivo (additive noise)

- Añadir distorsión o ruido en los datos originales.
- Por ejemplo, introducir el ruido siguiendo una distribución normal $N(0, p\sigma)$, donde
  - σ representa la desviación estándar de los datos originales
  - p es el parámetro que controla la cantidad de ruido

ID	Edad
1	29
2	48
3	21
4	36
5	45
6	58
7	72
8	22
9	25
10	43

ID	Edad
1	32
2	46
3	28
4	38
5	48
6	61
7	59
8	20
9	24
10	55

Métodos perturbativos

Micro-agregación (microaggregation)

- Crear grupos de datos según su similitud y reemplazar por el mismo valor (promedio, mediano, etc).
- Para cada valor específico de uno o más atributos existirán siempre un conjunto de registros.
- Dos casos principales:
  - micro-agregación univariante: aplica un único atributo.
  - micro-agregación multivariante: aplica a dos o más atributo al mismo tiempo.

ID	Edad
1	29
2	48
3	21
4	36
5	45
6	58
7	72
8	22
9	25
10	43

ID	Edad
1	27
2	46
3	21
4	39
5	46
6	65
7	65
8	21
9	27
10	39

Métodos perturbativos

Intercambio de rango (rank swapping)

- Intercambiar aleatoriamente los valores de un mismo atributo entre distintos registros.
- Ordena todos los valores del atributo y realiza el intercambio entre valores que se encuentren dentro de un rango acotado para preservar la utilidad.

ID	Edad
1	29
2	48
3	21
4	36
5	45
6	58
7	72
8	22
9	25
10	43

ID	Edad
1	25
2	45
3	22
4	43
5	48
6	72
7	58
8	21
9	29
10	36

Métodos no perturbativos

Generalización (generalization)

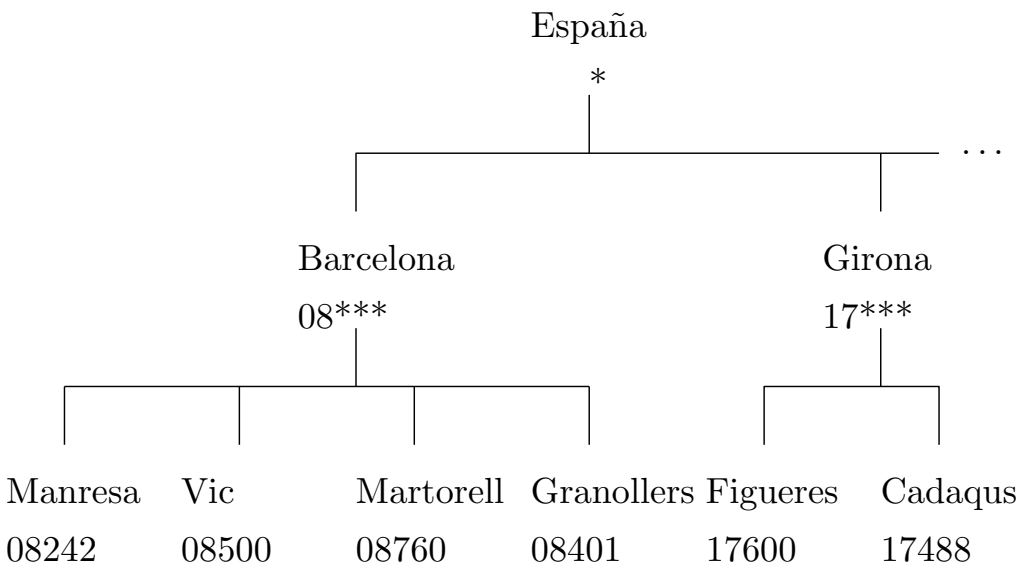
- No introducen ruido o distorsión. La información protegida continúa siendo totalmente verdadera.
- Se generalizan o suprimen algunas partes de la información.
- Dos casos básicos:
  - atributos numéricos: creación de rangos
  - atributos nominales: creación de jerarquías

ID	Edad
1	29
2	48
3	21
4	36
5	45
6	58
7	72
8	22
9	25
10	43

ID	Edad
1	[20,30)
2	[30,50)
3	[20,30)
4	[30,50)
5	[30,50)
6	[50,80)
7	[50,80)
8	[20,30)
9	[20,30)
10	[30,50)

Métodos no perturbativos

Generalización (generalization)



ID	Edad
1	08500
2	17600
3	08242
4	25128
5	17488
6	08401
7	08760
8	43840
9	43500
10	25310

ID	Edad
1	08***
2	17***
3	08***
4	25***
5	17***
6	08***
7	08***
8	43***
9	43***
10	25***

k-Anonimidad en tablas

Definición

Un conjunto de datos cumple el modelo de la k-**anonimidad** si, y sólo si, para cualquier combinación de atributos casi-identificadores existen k o más registros que comparten los mismos valores. Por lo tanto, la probabilidad de identificación de un usuario en un conjunto de datos k-anónimo con respecto a los casi-identificadores es de como máximo $\frac{1}{k}$.

- Es una condición que debe ser satisfecho por el conjunto de datos protegido.
- Generalmente, conseguimos cumplir la k-anonimidad a través de los métodos de protección o enmascaramiento que hemos visto en las secciones anteriores.

Métodos no perturbativos

Ejemplo de k-anonimidad

ID	CP	H/M	Edad	Enferm.
1	08500	H	25	Cáncer
2	17600	M	45	Hepatitis
3	08242	H	22	Gripe
4	25128	H	43	Cáncer
5	17488	M	48	Diabetes
6	08401	M	72	Gripe
7	08760	M	58	Hepatitis
8	43840	M	21	Cáncer
9	43500	H	29	Diabetes
10	25310	M	36	Gripe

Cuadro 3: Tabla original

ID	CP	H/M	Edad	Enferm.
1	08***	H	25	Cáncer
2	17***	M	46,5	Hepatitis
3	08***	H	25	Gripe
4	25***	*	39,5	Cáncer
5	17***	M	46,5	Diabetes
6	08***	M	65	Gripe
7	08***	M	65	Hepatitis
8	43***	*	23,5	Cáncer
9	43***	*	23,5	Diabetes
10	25***	*	39,5	Gripe

Cuadro 4: Tabla k-anónima, con $k = 2$

k-Anonimidad en tablas

Notas importantes:

- Para el enmascaramiento de los datos hemos aplicado generalización en el "código postal", supresión en el "género" y micro-agregación univariante en la "edad".
- La tabla puede ser publicada con la certeza de que un atacante sólo podrá identificar a un usuario con una probabilidad de, como máximo, $\frac{1}{2}$.
- Meyerson y Williams[5] demostraron que la obtención de datos k-anónimos óptimos para un conjunto multidimensional de casi-identificadores es *NP-Hard*.

[5] A. Meyerson and R. Williams. "On the complexity of optimal k-anonymity". In Proceedings of the 23 ACM SIGMOD-SIGACTSIGART Symposium on Principles of Database Systems, pp. 223–228, New York, NY, USA, 2004. ACM.

Las redes y los grafos

Tipos de redes o grafos básicos

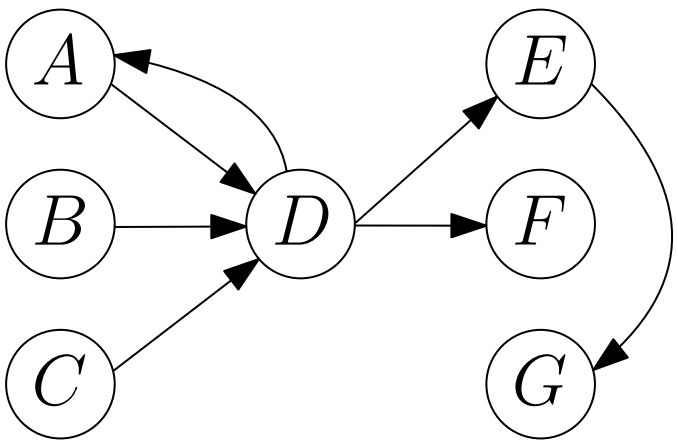


Figura 1: Grafo dirigido o asimétrico

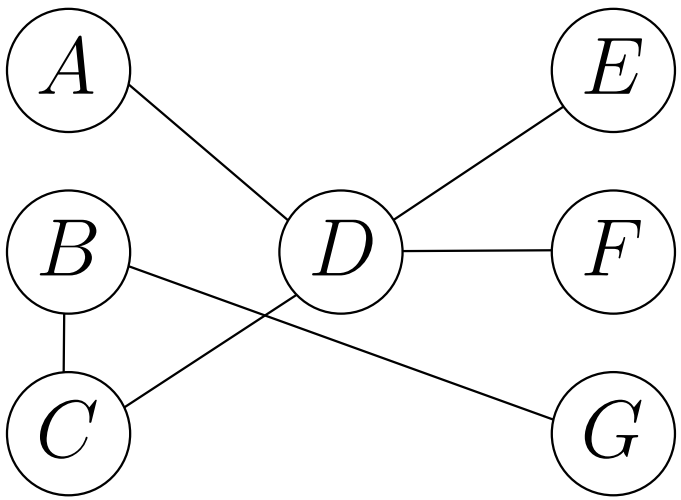


Figura 2: Grafo no dirigido o simétrico

Definición del problema

Anonimización simple (naïve anonimization)

- Re-identificación de un nodo a partir del subgrafo a distancia 1.

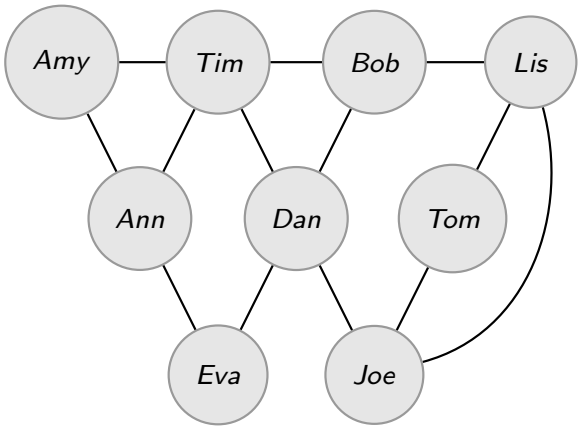


Figura 3: G

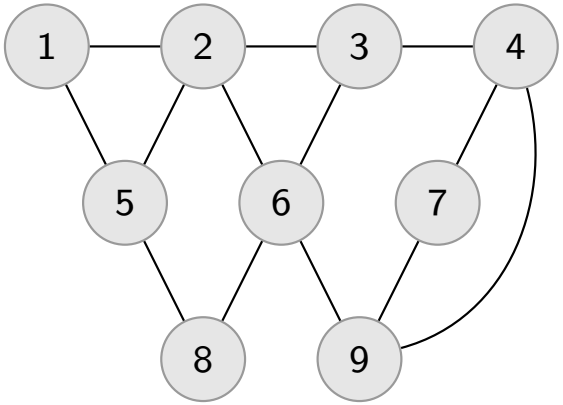


Figura 4: $\widetilde{G}$

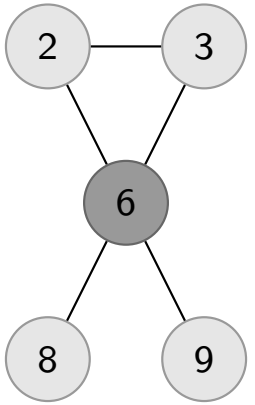


Figura 5: $\widetilde{G}_{Dan}$

Amenazas a la privacidad

Categorías principales de amenazas a la privacidad:

1. La **divulgación de la identidad** (*Identity disclosure*) ocurre cuando se revela la identidad de un individuo asociado con un vértice del grafo anónimo.
2. La **divulgación de los atributos** (*Attribute disclosure*) no busca identificar necesariamente un vértice, sino revelar atributos o datos sensibles del vértice. Los datos sensibles asociados a cada vértice se ven comprometidos.
3. La **divulgación de las relaciones** (*Link disclosure*) ocurre cuando se revela la relación sensible entre dos individuos.

Métodos de anonimización

Familias de técnicas de anonimización en grafos:

- **Modificación de aristas y vértices**: Estas técnicas transforman el grafo mediante modificaciones de aristas o vértices (añadiendo y/o eliminado) y luego publican los datos perturbados. Los datos se ponen así a disposición para cualquier tipo de análisis, sin restricciones.
- **Grafos inciertos** (*Uncertain graphs*): Este enfoque está basados en la adición o eliminación de aristas de forma "parcial", asignando una probabilidad de existir a cada arista de la red anónima. En lugar de crear o eliminar aristas, se considera el conjunto de todas las aristas posibles y se asigna una probabilidad de existir a cada una de ellas.
- **Métodos de generalización** (*Generalization*): Estos métodos buscan vértices similares y los agrupan en particiones, de forma que los detalles sobre los individuos quedan ocultos.

Modificación de aristas y vértices

Aproximaciones:

- **Métodos aleatorios**: se basan en la introducción de ruido aleatorio en los datos originales. Protegen contra la re-identificación de una manera probabilística.
- k-**Anonimidad y derivados**: modificación de aristas y vértices tiene como objetivo cumplir con determinadas restricciones de privacidad.

Estrategias básicas de modificación de aristas:

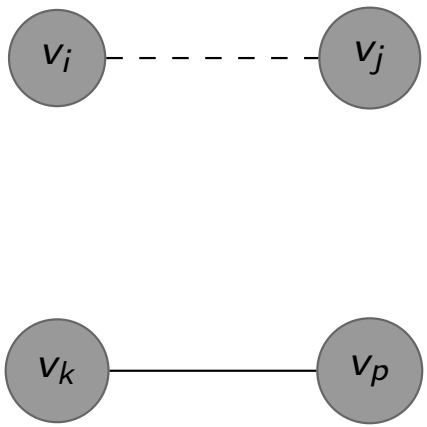


Figura 6: Edge add/del

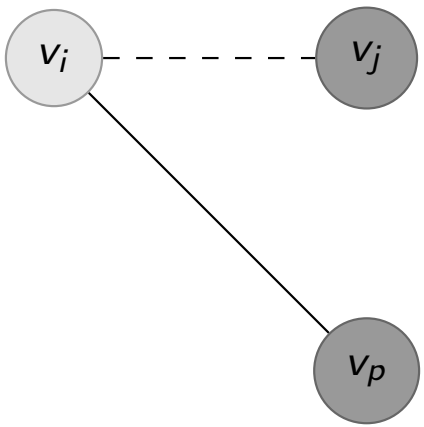


Figura 7: Edge rotation

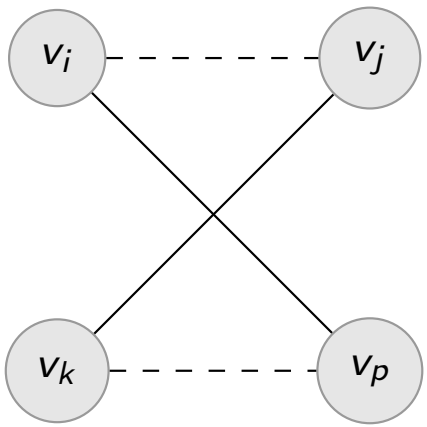


Figura 8: Edge switch

Modificación de aristas y vértices

Métodos aleatorios

- Introducción de ruido aleatorio en los datos originales.
- Protegen contra la re-identificación de una manera probabilística.

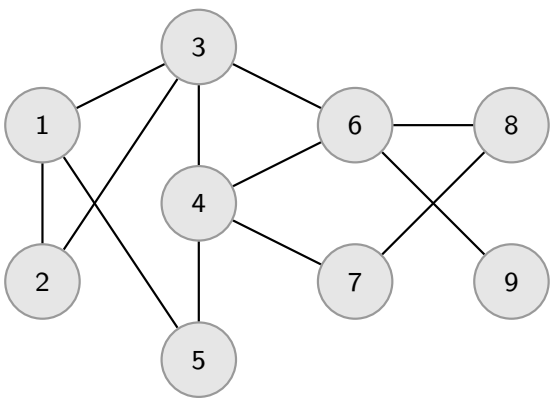


Figura 9: Original

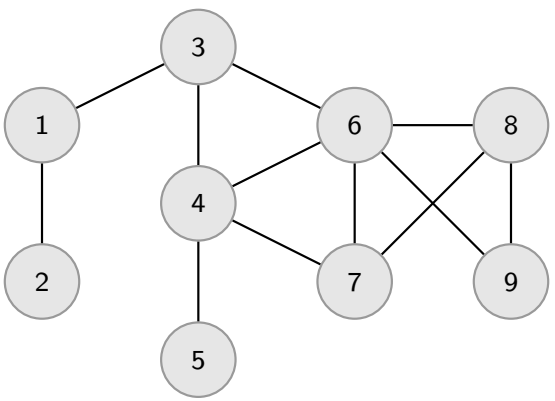


Figura 10: Add/del

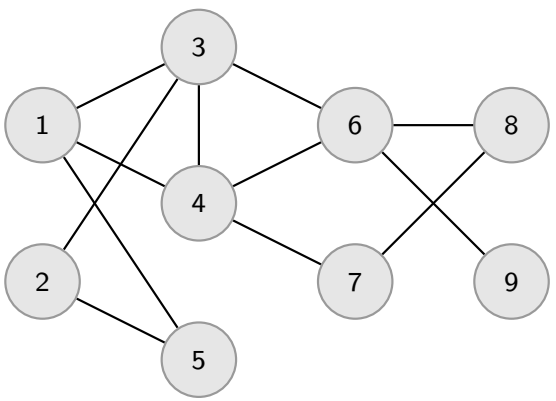


Figura 11: Switch

Modificación de aristas y vértices

k-Anonimidad y derivados

- Realizar las mínimas modificaciones que permitan cumplir con las restricciones de privacidad deseadas.
- k-anonimidad basada en el grado. Ejemplo redes 2-anónimas basadas en el grado a partir de modificaciones en las aristas y los vértices, respectivamente.

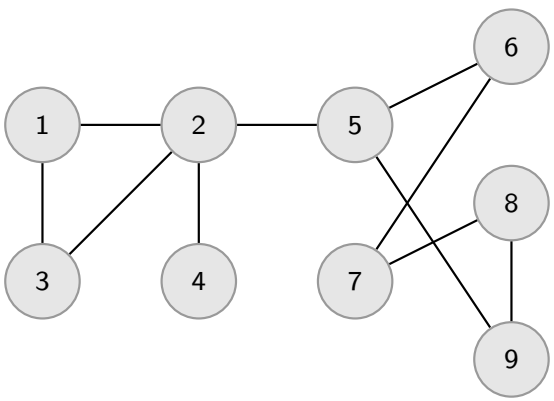


Figura 12: Original

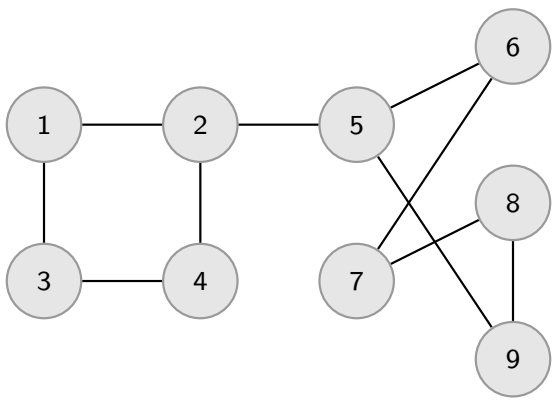


Figura 13: Modificación aristas

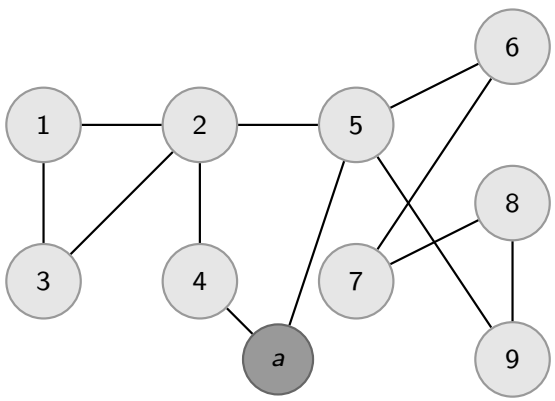


Figura 14: Vértices falsos

Grafos inciertos

Propiedades de los grafos inciertos

- $G = (V, p)$, donde $p : V_2 \to [0, 1]$ es una función que asigna las probabilidades existentes a todos las aristas posibles.
- Todas las aristas $\binom{n}{2}$ existen con una cierta probabilidad en el rango $[0, 1]$.

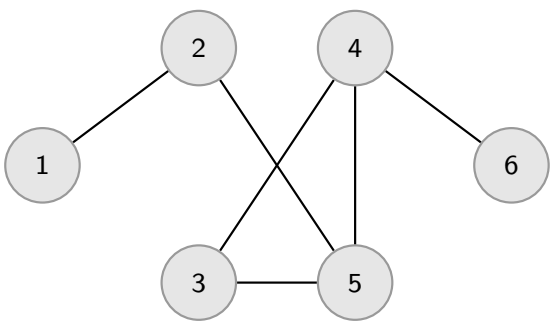


Figura 15: Original

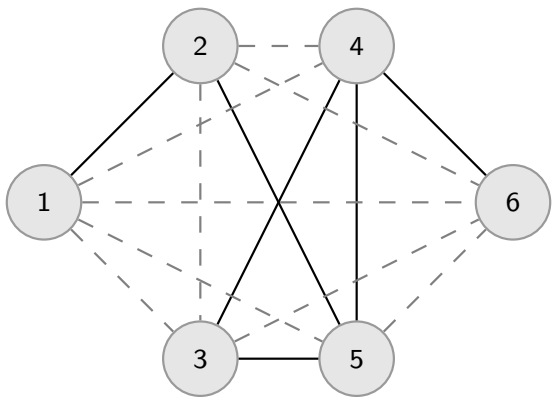


Figura 16: Grafo incierto

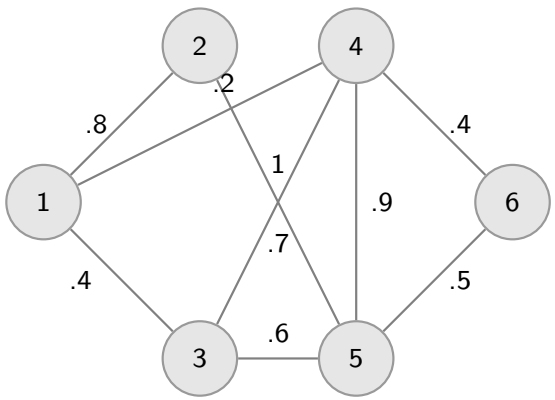


Figura 17: Grafo incierto anónimo

Métodos de generalización

Propiedades de la generalización

- Agrupar vértices y aristas en particiones llamadas *super-vértices* y *super-aristas*.
- El grafo generalizado contiene las estructuras de enlace entre las particiones, así como la descripción agregada de cada partición. No tiene la misma granularidad y escala que el grafo original.

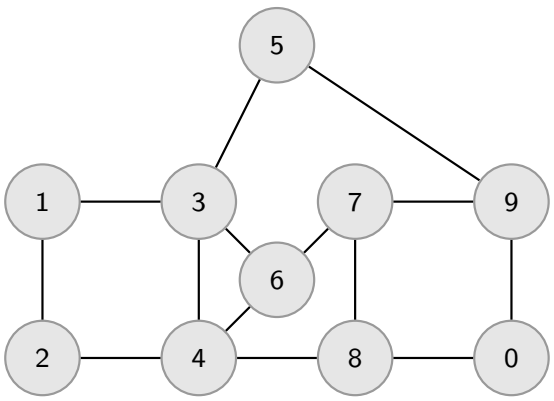


Figura 18: Original

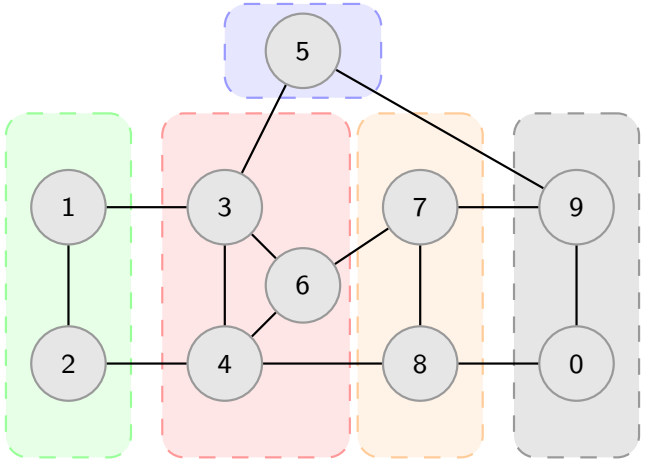


Figura 19: Conjunto particiones

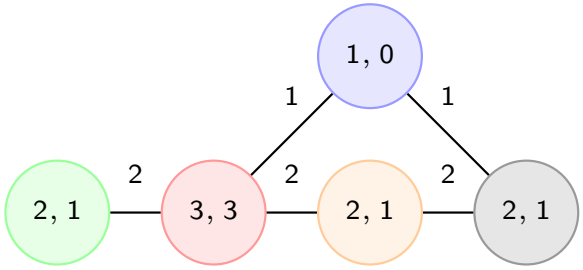


Figura 20: Generalización

Conclusiones

Conclusiones

1. La publicación y compartición de datos favorece el conocimiento global, p.ej. *open data*.
2. La preservación de la privacidad es un campo de investigación muy activo.
3. Cada tipo de datos requiere de métodos específicos, que pueda lidiar con las características de los datos.
4. El balance entre privacidad y utilidad de los datos es la clave para un buen método de anonimización.
5. El auge de las redes sociales, así como el *big data* están empujando este campo para crear nuevos modelos y técnicas que apliquen a las nuevas realidades:
   - Datos estructurados (tablas, registros de búsquedas, logs, etc)
   - Datos semi-estructurados (redes, json, xml, etc)
   - Datos no estructurados (documentos, imágenes, etc)

Final

Preguntas?

Jordi Casas-Roma
Universitat Oberta de Catalunya **UOC**
jcasasr@uoc.edu