Ús secundari de dades massives i privadesa

Josep Domingo-Ferrer

Universitat Rovira i Virgili

Càtedra de Privades@ de dades

josep.domingo@urv.cat

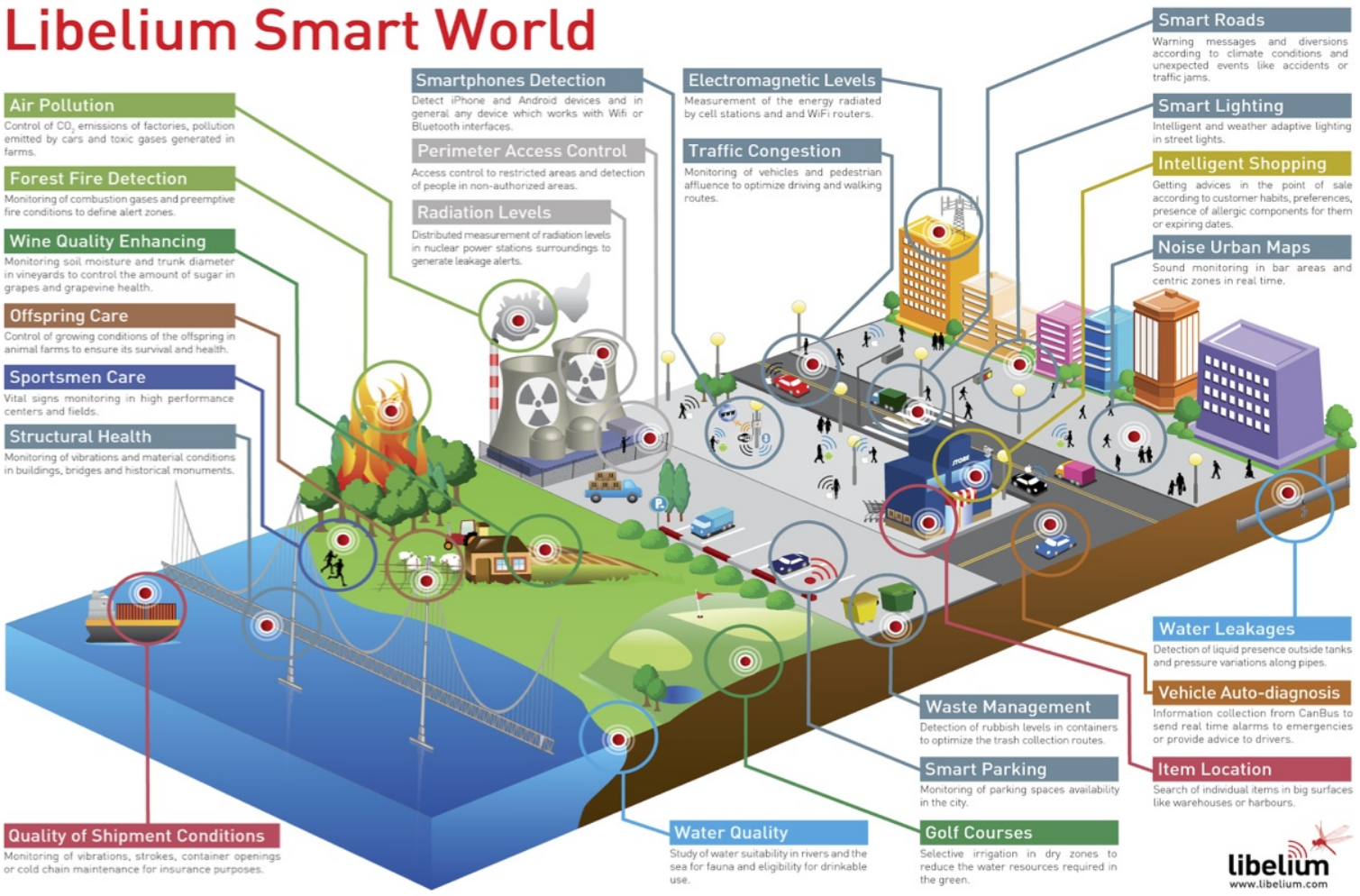
Barcelona, 9 de juny del 2017

Introducció

- Les megadades o dades massives han esdevingut una realitat amb el canvi de mil·leni.
- Qualsevol activitat humana deixa un rastre digital que algú recull i emmagatzema:
  - Sensors de la Internet de les Coses (IdC)
  - Aplicacions socials
  - Comunicació màquina-màquina
  - Vídeo mòbil, etc.

Exemple: megadades d'Internet en 1 minut

Exemple: megadades de ciutats intel·ligents i IdC



Trets distintius de les megadades

Volum El volum de dades de l'univers digital va a arribar a 9.500 milions de petaoctets (9.5×10^{24} octets) el 2015, amb un increment de 3.000 milions de petaoctets sobre el 2014 (Meeker 2016).

Velocitat Hi ha un nombre creixent de fluxos continus de dades provinents de sensors o de xarxes socials. Hom pot capturar dades en línia de milions d'esdeveniments per segon.

Varietat Les dades vénen de moltes fonts i en formats diferents (numèric, categòric, text no estructurat, àudio, vídeo, etc.).

Les megadades amenacen la privadesa

- Tot i que les megadades són un recurs valuosíssim en molts de camps, amenacen com més va més la privadesa dels subjectes les dades dels quals es recullen (sovint sense que ho sàpiguen).
- P.e. el model de predicció d'una cadena de supermercats va endevinar l'embaràs d'una adolescent abans no ho sabessin els seus pares (Duhigg 2012).

Control de revelació estadística

- Estadístics i informàtics s'han preocupat pel risc de revelació molt abans de l'aparició de les megadades.
- El secret estadístic o control de la revelació estadística (CRE, Hundepool *et al.* (2012)) cerca de permetre inferències útils sobre les dades publicades tot i preservant la privadesa dels subjectes als quals corresponen els registres.
- Hi ha tècniques per limitar el risc de revelació en microdades (fitxers dels quals els registres corresponen a subjectes individuals), en taules i en bases de dades en línia.
- La versió modificada per limitar-ne el risc s'anomena versió anonimitzada.

Utilitat o privadesa primer?

- L'anonimització centrada en la utilitat (canviant paràmetres iterativament fins que el risc de revelació sigui prou baix) és lenta i no té garanties formals de privadesa. És l'enfocament habitual en estadística oficial.
- L'anonimització centrada en la privadesa (basada a imposar un model de privadesa, com el k-anonimat, la t-proximitat o la privadesa ε-diferencial) sovint dóna poca utilitat o aparellabilitat.

Megadades, llei i ètica

- Les megadades s'obtenen recollint totes les dades possibles per extreure'n coneixement, possiblement amb mètodes innovadors.
- Això xoca amb la privadesa dels individus, especialment perquè el subjecte (consumidor, ciutadà) no sap que es recullen les seves dades.
- El proveïdor de serveis obté les dades com a resultat d'una transacció (p.e. compra en línia), com a retorn d'un servei gratuït (p.e. correu electrònic o xarxes socials) o com a requisit natural d'algun servei (situació si es fa servir GPS).

Protecció de dades personals en el dret de la UE

Les dades personals, més concretament la informació identificable personalment (IIP), són qualsevol informació relacionada amb una persona natural identificada or identificable.
Principis aplicables a la IIP abans de les megadades (grup de treball en protecció de dades de l'Art. 29, nou Reglament General de Protecció de Dades, cf. D'Acquisto *et al.* (2015)):

- **Legalitat** (consentiment obtingut o processament necessari per a: contracte o obligació legal o interessos vitals del subjecte o interès públic o interessos legítims del processador compatibles amb els drets del subjecte)
- **Consentiment** (senzill, específic, informat i explícit)
- **Limitació de propòsit** (legítim i especificat abans de la recollida)

Protecció de dades personals en el dret de la UE (II)

- **Necessitat i minimització de dades** (recollir només el que cal i guardar-ho només mentre calgui)
- **Transparència i obertura** (els subjectes han de rebre informació entenedora sobre la recollida i el processament)
- **Drets individuals** (accés, esmena, esborrament/oblit)
- **Seguretat de la informació** (cal protegir les dades recollides d'accés, processament o manipulació no autoritzats, i de la pèrdua o destrucció)
- **Responsabilitat** (qui recull i processa ha de poder demostrar que compleix els principis anteriors).
- **Protecció de dades per disseny i per defecte** (de bon principi, no com un afegit)

Conflicte entre principis i megadades personals

- Les megadades s'obtenen de recollir i aparellar dades de diverses fonts, sovint contínuament.
- Llevat que hom anonimitzi les dades personals, hi ha conflictes amb els principis:
  - **Limitació de propòsit**. Sovint es fan usos secundaris de les megadades que no es preveien en recollir-les.
  - **Consentiment**. Si el propòsit no és clar, no es pot obtenir consentiment.
  - **Legalitat**. Sense limitació de propòsit ni consentiment, la legalitat és dubtosa.
  - **Necessitat i minimització de dades**. Hom obté les megadades precisament acumulant dades per a usos potencials.
  - **Drets individuals**. Els individus no saben ni tan sols quines dades sobre ells es guarden.
  - **Responsabilitat**. Si no hi ha compliment de principis, no es pot demostrar.

Amenaces vinculades a la recollida de dades

- *Violació de dades*. Com més dades es recullen, més atractives són per a un atacant (avui Yahoo ha denunciat el robatori de mil milions de comptes).
- *Mal ús per part d'empleats* (Chen 2010). Els empleats de qui recull, emmagatzema o processa les dades en poden fer mal ús.
- *Ús secundari no desitjat*. P.e., les dades d'algú contrari als anticonceptius poden fer-se servir per elaborar-ne de nous.
- *Canvis en les pràctiques empresarials*. El compromís de privadesa d'un recol·lector pot canviar (p.e. Whatsapp ha decidit recentment i unilateral de compartir amb Facebook els números de mòbil dels seus usuaris).
- *Accés del govern sense les garanties legals degudes* (Solove 2011). P.e. la NSA dels EUA ha accedit a les dades dels usuaris de les grans empreses d'Internet.

L'anonimització com a solució

- L'anonimització és una possible solució per superar el conflicte entre megadades i privadesa.
- Com que els principis legals es refereixen a IIP, un cop anonimitzades les dades, hom pot pensar que ja no són IIP i que no els cal protecció.
- Tanmateix, l'anonimització de megadades és plena de reptes...

Reptes en anonimització de megadades

- Massa poca anonimització, p.e. només suprimir els identificadors directes, pot ser insuficient per impedir la reidentificació dels subjectes (Barbaro i Zeller 2006);
- Això és especialment problemàtic amb megadades, el volum i la varietat de les quals faciliten la reidentificació.
- Massa anonimització pot impedir d'aparellar dades sobre el mateix subjecte (o subjectes semblants) provinents de diferents fonts, cosa que entorpeix la construcció de megadades.

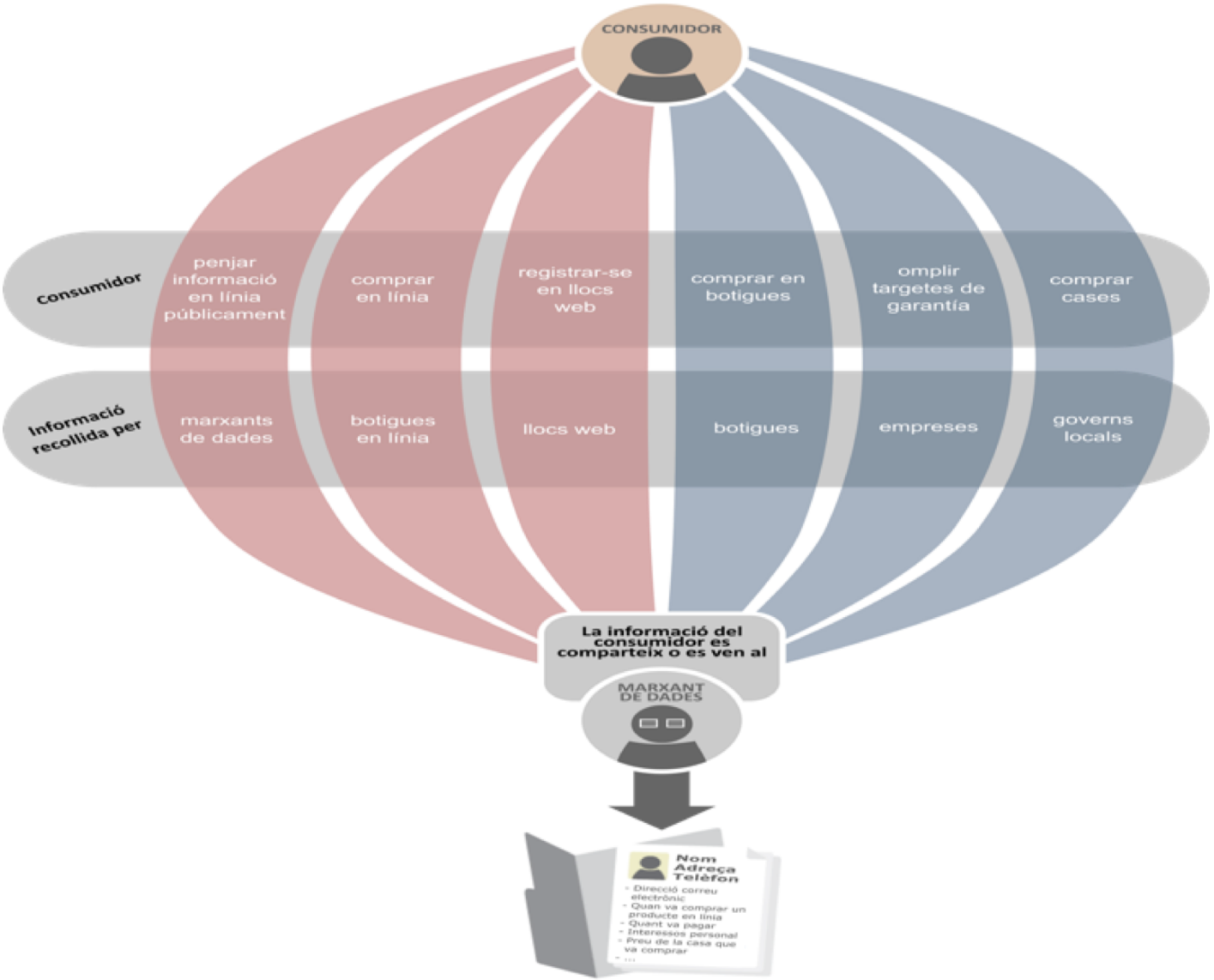
Els nihilistes: cal sacrificar la privadesa

- Sacrificar la privadesa a la seguretat. Governs (lluita antiterrorista. Empreses (identificació biomètrica d'empleats o clients, que envaeix la privadesa sense garantir necessàriament més seguretat).
- Sacrificar la privadesa a la funcionalitat. Aplicacions gratuïtes per a la web i els telèfons mòbils (motors de cerca; Google Calendar, Streetview, etc.).
- Sacrificar la privadesa a la funcionalitat i a la seguretat. Les empreses d'Internet poden filtrar als governs les dades que recullen amb aplicacions gratuïtes (Snowden sobre la NSA).

Els nihilistes pragmàtics: els marxants de dades (*data brokers*)

- No donen arguments: recullen totes les dades personals que poden (web, xarxes socials, etc.) o les compren (programes de fidelitat, comerç electrònic, etc.).
- Empaqueten tota la information corresponent a la mateixa persona per obtenir fitxes personals.
- Venen aquestes fitxes a qui les vulgui, normalment empreses de màrqueting personalitzat.
- Diversos marxants de dades operen als EUA, entre els quals Acxiom acumula dades sobre més de 700 milions de persones a tot el món (FTC 2014).
- Els marxants de dades amenacen la privadesa més que les empreses d'Internet, perquè són desconeguts del públic.

Activitat dels marxants de dades



Els nihilistes extrems

Stephen Brobst (director de tecnologia de Teradata)

- "Aspirar a tenir privadesa en la societat de les megadades és delirant."
- "El que poden demanar els subjectes és que els recol·lectors no facin un mal ús de les seves dades" (cosa que no podran verificar).

Moltes grans empreses contracten Teradata per analitzar la informació que acumulen sobre els seus clients i les seves transaccions, per fer-los ofertes personalitzades.

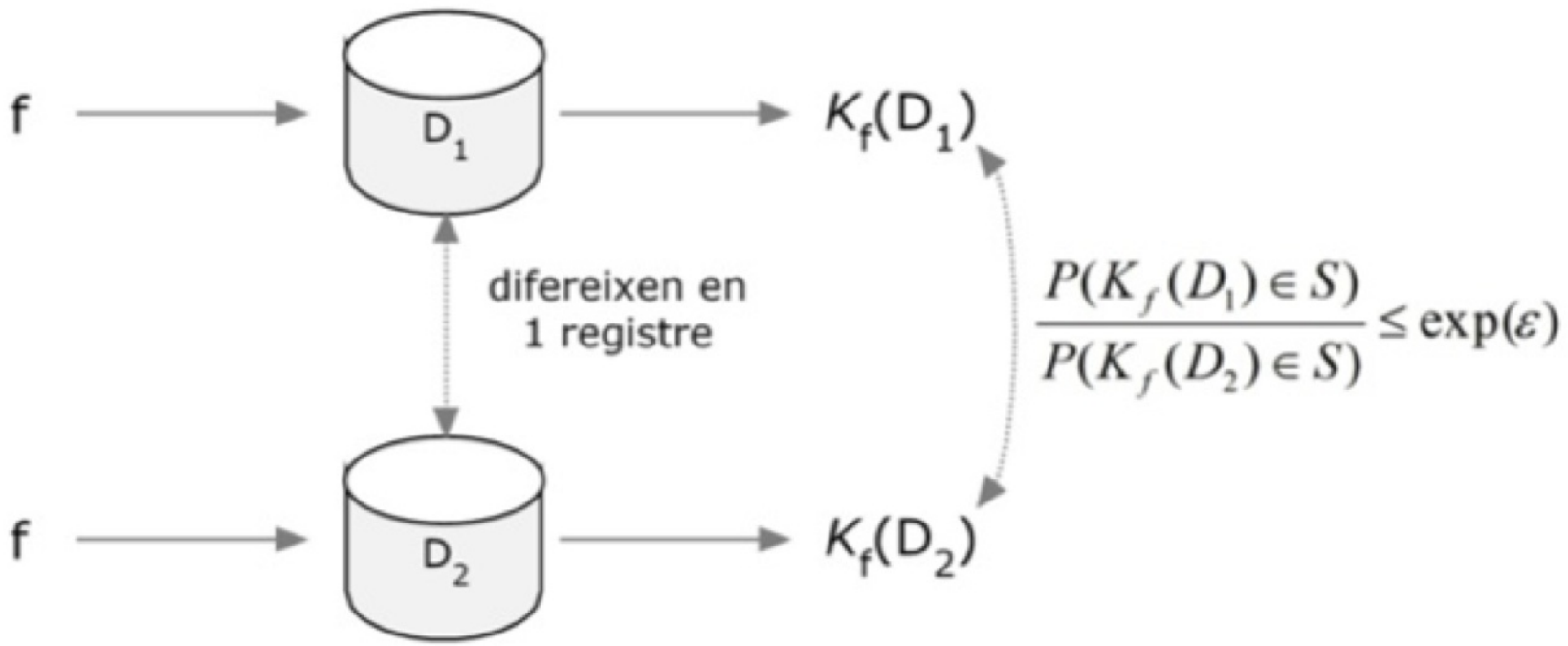
Els fonamentalistes: privadesa a costa d'inutilitzar les dades

- El control de la revelació estadística (CRE) va iniciar-se amb Dalenius (1977), un estadístic oficial. Cf. l'estat de la tècnica a Hundepool *et al.* (2012).
- Més tard, els informàtics van encunyar el terme "mineria de dades amb preservació de privadesa" (Agrawal i Srikant 2000), paral·lel al CRE de l'estadística oficial.
- Els informàtics també van inventar els models de privadesa, que especifiquen garanties *ex ante* de privadesa que es poden parametritzar.
- Els models de privadesa s'imposen fent servir un (o diversos) mètodes d'anonimització.
- Si llurs paràmetres són massa estrictes, poden inutilitzar les dades.

Models de privadesa: k-anonimat ($k = 2$)

Identificadors				Quasi-identificador					Confidencial
DNI	*Nom*	*Cognom 1*	*Cognom 2*	*Sexe*	*Codi postal*	*Edat*	*Data admissió*	*Data d'alta*	*Diagnòstic*
-	-	-	-	F	BCN	57	23/05/15	29/05/15	070.54
-	-	-	-	F	BCN	57	23/05/15	29/05/15	311.00
-	-	-	-	F	TGN	79	29/05/15	05/06/15	070.54
-	-	-	-	F	TGN	79	29/05/15	05/06/15	401.90
-	-	-	-	M	TGN	57	18/03/15	30/03/15	305.10
-	-	-	-	M	TGN	57	18/03/15	30/03/15	592.10

Models de privadesa: privadesa ε-diferencial



Privadesa diferencial i fonamentalisme

- Si l'absència, presència, o modificació de qualsevol registre no s'ha de notar en les respostes a les consultes, cal afegir-los molt de soroll $\Longrightarrow$ llur utilitat és dubtosa.
- P.e., si la presència o absència de l'únic milionari d'un poble no s'ha de notar quan es consulta la renda màxima del poble, el màxim que retorni la consulta no pot ésser real.

El camí de mig: desiderata en anonimització de megadades

- Les megadades anonimitzades que es publiquen haurien de donar resultats semblants als que s'obtindrien amb les dades originals per a un bon grapat d'anàlisis exploratòries.
- No haurien de permetre la reconstrucció inequívoca del perfil de cap subjecte.
- A banda de preservar la utilitat analítica, un model de privadesa per a megadades hauria de satisfer (Soria-Comas i Domingo-Ferrer 2015):
  - Componibilitat;
  - Cost computacional (quasi-)lineal;
  - Aparellabilitat.

Componibilitat

- Un model de privadesa és componible si la seva garantia de privadesa es manté (potser amb limitacions) després d'aplicar-lo repetidament.
- És a dir, un model *no* és componible si ajuntar fitxers publicats independentment, cadascun dels quals satisfà el model, pot violar el model.
- La componibilitat pot avaluar-se entre fitxers anonimitzats que satisfan el mateix model de privadesa o models diferents; i també entre un fitxer anonimitzat i un de no anonimitzat (cas més exigent).
- La componibilitat cal per afrontar la velocitat i la varietat de les megadades.

Cost computacional (quasi-)lineal

- El cost baix cal per afrontar el volum de les megadades.
- Normalment, un model de privadesa es pot satisfer amb diversos mètodes de CRE.
- El cost computacional depèn del mètode triat.
- El cost desitjable és $O(n)$ o com a molt $O(n \log n)$, per a un fitxer de n registres.
- Per a mètodes amb cost més alt, pot ser útil el blocatge, tot i que pot perjudicar la utilitat i/o la privadesa de les dades resultants.

Aparellabilitat

- En megadades, la informació sobre un subjecte concret es recull de diverses fonts (varietat de les megadades).
- Per tant, la capacitat d'aparellar registres que corresponen al mateix individu o a individus semblants és crítica
  $\Longrightarrow$ anonimitzar les dades a la font hauria de preservar l'aparellabilitat fins a cert punt.
- Però... aparellar registres corresponents al mateix subjecte en redueix la privadesa
  $\Longrightarrow$ la precisió dels aparellaments hauria de ser més petita amb fitxers anonimitzats que amb fitxers originals.

Protecció de megadades amb k-anonimat

- En megadades, costa de determinar el conjunt dels atributs quasi-identificadors (QI), que poden fer-se servir per aparellar amb fitxers identificats externs.
- L'opció més segura és considerar que tots els atributs són QI.

Resum de k-anonimat per a megadades

- Perquè el k-anonimat sigui componible, els controladors que comparteixen subjectes s'han de coordinar o han de seguir estratègies adequades.
- Hi ha heurístiques quasi-lineals per al k-anonimat.
- L'aparellabilitat és possible al nivell de classe k-anònima.
- Amb un cert esforç de coordinació, el k-anonimat és una opció raonable per anonimitzar megadades.

Protecció de megadades amb privadesa diferencial

- La privadesa ε-diferencial (PD) ofereix garanties fortes de privadesa.
- Com més petit és ε, més privadesa.
- La PD es pot assolir afegint soroll o generant dades sintètiques a partir d'un model privat diferencialment (p.e. un histograma).
- Un fitxer sintètic pot ser-ho parcialment o totalment.
- En la síntesi parcial, només se substitueixen per dades sintètiques els valors que es consideren massa sensibles.

Resum de PD per a megadades

- La PD té bones propietats de componibilitat, que poden anar bé per anonimitzar dades dinàmiques.
- La PD també té un cost computacional baix, que pot anar bé per a fitxers molt grans.
- L'aparellament de fitxers PD només és possible si els fitxers comparteixen atributs que no s'han alterat.
- El problema més gros de la PD és que la utilitat de les dades privades ε-diferencialment per a anàlisi exploratòria és pràcticament zero per als valors de ε que donen garanties significatives de privadesa (típicament, $\varepsilon \leq 1$).

Conclusions

- Hi ha un debat de si les dades massives són compatibles amb la privadesa dels ciutadans.
- Hi ha dues posicions extremes: el nihilisme i el fonamentalisme.
- Hem explorat un camí del mig entre aquests extrems.
- Hem formulat algunes propietats desitjables dels models de privadesa per a megadades (componibilitat, cost computacional baix i aparellabilitat).
- Hem examinat fins a quin punt els dos models de privadesa principals (k-anonimat i privadesa ε-diferencial) satisfan aquestes propietats.

Conclusions (II)

- El k-anonimat satisfà els requisits amb penes i treballs:
  - La componibilitat exigeix coordinació entre els diversos protectors de dades;
  - La complexitat pot arribar a ser quasi-lineal;
  - L'aparellabilitat és al nivell de classe anònima, però no de registre.
- La privadesa diferencial satisfà millor les propietats, però l'argument que la descarta per a megadades és que en destrueix la utilitat per a anàlisis exploratòries.

Línies de recerca obertes

- Hi ha molta feina a fer per fressar aquest camí del mig.
- Calen models de privadesa que ofereixin componibilitat, complexitat baixa, aparellabilitat i preservació d'utilitat per a anàlisis exploratòries.
- La varietat de les megadades va més enllà dels fitxers formats per registres: inclou vídeo, àudio, text no estructurat, etc., l'anonimització dels quals és un gran repte.
- El models de privadesa i els mètodes CRE han de poder afrontar la velocitat i el volum: anonimitzar megadades dinàmiques o en flux és un territori pràcticament inexplorat.

Referències I

S. Agrawal i J.R. Haritsa (2005) A framework for high-accuracy privacy-preserving data mining, a *ICDE'05*, IEEE, p. 193-204.

R. Agrawal i R. Srikant (2000) Privacy-preserving data mining, a *ACM SIGMOD'00*, p. 439-450.

M. Barbaro i T. Zeller (2006) A face is exposed for AOL searcher no. 4417749, *New York Times*.

A. Chen (2010) Gcreep: Google engineer stalked teens, spied on chats, *Gawker*.

G. Cormode, C. Procopiuc, D. Srivastava, E. Shen i T. Yu (2012) Differentially private spatial decompositions, a *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering-ICDE'12*, Washington, DC, EUA, p. 20-31. IEEE Computer Society.

Referències II

G. D'Acquisto, J. Domingo-Ferrer, P. Kikiras, V. Torra, Y.-A. de Montjoye i A. Bourka (2015) *Privacy by Design in Big Data — An overview of privacy enhancing technologies in the era of big data analytics*, European Union Agency for Network and Information Security (ENISA).

T. Dalenius (1977) Towards a methodology for statistical disclosure control. *Statistik Tidskrift* 15:429-444.

J. Domingo-Ferrer i K. Muralidhar (2016) New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users, *Information Sciences* 337-338:11-24.

J. Domingo-Ferrer, D. Sánchez i J. Soria-Comas (2016) Co-utility: self-enforcing collaborative protocols with mutual help, *Progress in Artificial Intelligence* 5(2):105-110.

Referències III

J. Domingo-Ferrer i J. Soria-Comas (2016) Anonymization in the time of big data, a *Privacy in Statistical Databases-PSD 2016*, Springer, p. 225-236.

J. Domingo-Ferrer i V. Torra (2005) Ordinal, continuous and heterogeneous k-anonymity through microaggregation, *Data Mining and Knowledge Discovery* 11(2):195-212.

C. Duhigg (2012) How companies learn your secrets, *New York Times Magazine*, Feb. 16.

C. Dwork (2006) Differential privacy, a *ICALP'06*, LNCS 4052, Springer, p. 1-12.

C. Dwork i G. N. Rothblum (2016) Concentrated differential privacy (v2), March 16, arXiv:1603.01887v2.

M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page i T. Ristenpart (2014) Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing, a *Proc. of the 23rd USENIX Security Symposium*, San Diego CA, EUA, p. 17-32.

Referències IV

FTC (2014) *Data Brokers: A Call for Transparency and Accountability*, US Federal Trade Commission.

A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte-Nordholt, K. Spicer i P.-P. de Wolf (2012) *Statistical Disclosure Control*, Wiley.

N. Li, T. Li i S. Venkatasubramanian (2007) t-Closeness: privacy beyond k-anonymity and l-diversity, a *ICDE'07*, p. 106-115.

A. Machanavajjhala i D. Kiefer (2015) Designing statistical privacy for your data, *Communications of the ACM* 58(3):58-67.

A. Machanavajjhala, D. Kifer, J. Gehrke i M. Venkitasubramaniam (2007) l-Diversity: privacy beyond k-anonymity, *ACM Trans. Knowl. Discov. Data* 1(1):3.

Referències V

A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke i L. Vilhuber (2008) Privacy: theory meets practice on the map, a *Proceedings of the 2008 IEEE 24th Intl. Conf. on Data Engineering-ICDE'08*, Washington, DC, USA. IEEE Computer Society, p. 277–286.

M. Meeker (2016) *2016 Internet Trends* `http://www.kpcb.com/blog/2016-internet-trends-report`

P. Mohan, A. Thakurta, E. Shi, D. Song i D. E. Culler (2012) GUPT: privacy preserving data analysis made easy, a *Proc. of ACM SIGMOD'12*, Scottsdale AZ.

P. Samarati i L. Sweeney (1998) *Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression*, Technical Report, SRI International.

Referències VI

D. Sánchez, J. Domingo-Ferrer i S. Martínez (2014) Improving the utility of differential privacy via univariate microaggregation, a *Privacy in Statistical Databases-PSD 2014*, p. 130-142. Springer.

D. J. Solove (2011) *Nothing to Hide: the False Tradeoff Between Privacy and Security*, New York: Yale University Press.

C. Song i T. Ge (2014) Aroma: a new data protection method with differential privacy and accurate query answering, a *CIKM'14*, ACM, p. 1569-1578.

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

Referències VII

J. Soria-Comas i J. Domingo-Ferrer (2015) Big data privacy: challenges to privacy principles and models, *Data Science and Engineering* 1(1):21-28.

J. Soria-Comas i J. Domingo-Ferrer (2015b) Co-utile collaborative anonymization of microdata, a *MDAI 2015*, LNCS 9321, Springer, p. 192-2016.

S. L. Warner (1965) Randomized response: a survey technique for eliminating evasive answer bias, *J. Am. Stat. Assoc.* 60:63-69.

X. Xiao i Y. Tao (2007) M-Invariance: towards privacy-preserving re-publication of dynamic datasets, a *SIGMOD'07*, ACM, p. 689-700.

J. Xu, Z. Zhang, X. Xiao, Y. Yang i G. Yu (2012) Differentially private histogram publication, a *Proceedings of the 2012 IEEE 28th Intl. Conf. on Data Engineering-ICDE'12*, Washington, DC, USA. IEEE Computer Society, p. 32–43.

Referències VIII

J. Zhang, G. Cormode, C.M. Procopiuc, D. Srivastava i X. Xiao (2014) Privbayes: private data release via Bayesian networks, a *Proceedings of the 2014 ACM SIGMOD Intl. Conf. on Management of Data, SIGMOD'14*, New York, NY, USA. ACM, p. 1423–1434.