Données massives et modèles de vie privée

Josep Domingo-Ferrer

Universitat Rovira i Virgili, Tarragona
CYBERCAT-Center for Cybersecurity Research of Catalonia

josep.domingo@urv.cat

Paris, le 30 mai 2018

Universitat Rovira i Virgili

Introduction

- Big data have come true with the new millennium.
- Any human activity leaves a digital track that someone collects and stores:
  - Sensors of the Internet of Things
  - Social media
  - Machine-to-machine communication
  - Mobile video, etc.

Desiderata in big data anonymization

- Anonymized big data that are published should yield results similar to those obtained on the original big data for a broad range of exploratory analyses.
- They should not allow unequivocal reconstruction of any subject's profile.
- A privacy model for big data should satisfy at least (Soria-Comas and Domingo-Ferrer 2015):
  - Composability
  - (Quasi-)linear computational cost
  - Linkability

Composability

- A privacy model is composable if its privacy guarantee holds (perhaps in a limited way) after repeated application.
- In other words, a privacy model is not composable if pooling independently released data sets, each of which satisfies the model separately, can lead to a violation of the model.
- Composability can be evaluated between data sets satisfying the same privacy model, different privacy models, or between an anonymized data set and a non-anonymized data set (the latter is the most demanding case).
- Composability is needed to cope with the velocity and variety features of big data.

Universitat Rovira i Virgili

(Quasi-)linear computational cost

- Low cost is needed to cope with the volume feature of big data.
- Normally, there are several SDC methods that can be used to satisfy a privacy model.
- The computational cost depends on the selected method.
- The desirable costs would be $O(n)$ or at most $O(n \log n)$, for a data set of n records.
- For methods with higher cost, blocking can be used, but it can damage the utility and/or privacy of the resulting data.

Linkability

- In big data, the information on a particular subject is collected from several sources (variety feature of big data).
- Hence, the ability to link records corresponding to the same individual or to similar individuals is critical.
- Thus, anonymizing data at the source should preserve linkability to some extent.
- But... linking records corresponding to the same subject decreases the subject's privacy
  $\Longrightarrow$ the accuracy of linkage should be lower with anonymized data sets than with original data sets.

Privacy models: k-anonymity

k-Anonymity (Samarati & Sweeney 1998)

A data set is said to satisfy k-anonymity if each combination of values of the quasi-identifier attributes in it is shared by at least k records (k-anonymous class).

$\Longrightarrow$ Usually enforced via generalization and suppression in quasi-identifiers, but also reachable via microaggregation (Domingo-Ferrer and Torra 2005)

Universitat Rovira i Virgili

Privacy models that extend k-anonymity

l-Diversity (Machanavajjhala *et al.* 2007)

A data set is said to satisfy l-diversity if, for each group of records sharing a combination of quasi-identifier attributes, there are at least l "well-represented" values for each confidential attribute.

t-Closeness (Li *et al.* 2007)

A data set is said to satisfy t-closeness if, for each group of records sharing a combination of quasi-identifier attributes, the distance between the distribution of the confidential attribute in the group and the distribution of the attribute in the whole data set is no more than a threshold t.

Big data protection under k-anonymity

- In a context of big data, it is hard to determine the subset of QI attributes (attributes that can be used by an attacker to link with external identified databases).
- The safest option is to consider that all attributes are QI attributes.

Composability of k-anonymity

- k-Anonymity was designed to protect a single data set and is not composable in principle.
- If several k-anonymous data sets have been published that share some subjects, the attacker can mount an intersection attack to discard some records in the k-anonymous classes as not corresponding to the target subject (based on the latter's confidential attributes).
- To reach composability, the controllers ought to coordinate so that, for the subjects shared by two data sets, their k-anonymous classes contain the same k subjects.
- If such coordination is infeasible, see Domingo-Ferrer and Soria-Comas (2016) for alternative strategies.

Intersection attack against k-anonymity

$R_1, \ldots, R_n \leftarrow n$ independent data releases
$P \leftarrow$ population consisting of subjects present in all $R_1, \ldots, R_n$
for each individual i **in** P **do**
    for $j = 1$ **to** n **do**
        $e_{ij} \leftarrow$equivalence class of R_j associated to i
        $s_{ij} \leftarrow$set of confidential values of e_{ij}
    end for
    $S_i \leftarrow s_{i1} \cap s_{i2} \cap \ldots \cap s_{in}$
end for
return $S_1, \ldots, S_{|P|}$

Computational cost of k-anonymity

- k-Anonymity is attained by modifying the values of QI attributes either by combining generalization and suppression (Samarati and Sweeney 1998) or via microaggregation (Domingo-Ferrer and Torra 2005).
- Optimal generalization/suppression and optimal microaggregation are NP problems.
- Using heuristics and blocking one can reach $O(n \log n)$ complexities, where n is the number of records.

Linkability of k-anonymity

- For a subject known to be in two k-anonymous data sets, we can determine and link the corresponding k-anonymous classes containing her.
- If some of the confidential attributes are shared between the data sets, the linkage accuracy improves (one can link within k-anonymous classes).

Summary on k-anonymity for big data

- For k-anonymity to be composable, the controllers sharing subjects must coordinate or follow suitable strategies.
- There are quasi-linear heuristics for k-anonymity.
- Linkability is possible at least at the k-anonymous class level.
- With some coordination effort, k-anonymity is a reasonable option to anonymize big data.

Privacy models: ε-differential privacy

ε-Differential privacy (Dwork 2006)

A randomized query function F gives ε-differential privacy if, for all data sets D_1, D_2 such that one can be obtained from the other by modifying a single record (neighbor data sets), and all $S \subset Range(F)$

$$\Pr(F(D_1) \in S) \leq \exp(\varepsilon) \times \Pr(F(D_2) \in S).$$

- Usually enforced via Laplacian noise addition.
- Later extended for data set publishing (Soria-Comas *et al.* 2014; Xiao *et al.* 2007; Xu *et al.* 2012; Zhang *et al.* 2014).

Universitat Rovira i Virgili

Big data protection under differential privacy

- ε-Differential privacy (DP) offers strong privacy guarantees.
- The smaller ε, the more privacy.
- DP can be reached via noise addition or by generating synthetic data from a differentially privacy model (e.g. a histogram).
- A synthetic data set can be either partially or fully synthetic.
- In partial synthesis, only values deemed too sensitive are replaced by synthetic data.

Composability of DP: sequential composition

Sequential composition refers to a sequence of computations, each of them providing differential privacy in isolation, providing also differential privacy in sequence.

Theorem

Let $\kappa_i(D)$, for some $i \in I$, be computations over D providing ε_i-differential privacy. The sequence of computations $(\kappa_i(D))_{i \in I}$ provides $(\sum_{i \in I} \varepsilon_i)$-differential privacy.

Universitat Rovira i Virgili

Composability of DP: parallel composition

Parallel composition refers to several ε-differentially private computations each on data from a disjoint set of subjects yielding ε-differentially private output on the data from the pooled set of subjects.

Theorem

Let $\kappa_i(D_i)$, for some $i \in I$, be computations over D_i providing ε-differential privacy. If each D_i contains data on a set of subjects disjoint from the sets of subjects of D_j for all $j \neq i$, then $(\kappa_i(D_i))_{i \in I}$ provides ε-differential privacy.

Universitat Rovira i Virgili

Composability of DP for data sets

- *Sequential composition*. The release of ε_i-differentially private data sets D_i, for some $i \in I$, is $(\sum_{i \in I} \varepsilon_i)$-differentially private. That is, by accumulating differentially private data about a set of individuals, differential privacy is not broken but the level of privacy decreases.
- *Parallel composition*. The release of ε-differentially private data sets D_i refering to disjoint sets of individuals, for some $i \in I$, is ε-differentially private.

Computational cost of DP

- DP by noise addition has linear cost $O(n)$.
- It has been suggested to use other methods to attain DP with improved utility:
  - Data synthesis (Cormode *et al.* 2012; Zhang *et al.* 2014) has a higher computational complexity.
  - Microaggregation step prior to noise addition (Sánchez *et al.* 2014; Soria-Comas *et al.* 2014) has complexity $O(n^2)$ or $O(n \log n)$ depending on whether blocking is used.

Linkability of DP

- In general, there is no linkability between two DP data sets generated via noise addition or as fully synthetic data.
- Partially synthetic data sets, although they do not satisfy strict DP, allow accurate linkage.

Summary on DP for big data

- DP has good composability properties, which may be suitable to anonymize dynamic data.
- DP has also a low computational cost, which may be suitable for very large data sets.
- Linkability across differentially private data sets is only feasible if the data sets share unaltered attributes.
- The main problem with DP is that it does not provide significant utility for exploratory analyses unless the ε parameter is quite large.

Connections between privacy models

We show in Domingo-Ferrer and Soria-Comas (2018) that the following privacy models are interconnected around the principles of deniability and permutation

- Randomized response
- Post-randomization
- Differential privacy
- t-Closeness

Randomized response (RR)

Let X be an attribute containing the answer to a sensitive question. If X can take r possible values, then the randomized response Y (Greenberg *et al.* 1969) reported by the respondent instead of X is computed using

$$\mathbf{P} = \begin{pmatrix} p_{11} & \cdots & p_{1r} \\ \vdots & \vdots & \vdots \\ p_{r1} & \cdots & p_{rr} \end{pmatrix}$$

where $p_{uv} = \Pr(Y = v | X = u)$, for $u, v \in \{1, \ldots, r\}$ denotes the probability that the randomized response is v when the respondent's true attribute value is u.

Randomized response: estimates

- Let $\pi_1, \ldots, \pi_r$ be the proportions of respondents whose true values fall in each of the r categories of X.
- Let $\lambda_v = \sum_{u=1}^{r} p_{uv}\pi_u$ for $v = 1, \ldots, r$, be the probability of the reported value Y being v.
- Let $\lambda = (\lambda_1, \ldots, \lambda_r)^T$ and $\pi = (\pi_1, \ldots, \pi_r)^T$.
- Then $\lambda = \mathbf{P}^T\pi$.
- If $\hat{\lambda}$ is the vector of sample proportions corresponding to λ and $\mathbf{P}$ is nonsingular:

$$\hat{\pi} = (\mathbf{P}^T)^{-1}\hat{\lambda}.$$

Universitat Rovira i Virgili

The privacy model of randomized response: plausible deniability

The privacy guarantee RR offers to respondents are plausible deniability and secrecy:

- By the Bayes' formula:

$$\hat{p}_{vu} = \Pr(X = u | Y = v) = \frac{p_{uv}\pi_u}{\sum_{u'=1} p_{u'v}\pi_{u'}}.$$

- Given a reported $Y = v$, deniability can be measured as

$$H(X|Y = v) = -\sum_{u=1}^{r} \hat{p}_{vu} \log_2 \hat{p}_{vu}.$$

- If the probabilities within each column of $\mathbf{P}$ are identical, then $\hat{p}_{vu} = \pi_u$, for $u, v \in \{1, \ldots, r\}$, and $H(X|Y = v) = H(X)$ for any v, and thus $H(X|Y) = H(X)$ (Shannon's perfect secrecy).
- The price paid for perfect secrecy is a singular matrix $\mathbf{P}$, so no unbiased estimator $\hat{\pi}$ can be computed.

Randomized response: a local version of PRAM

- Matrix **P** looks exactly as the PRAM transition matrix.
- The main difference is that in RR randomization is done by the respondent, whereas in PRAM it is done by the data controller.
- Thus, RR is a local anonymization method *avant la lettre*: when RR was invented, the notion of anonymization did not exist, let alone local anonymization.

Randomized response and differential privacy

Wang *et al.* (2016) show that RR is ϵ-differentially private if

$$e^{\epsilon} \geq \max_{v=1,\ldots,r} \frac{\max_{u=1,\ldots,r} p_{uv}}{\min_{u=1,\ldots,r} p_{uv}}.$$

We can assert:

- If the maximum ratio between the probabilities in a column of **P** is bounded by e^{ϵ}, the influence of the real value X on the reported value Y is limited.
- When $\epsilon = 0$, in the above bound, the probabilities within each column of **P** are identical, and RR provides perfect secrecy.
- Thus, DP with strictest privacy ($\epsilon = 0$) offers perfect secrecy.

Universitat Rovira i Virgili

Explaining large ϵ in DP using deniability

- When one takes not-so-small ϵ, the intuition of DP is unclear: it is no longer tenable that the presence or absence of any single record is unnoticeable.
- The connection of DP with RR and hence with deniability helps understanding what large ϵ implies.
- E.g., if $\epsilon = 2$, in some columns of **P** the probability ratio may be as large as $e^2 = 7.389$. If $r = 2$, one might have a column with $p_{1v} = 0.7389$ and $p_{2v} = 0.1$. Thus, after reporting $Y = v$, the most likely value is $X = 1$ and there is only a small margin to deny it. Thus, clearly $\epsilon = 2$ does not seem to offer enough privacy.

Universitat Rovira i Virgili

Differential privacy and t-closeness

Given two distribution F_1 and F_2, consider the distance

$$d(F_1, F_2) = \max_{i=1,2,\cdots,t} \left\{ \frac{\Pr_{F_1}(x_i)}{\Pr_{F_2}(x_i)}, \frac{\Pr_{F_2}(x_i)}{\Pr_{F_1}(x_i)} \right\}.$$

Proposition (Domingo-Ferrer and Soria-Comas, 2015) *Let $k_I(D)$ be the function that returns the view on subject I's sensitive attributes given a data set D. If D satisfies $\exp(\varepsilon/2)$-closeness when using the above distribution distance, then $k_I(D)$ satisfies ε-differential privacy. In other words, if we restrict the domain of k_I to $\exp(\varepsilon/2)$-close data sets, then we have ε-differential privacy for k_I.*

DP and intruder's knowledge gain via t-closeness

- The previous proposition can explain DP in terms of the intruder's knowledge gain on the sensitive attribute value of a target respondent if the intruder can determine the respondent's cluster.
- E.g. take DP with $\epsilon = 2$. By the proposition, the probability weight attached to a certain value of a sensitive attribute X can grow by a factor $e \approx 2.718$ if the target individual's cluster is learnt by the intruder.

DP and intruder's knowledge gain via t-closeness (II)

- To decide whether a probability has grown too much, consider that the reported value v is the cluster identifier and probabilities $\hat{p}_{vu} = \Pr(X = u | Y = v)$, for $u = 1, \ldots, r$ are the probabilities assigned by the cluster-level distribution to the values of the sensitive attribute within the cluster.
- Determining the real X given the reported Y becomes determining the target respondent's sensitive value X given the target respondent's cluster Y.
- We can use a deniability argument to assess whether the cluster-level distribution is too inhomogeneous.

Example deniability argument to assess cluster-level distribution

- Take $\epsilon = 2$ and assume the sensitive attribute can take $r = 5$ different values, with uniform data set-level distribution (prob. $1/5$ for each value).
- A cluster-level distribution with one value having relative frequency $1/5 \times \exp(1) = 0.5436$ and the remaining four values 0.1141 satisfies $\exp(1) - closeness$.
- The cluster-level distribution makes guessing the sensitive attribute value much easier than the data set-level distribution (thus $\epsilon = 2$ does not offer enough protection).

Reverse mapping

Domingo-Ferrer and Muralidhar (2016):

Require: Original attribute $X = \{x_1, x_2, \cdots, x_n\}$
Require: Anonymized attribute $Y = \{y_1, y_2, \cdots, y_n\}$
  for $i = 1$ to n **do**
    Compute $j = \text{Rank}(y_i)$
    Set $z_i = x_{(j)}$ (where $x_{(j)}$ is the value of X of rank j)
  end for
  return $Z = \{z_1, z_2, \cdots, z_n\}$

The permutation paradigm

- The output Z is a permutation of X and has the same rank order as Y.
- Thus any anonymization procedure can be viewed as a permutation (X into Z) followed by residual noise addition (Z into Y) that does not alter ranks.

PRAM and the permutation paradigm

- PRAM does not permute attribute values in the data set, rather it permutes in the *domain* of attributes.
- Hence, PRAM should be viewed in terms of the permutation paradigm as permutation plus noise.
- Hence, RR can also be viewed as permutation, and so can DP and so can t-closeness.

Conclusions and further research

- There is a debate on whether big data are compatible with the privacy of citizens.
- We have stated the desirable properties of privacy models for big data (composability, low computation, linkability).
- We have examined how well the two main privacy models (k-anonymity and ε-differential privacy) satisfy those properties.
- None of them is entirely satisfactory, although k-anonymity seems more amenable to big data protection.
- We highlighted connections between the main privacy models that might result in synergies between them in order to tackle big data:
  - The principles underlying all those models are deniability and permutation.

Universitat Rovira i Virgili

References I

G. Cormode, C. Procopiuc, D. Srivastava, E. Shen and T. Yu (2012) Differentially private spatial decompositions, in *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering-ICDE12*, Washington, DC, EUA, pp. 20-31. IEEE Computer Society.

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

J. Domingo-Ferrer and J. Soria-Comas (2018) Connecting randomized response, post-randomization, differential privacy and *t*-closeness via deniability and permutation. https://arxiv.org/abs/1803.02139

Universitat Rovira i Virgili

References II

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

C. Dwork (2006) Differential privacy, in *ICALP06*, LNCS 4052, Springer, pp. 1-12.

B. G. Greenberg, A.-L. A. Abul-Ela, W. R. Simmons and D. G. Horvitz (1969) The unrelated question randomized response model: theoretical framework. *Journal of the American Statistical Association*, 64(326):520-539.

N. Li, T. Li and S. Venkatasubramanian (2007) t-Closeness: privacy beyond k-anonymity and l-diversity, in *ICDE07*, pp. 106-115.

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

References III

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

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

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

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

References IV

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

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

J. Xu, Z. Zhang, X. Xiao, Y. Yang and G. Yu (2012) Differentially private histogram publication, in *Proceedings of the 2012 IEEE 28th Intl. Conf. on Data Engineering-ICDE'12*, Washington, DC, USA. IEEE Computer Society, pp. 3243.

Y. Wang, X. Wu, and D. Hu (2016) Using randomized response for differential privacy preserving data collection. In *EDBT/ICDT 2016 Joint Conference*, Bordeaux, France.

References V

J. Zhang, G. Cormode, C.M. Procopiuc, D. Srivastava and X. Xiao (2014) Privbayes: private data release via Bayesian networks, in *Proceedings of the 2014 ACM SIGMOD Intl. Conf. on Management of Data, SIGMOD'14*, New York, NY, USA. ACM, pp. 14231434.