Partage et Secret de l'Information de Santé – Nancy 15/10/10

L'Anonymisation des données du DMP

Etat des lieux du “Privacy Preserving Data Publishing”

T. Allard & *B.Nguyen*

Institut National de Recherches en Informatique et Automatique (INRIA)

Secure and Mobile Information Systems Team

& Université de Versailles St-Quentin

Programme DEMOTIS

Summary

- Introduction
- Data Partitionning Family
- The continuous release problem
- Data Perturbation Family

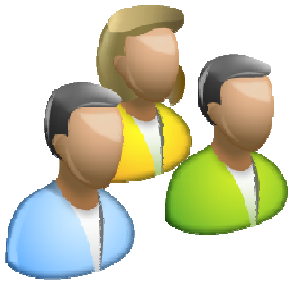
A Brief History...

- Protecting data about individuals goes back to the 19th century: Carrol Wright (Bureau of Labor Statistics, 1885)
- Stanley Warner: Interviews for market surveys (1965)
- Tore Dalenius (1977) definition of disclosure : “If the release of the statistic S makes it possible to determine the microdata value more accurately than without access to S, a disclosure has taken place.”
- Rakesh Agrawal (1999) invents Privacy Preserving Data Mining
- ... 2010 differential privacy
- However : the definition of anonymous information is vague !

Disclaimer

- Context : statistical study... probably too limitative for many practicians.
- We do not cover the problem of studies that are directly done by the hospital that collects the data
- Data can only be exported if it is “anonymized”.

Privacy Preserving Data Publishing (PPDP) Principle



Raw data

Anonymized data
(or *sanitized …*)

Individuals
ex: patients

Publisher
(trusted)
ex: hospital

Recipients *(no assumption of trust)*
ex: Pharma

Pseudonymization: A naïve privacy definition

SSN	Activity	Age	Diag
123	"M2 Stud."	22	Flu
456	"M1 Stud."	25	HIV
789	"L3 Stud."	20	Flu

Raw data

Pseudo	Activity	Age	Diag
ABC	"M2 Stud."	22	Flu
MNO	"M1 Stud"	25	HIV
XYZ	"L3 Stud."	20	Flu

Pseudonymized Data

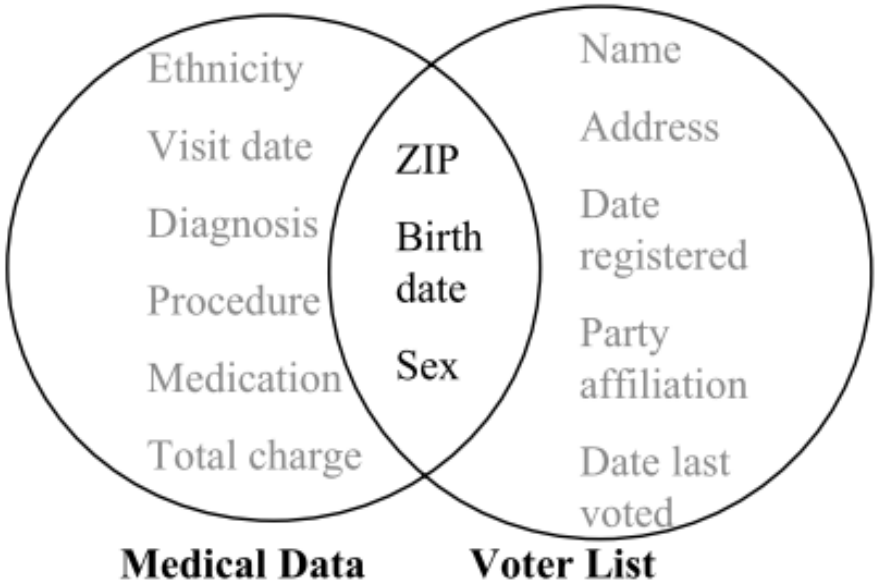
Individuals

Publisher *(trusted)*

Recipients *(no assumption of trust)*

Pseudonymization is not safe !

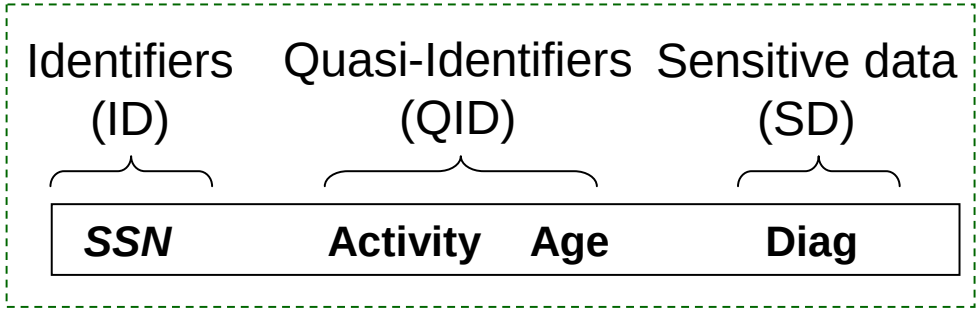
- Sweeney [1] shows the existence of *quasi-identifiers*:
  - Medical data were « anonymized » and released;
  - A voter list was publicly available;
  - → Identification of medical records of Governor Weld by joining datasets on the *quasi-identifiers.*



  - In the US census of 1990: « 87% of the population in the US had **characteristics that likely made them unique** based only on {5-digit Zip, gender, date of birth} » [1].

Data Partitioning Family

Data classification



- For each tuple:
  - Identifiers must be removed;
  - The link between a quasi-identifier and its corresponding sensitive data must be *obfuscated* but remain *true*

k-Anonymity

- Form groups of (at least) k tuples indistinguishable wrt their quasi-identifier:

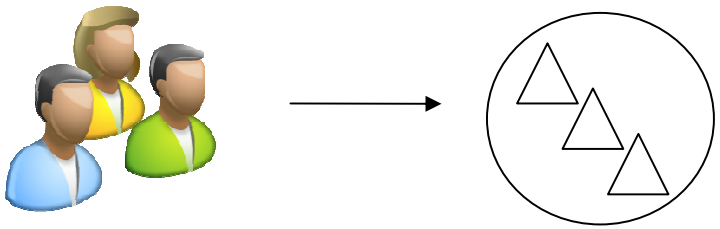
Name	**Activity**	**Age**	**Diag**
Sue	"M2 Stud."	22	Flu
Bob	"M1 Stud."	21	HIV
Bill	"L3 Stud."	20	Flu

Raw data

Activity	**Age**	**Diag**
"Student"	[20, 22]	Flu
"Student"	[20, 22]	HIV
"Student"	[20, 22]	Flu

3-anonymous data

Record linkage probability: 1/k



Questions : k-anonymité

- Cas d'école:

"Student" [20, 22] → {Flu, HIV, Flu}
"Teacher" [24, 27] → {Cancer, Cold, Cancer}

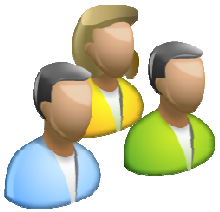
  - On peut voir les attributs du QID comme les axes d'analyse des données sensibles…
    - Peut on les déterminer avant l'anonymisation?
    - Typiquement, que contiennent les données sensibles? Leur cardinalité?
  - Si les groupes se chevauchent:

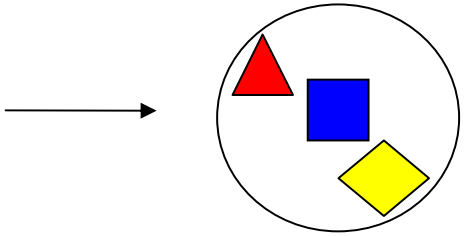
"Student" [20, 25] → {…}
"Teacher" [24, 27] → {…}

  - Quelles propriétés doivent être vérifiées pour qu'elles soient utilisables selon vous?

- Habituellement, faites-vous des croisements multi-sources? Comment?

L-diversity





- Ensure that each group has « enough diversity » wrt its sensitive data;

Name	Activity	Age	Diag
Pat	"MC"	27	Cancer
Dan	"PhD"	26	Cancer
San	"PhD"	24	Cancer

Raw data

Activity	Age	Diag
"Teacher"	[24, 27]	Cancer
"Teacher"	[24, 27]	Cancer
"Teacher"	[24, 27]	Cancer

3-anonymous data

Name	Activity	Age	Diag
Sue	"M2 Stud."	22	Flu
Pat	"MC"	27	Cancer
Dan	"PhD"	26	Cancer
San	"PhD"	24	Cancer
John	"M2 Stud"	22	Cold

Raw data

Activity	Age	Diag
"University"	[22, 27]	Flu
"University"	[22, 27]	Cold
"University"	[22, 27]	Cancer
"University"	[22, 27]	Cancer
"University"	[22, 27]	Cancer

3-diverse data

Attribute linkage probability: 1/L

Questions : l-diversité

- La l-diversité empêche typiquement:

"Teacher" [24, 27] → {Cancer, Cancer, Cancer}

- Quid de l'utilité de classes l-diverses?

t-closeness

- Distribution of sensitive values within each group ≈ Global distribution (factor *t*);
- Example:

Non-Sensitive		Sensitive	
Age	Gender	Disease	Count
< 40	M	Flu	400
< 40	M	Cancer	200
≥ 40	M	Flu	400
≥ 40	M	Cancer	200
≥ 40	F	Flu	400
≥ 40	F	Cancer	200

Limited gain in knowledge

The continuous release problem

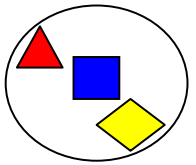
Limits of Data Partitioning

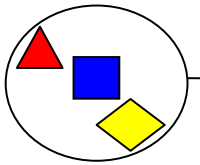
m-invarience [5]

	Name	Activity	Age	Diag
t1	***Bob***	**"M1 Stud."**	**21**	**HIV**
	Bill	"L3 Stud."	20	Flu
	Jim	"M2 Stud"	23	Cancer
t2	***Bob***	**"M1 Stud."**	**21**	**HIV**
	Helen	"L1 Stud."	18	Cold
	Jules	"L1 Stud"	19	Dysp.

Activity	Age	Diag
"Student"	[20, 23]	Flu
"Student"	[20, 23]	**HIV**
"Student"	[20, 23]	Cancer
"Student"	[19, 21]	**HIV**
"Student"	[19, 21]	Cold
"Student"	[19, 21]	Dysp

- Current models: « each group must be (relatively) invariant »;
  - May require introducing fake data;
  - Make hard « case histories »;
- Our direction: sampling
  - No fake data;
  - « Case histories » hard also;





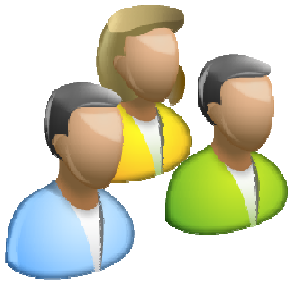
Questions: m-invarience

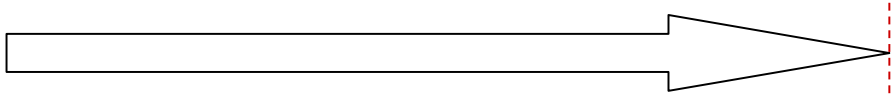
- Quels sont les cadres applicatifs du continuous release?
  - Suivi individuel de chaque dossier?
  - Suivi d'une population?
  - …?
- La dichotomie entre données transientes/permanentes est-elle pertinente?

Data Perturbation Family

Local Perturbation

Local Perturbation





Anonymized data
(or *sanitized …*)

Individuals

Recipients *(no assumption of trust)*

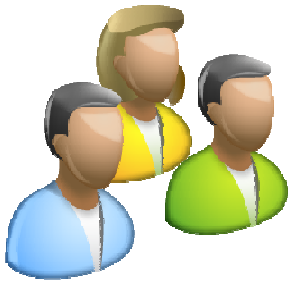
Mechanism

- Define a count query and send it to individuals;
- Each individual perturbs his answer with a known probability p;
- Receive the perturbed answers r_{pert} and compute an estimate of the real count r_{est};
- There are formal proofs of correctness: $r_{est} = (r_{pert} - p)/(1 - 2p)$

Data Perturbation Family

Input Perturbation

Input Perturbation



Raw data

Anonymized data
(or *sanitized …*)

Individuals

Publisher *(trusted)*

Recipients *(no assumption of trust)*

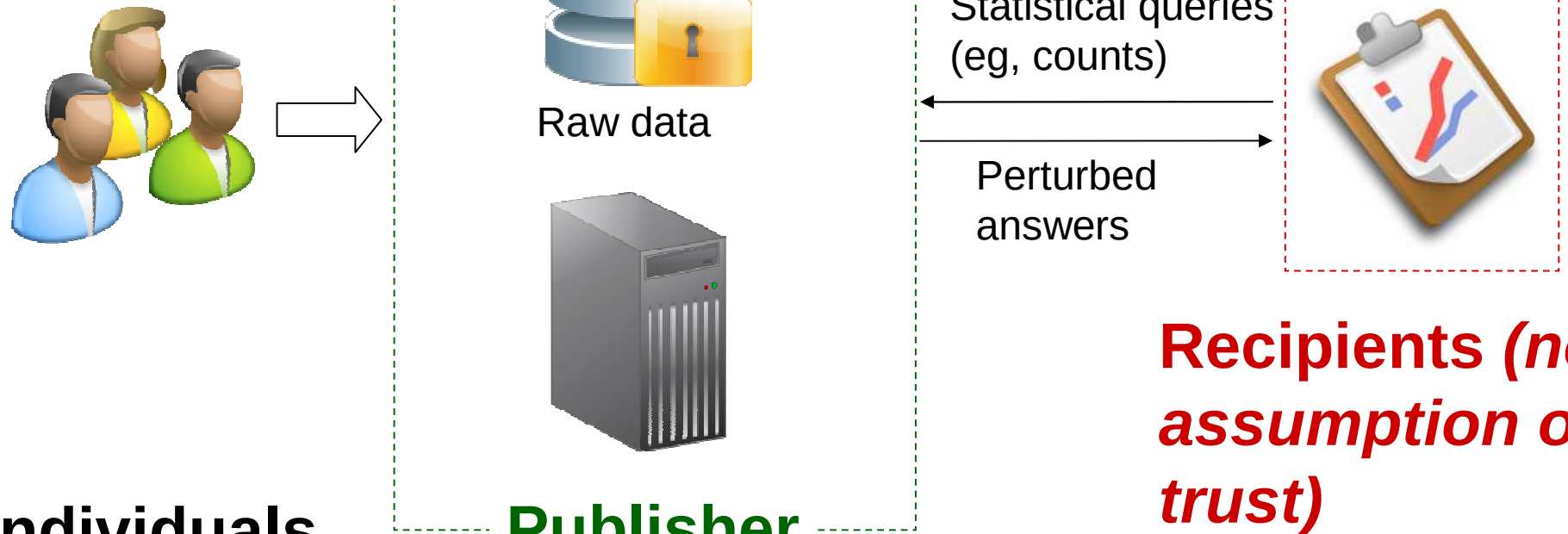
Mechanism (not detailled)

- Similar to Local Perturbation except that data is not perturbed independently;
- We can expect smaller errors;

Data Perturbation Family

Statistics Perturbation

Statistics Perturbation (interactive setting)



Individuals

Publisher *(trusted)*

Recipients *(no assumption of trust)*

Statistics perturbation

- Define a statistical query (eg, a count): Q_i;
- The server answers a count perturbed according to the query sensitivity: $Q_1 + \eta_1$;

- The error magnitude is low;
- The total number of queries is bounded;

Questions: pseudo-random

- De quel ordre est le nombre typique:
  - D'attributs dans une requête?
  - De valeurs possibles par attributs?
  - De requêtes d'une étude épidémiologique?
- Les données auxquelles vous avez accès contiennent-elles déjà des erreurs non intentionnelles?
- De quels estimateurs statistiques avez-vous besoin?
- « Deviner » les statistiques d'intérêt sans « voir » les données est-il réaliste?

Conclusion

- The difficulty of bridging the gap between the medical needs of very precise data, and legal constraints on privacy protection and anonymization.
- Participation of users in many studies depends on their security guarantees.
- Using secure hardware could convince patients to participate more in widespread studies. (our current work)

Merci!

References

[1] L. Sweeney. k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 10(5), 2002.

[2] Xiaokui Xiao , Yufei Tao, Anatomy: simple and effective privacy preservation, Proceedings of the 32nd international conference on Very large data bases, September 12-15, 2006, Seoul, Korea.

[3] Ashwin Machanavajjhala , Daniel Kifer , Johannes Gehrke , Muthuramakrishnan Venkitasubramaniam, L-diversity: Privacy beyond k-anonymity, ACM Transactions on Knowledge Discovery from Data (TKDD), v.1 n.1, p.3-es, March 2007.

[4] Ninghui Li, Tiancheng Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 106--115, April 2007.

[5] Xiaokui Xiao , Yufei Tao, M-invariance: towards privacy preserving re-publication of dynamic datasets, Proceedings of the 2007 ACM SIGMOD international conference on Management of data, June 11-14, 2007, Beijing, China

[6] S. L. Warner, “Randomized Response: A survey technique for eliminating evasive answer bias,” Journal of the American Statistical Association, 1965.

References

[7] Nina Mishra , Mark Sandler, Privacy via pseudorandom sketches, Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, June 26-28, 2006, Chicago, IL, USA

[8] Alexandre Evfimievski , Johannes Gehrke , Ramakrishnan Srikant, Limiting privacy breaches in privacy preserving data mining, Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, p.211-222, June 09-11, 2003, San Diego, California.

[9] Duncan G.T., Jabine T.B., and De Wolf V.A. (eds.). Private Lives and Public Policies. Report of the Committee on National Statistics' Panel on Confidentiality and Data Access. National Academy Press, WA, USA, 1993.

[10] J. Gouweleeuw, P. Kooiman, L. C. R. J. Willenborg, and P.-P. de Wolf, "The Post Randomisation Method for Protecting Microdata," QUESTIIO, vol. 22, no. 1, 1998.

[11] C. Dwork, A Firm Foundation for Private Data Analysis, *To appear in* Communications of the ACM, 2010.

[12] S. E. Fienberg and J. McIntyre, "Data swapping: Variations on a theme by Dalenius and Reiss," in *Privacy in Statistical Databases*, pp. 14-29, 2004.

[13] Bee-Chung Chen, Daniel Kifer, Kristen LeFevre and Ashwin Machanavajjhala, Privacy-Preserving Data Publishing, In *Foundations and Trends in Databases,* vol.2, issue 1–2 , January 2009.

Basic mechanism [6]

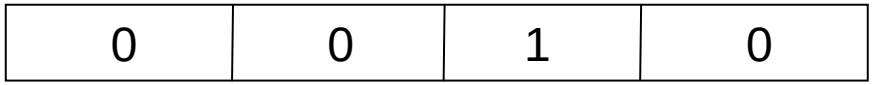
- Survey context (Warner, 1965);
- **Sensitive question:** Have you ever driven intoxicated?
- **Response:** truthful with probability *p* , lie with probability *(1-p);*
- Estimator:
  - Let π be the fraction of the population for which the true response is « *Yes* »
  - Expected proportion of « Yes »:
    $P(Yes) = (\pi * p) + (1 - \pi)*(1 - p)$
    → $\pi = [P(Yes) - (1 - p)] / (2p - 1)$
  - If m/n individuals answered « yes », π_{est} estimates π:
    $\pi_{est} = [m/n - (1 - p)] / (2p - 1)$

Extended mechanism [7]

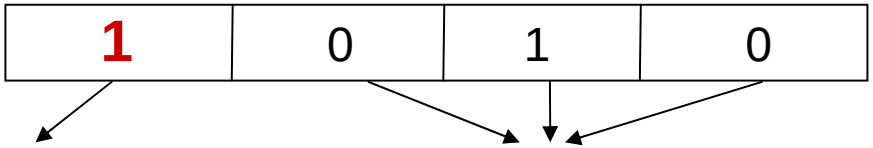
- (Mishra, 2006)
- **Server: defines queries:**
  - A conjunction of values (eg, people that have « HIV+ = true » and « aids = false»);
  - And wants to know the fraction of individuals that agree with the conjunction;

Extended mechanism [7] cont'

- **Individuals: each one receives the values of each conjunction :**
  - Eg : the conjunction contains « HIV+ » and « aids »;
  - Generates the vector of all the possible answers (« HIV+ = true » and « aids = true », « HIV+ = true » and « aids = false », …) with his answer set to 1:



  - And flips each element of the vector with probability p;



Flipped (probability *p*) Not flipped (probability *(1 - p)*)

Extended mechanism [7] cont’

- **Server: receives the pertubed vectors:**
  - Estimate the count result:
    - r_{pert} = number of perturbed vectors that agree with the conjunction;
    - $r_{est} = (r_{pert} – p)/(1 – 2p)$;
  - The true result r is proven to be « not too far away » from r_{est} ;