IBM Research – Zurich

Dr. Günter Karjoth
26. August 2013

Sommerakademie Kiel

Privacy by Design für Big Data

Privacy by Design (PbD)

- proposed by Ann Cavoukin, Privacy Commissioner Ontario
- mostly defined as general concepts
  - safeguards are incorporated into a systems and products from the very beginning of the development process
  - 9 PbD application areas including 'Big Data and Data Analytics'
- interpretation of PbD principles requires specific engineering expertise, contextual analysis, and a balancing of multilateral security and privacy interests

Ann Cavoukian: Privacy by Design – The 7 Foundational Principles.
Implementation and Mapping of Fair Information Practices. January 2011

PbD – The 7 Foundational Principles

1) **Proactive** not Reactive;
   - **Preventative** not Remedial
2) Privacy as the **Default Setting**
3) Privacy **Embedded** into Design
4) Full Functionality
   - **Positive-Sum**, not Zero-Sum
5) End-to-End Security
   - **Full Lifecycle Protection**
6) **Visibility** and **Transparency**
   - Keep it **Open**
7) **Respect** for User Privacy
   - Keep it **User-Centric**

☞ What are the concrete techniques or technologies?

Data Minimization and De-Identification

The use of strong de-identification techniques, data aggregation and encryption techniques are absolutely critical.

- De-identification in the context of legislation
- Re-identification risk assessment

- integration of legal and economic aspects
- practability / usability

What is the necessary level of de-identification?

The following questions may help to determine the identification risk factors.

- What kind of information is contained in the data set?
- Who will have access to the data set, and why?
- Are there any unique or uncommon characteristics (quasi-identifiers)?
- Will the data be a target of re-identification?
- What other data could be used to link with the data to re-identify indviduals?
  - ‘motivated intruder’ test
  - ‘re-identification in the round’
- What harm may result if individuals are re-identified?

Steps to manage the risk of re-identification (mitigation)

- Automated de-identification programs
  - remove or replace direct identifiers (masking)
  - generalize characteristic attributes to achieve a certain level of protection (k-Anonymity)
- The data recipient is bound by contract that limits the use and distribution of the disclosed information.
- Access to the data is limited (no data copy; partial view). Data owner may run the analysis itself and only return the result.

Content

The Australian Public Service Big Data Strategy

Protection of privacy

- incorporate “privacy by design” into big data analytics projects, and proactively ensure the privacy of the individual’s data and information; and
- adopt better practice methodologies that address the potential risk to privacy posed by big data analytics and “the mosaic effect”.

Principles

- All data sharing will conform to the relevant legislative and business requirements.
- Agencies are encouraged to conduct Privacy Impact Assessments (PIA) for any new big data projects and publish these PIAs (or modified versions if necessary).

☞ Action 1: Develop big data better practice guidance [by March 2014]

ISACA Whitepaper on PRIVACY & BIG DATA

Enterprises need a robust data-privacy solution to prevent data breaches and enforce data security in a complex IT environment. The solution should empower enterprises to:

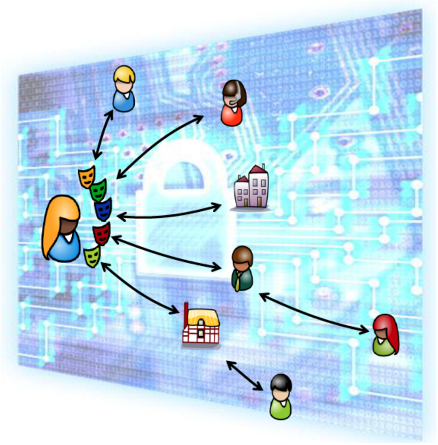
- Identify all sensitive data
- Ensure that sensitive data are identified and secured
- Demonstrate compliance with all applicable laws and regulations
- Proactively monitor the data and IT environment
- React and respond faster to data or privacy breaches with incident management

Content

Anonymous Certificates

☞ Prevent the disclosure of identity at the point of data collection; .e.g, cashier.

- Cryptografic mechanisms anonymise (personal) data for secure authentification without exposure of identity.
- attribute-based credentials enable the proof of properties instead of values; e.g., age above 18 years
- privacy-enabled authorization



Privacy-Enhancing Cryptography

- Secure Multiparty Computation
  - ☞ information sharing across private repositories
- Private Information Retrieval
  - ☞ searchers privacy – protecting the query search criteria
- Search over encrypted data (key words, order-preserving, . . .)
  - ☞ Cloud computing (?)
- Format-preserving encryption
  - ☞ masking

☞ Good in limiting the amount of data that users can acquire (collect)

Anonymous Data Analysis

Record #100031
Khalid Al-Midhar
Saudia Arabia
DOB: 07/12/76

one-way hash

Source:Agency #101
Record #100031
Name:cbd034409c22929518fa494f99dc9964
Citizen:b835b521c29f399c78124c4b59341691
DOB: 799709b2e5f26f796078fd815bebf724

?

#VX1RU9
Khaleed Al-midhar
San Francisco
DOB: 12/07/76
ID: 33000102334

Source: J.X. Dempsey and P. Rosenzweig: Technologies That Can Protect Privacy as Information Is Shared to Combat Terrorism, 2004. Available at `www.heritage.org/Research/HomelandDefense/lm11.cfm`

Anonymous Data Analysis (cont'd)

Data Standardization

"Robert"	"Robert"	"4ffe35db90d94c6041fb8ddf7b44df29"
"ROBERT"	"Robert"	"4ffe35db90d94c6041fb8ddf7b44df29"
"Rob"	"Robert"	"4ffe35db90d94c6041fb8ddf7b44df29"
"Bob"	"Robert"	"4ffe35db90d94c6041fb8ddf7b44df29"
"Bobby"	"Robert"	"4ffe35db90d94c6041fb8ddf7b44df29"

Variations

07/12/76	07/12/76	"799709b2e5f26f796078fd815bebf724"
	12/07/76	"8ceb0fe202b794c27694a83a5ad91df4"
	1976	"dd055f53a45702fe05e449c30ac80df9"

☞ dictionary attacks

Content

Data Linkage Problem

How to prevent users to know the private information of an individual by linking some public or easy-to-know database with the data they had received legally.

Challenges

- To check all possible kinds of knowledge that can be derived from the to-be-disclosed data (protection)
  - refuse the query
  - modify return data (masking, swapping values, rounding, additive noise, . . .)
- To achieve a balance between privacy protection and data availability (utility)

- Utility: accurate statistical info is released to users
- Privacy: each individual's sensitive information remains “hidden”

k-Anonymity

Each release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least k respondents.

- set by the data holder, possibly as the result of a negotiation with other parties
- satisfaction requires knowing how many individuals each released tuple matches

☞ How to produce a version of private table PT that satisfies k-anonymity wrt quasi-identifier QI ?

A Table

Race	Date of Birth	Sex	ZIP	Marital Status	Health Problems
asian	09/27/64	female	94139	divorced	hypertension
asian	09/30/64	female	94139	divorced	obesity
asian	04/18/64	male	94139	married	chest pain
asian	04/15/64	male	94139	married	obesity
black	03/13/63	male	94138	married	hypertension
black	03/18/63	male	94138	married	shortness of breath
black	09/13/64	female	94141	married	shortness of breath
black	09/07/64	female	94141	married	obesity
white	05/14/61	male	94138	single	chest pain
white	05/08/61	male	94138	single	obesity
white	09/15/61	female	94142	widow	shortness of breath

. . . and its minimal generalization

Race	DoB	Sex	ZIP	Marital Status
asian	64	-	941**	-
asian	64	-	941**	-
asian	64	-	941**	-
asian	64	-	941**	-
black	63	-	941**	-
black	63	-	941**	-
black	64	-	941**	-
black	64	-	941**	-
white	61	-	941**	-
white	61	-	941**	-
white	61	-	941**	-

$GT_{[0,2,1,2,2]}$

Race	DoB	Sex	ZIP	Marital Status
person	[60–64]	F	9413*	been married
person	[60–64]	F	9413*	been married
person	[60–64]	M	9413*	been married
person	[60–64]	M	9413*	been married
person	[60–64]	M	9413*	been married
person	[60–64]	M	9413*	been married
person	[60–64]	F	9414*	been married
person	[60–64]	F	9414*	been married
person	[60–64]	M	9413*	single
person	[60–64]	M	9413*	single
person	[60–64]	F	9414*	been married

$GT_{[1,3,0,1,1]}$

☞ The computation of an optimal *k*-Anonymity table is an NP-hard problem, independent of the granularity level.

Meyerson and Williams: On the complexity of optimal k-anonymity. ACM Symp. on Principles of Database Systems, 2004
G. Aggarwal et al.: Anonymizing Tables. ICDT 2005: 246–258

> Like everything else in security, anonymity systems shouldn't be fielded before being subjected to adversarial attacks. We all know that it's folly to implement a cryptographic system before it's rigorously attacked; why should we expect anonymity systems to be any different? And, like everything else in security, anonymity is a trade-off. There are benefits, and there are corresponding risks.
>
> Bruce Schneier

Attack Examples

Homogeneity Attack

Reto

Zip	Age
74678	26

Background Knowledge Attack

Satoshi (Japanese)

Zip	Age
74673	36

A 3-anonymous patient table

Zipcode	Age	Salary	Disease
746**	2*	20K	Heart Disease
746**	2*	30K	Heart Disease
746**	2*	40K	Heart Disease
7490*	≥ 40	50K	Gastritis
7490*	≥ 40	100K	Flu
7490*	≥ 40	70K	Bronchitis
746**	3*	60K	Heart Disease
746**	3*	80K	Cancer
746**	3*	90K	Cancer

k-Anonymity may fail if

- Sensitive values in an equivalence class lack diversity
- The attacker has background knowledge

l-Diversity Principle

A $q*$-block is a set of tuples in $T*$ whose non-sensitive attribute values generalize to $q*$.

> A $q*$-block is l-diverse if it contains at least l “well-represented” values for the sensitive attribute S.
> A table is l-diverse if every q-block is l-diverse.

- if there are l “well-represented” sensitive values in a q*-block then the attacker needs l-1 damaging pieces to infer a positive disclosure
- There are different instantiations of the l-diversity principle, e.g. **Entropy-l-Diversity** (information-theoretic notion):

$$-\sum_{s \in S} p_{(q^*,s)} \cdot log(p_{(q^*,s')} \geq log(l)$$

where $p_{(q^*,s))} = \frac{n_{(q^*,s)}}{\sum_{s' \in S} n_{(q^*,s')}}$ is the fraction of tuples in the q^*-block with sensitive attribute value equal to s.

t-Closeness

Definition of *l*-Diversity does not take into account

- the frequency distribution of the values in the sensitive attribute domain;
- the possible semantic relationships among sensitive attribute values; and
- the different sensitivity degree associated with different values of the sensitive attribute domain.

☞ skewness attacks and similarity attacks

> t-Closeness requires that the frequency distribution of the sensitive attribute values in each equivalence class has to be close (i.e., with distance less than a fixed threshold t) to the frequency distribution of the sensitive attribute values in the released microdata table.

Comparsion

Age	Gender	Condition
$[26-27]$	Male	Flu
$[26-27]$	Male	Flu
$[23-25]$	*	Cold
$[23-25]$	*	Diabetes
22	Male	Flu
22	Male	Cancer

$k=2$

Age	Gender	Condition
$[25-27]$	Male	Flu
$[25-27]$	Male	Flu
$[25-27]$	Male	Cold
$[22-24]$	*	Diabetes
$[22-24]$	*	Flu
$[22-24]$	*	Cancer

$k=3,\ E \geq log(1.9)$

Age	Gender	Condition
$[22-27]$	*	Flu
$[22-27]$	*	Flu
$[22-27]$	*	Cold
$[22-27]$	*	Diabetes
$[22-27]$	*	Flu
$[22-27]$	*	Cancer

$k=6,\ E \geq log(1.9),\ t$

- *l*-Diversity is hard to be achieved if one of the sensible values is very common; e.g., 90 % have “heart problems”.
- runtime complexity of *k*-Anonymity and *l*-Diversity are similar
- if some positive disclosures are acceptable it might be possible to be less conservative

Differential Privacy

Previous appproaches implicitly assume that the privacy of individuals **not included** in the dataset is **not at risk**.

> A data release is considered safe if the inclusion in the dataset of tuple t_p, related to respondent p, does not change the probability that a malicious recipient can correctly identify the sensitive attribute value associated with p.

☞ The techniques proposed in the literature to guarantee differential privacy are based on the addition of noise, and therefore **do not preserve data truthfulness**.

C. Dwork. Differential privacy. In *ICALP 2006*. LNCS, vol. 4052, pp. 1-12. Springer, 2006.

What Data Must be Anonymized?

- Relational data
  - Registration and demographic data
- Transactional (set-valued) data
  - Billing information
- Sequential data
  - DNA
- Trajectory
  - mobile phone traces
- Graph
  - Social Networks
- Text data
  - Clinical notes, tweets

Electronic Medical Records			
Name	YOB	ICD	DNA
Jim	1955	493.00, 185	C . . . T
Mary	1943	185, 157.3	A . . . G
Mary	1943	493.01	C . . . G
Carol	1965	493.02	C . . . G
Anne	1973	157.9, 493.03	G . . . C
Anne	1973	157.3	A . . . T

CLINICAL HISTORY: 77 year old female with a history of B-cell lymphoma (Marginal zone, SH-02-22222, 6/22/01). Flow cytometry and molecular diagnostics drawn.

Data Uniqueness

How much personal data you need to know for unique re-identification:

- (YoB, gender, 3-digit ZIP code) – 0.04 % of US citizens
- (DoB, gender, 5-digit ZIP code) – 87 % of US citizens
- 2 spatio-temporal points – 50 %
- 4 spatio-temporal points – 95 %
- 2 ICD codes – $> 90\%$

The de-identification process is risk-based. It balances the the need for protection with the usefulness of the data.

Text De-identification

Clinical history

77 year old female with a history of B-cell lymphona (Marginal zone, SH-02-22222, 6/22/01). Flow cytometry and molecular diagnostics drawn.

- Detect personal identifiers (e.g., name, record#, SSN)
- Replace or remove the discovered personal identifiers

☞ Preserve integrity of information while personal identity is effectively concealed.

Techniques

- white lists (high-frequency words are preserved in their original location)
- rule-based and dictionary-based (pattern matching)
- statistical learning

Content

Summary

PbD principles remind you to introduce privacy upfront; privacy risks are best managed proactively.

- Data minimization whenever possible.
  - Cryptography helps but is not a general solution.
- Data anonymization
  - De-identification of data does not necessarily give a (strict) guaranty of anonymity!
    ☞ Attack analysis is an important part of anonymization
    ☞ Privacy threats: Disclosure of identity, sensitive attributes, and inferential knowledge
  - Unstructured data makes it harder to name identifiers and “quasi-identifiers”
    ☞ Masking (redaction) extended with generalisation
- However, PbD principles are very general and a specific interpretation for Big Data is not yet available.

Outlook

- Homomorphic Encryption

A new form of encryption that allows computations to be carried out on encrypted data to obtain an encrypted result.

☞ Slooooow

- Privacy by Design 2.0

"Smart data": Intelligent "smart agents" will be developed and embedded into IT systems virtuelly – thereby creating "SmartData" – allowing one's data to protect itself.

☞ At the veeeery beginning

SmartData – Privacy Meets Evolutionary Robotics. Editors: Inman Harvey et al. Springer, 2013

Literature (i)

- Ann Cavoukian: Privacy by Design – The 7 Foundational Principles. Implementation and Mapping of Fair Information Practices. January 2011
- A. Cavoukian and K. El Emam: Dispelling the Myths Surrounding de-identification: Anonymization remains a strong tool for protecting privacy. June 2011. Available at `www.ipc.on.ca/images/Resources/anonymization.pdf`
- A. Cavoukian and J. Jonas: Privacy by Design in the Age of Big Data. June 8, 2012. Available at `http://privacybydesign.ca/content/uploads/2012/06/pbd-big_data.pdf`
- BITKOM: Management von Big-Data-Projekten (Leitfaden). June 18, 2013. Available at `http://www.bitkom.org/files/documents/LF_big_data2013_web.pdf`
- Privacy & Big Data. An ISACA White Paper, August 2013. Available at `www.isaca.org/Knowledge-Center/Research/ResearchDeliverables/Pages/Privacy-and-Big-Data.aspx`
- The Australian Public Service Big Data Strategy. August 2013. Available at `http://agict.gov.au/sites/default/files/Big%20Data%20Strategy.pdf`

Literature (ii)

- E. Buchmann: Anonymitätsmasse für Personendaten. S. 166–171, digma 2011.4.
- Big Data. digma 2013.1
- P. Samarati: Protecting Respondents' Identities in Microdata Release. IEEE Trans. on Knowledge and Data Engineering. 13(6), 2001; 1010–1027.
- L. Sweeney: k-anonymity: a model for protecting privacy. Int. Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557–570.
- T. Rosamilia: Privacy of Data, a business perspective. `http://www.almaden.ibm.com/institute/pdf/2003/TomRosamilia.pdf`
- P. Ohm: Broken Promises of Privacy – Responding to the Surprising Failure of Anonymization. UCLA Law Review, Vol. 57, p. 1701, 2010 U of Colorado Law Legal Studies Research Paper No. 9-12. Available at SSRN: `http://ssrn.com/abstract=1450006`

Vielen Dank für ihre Aufmerksamkeit.

Fragen?

Acknowledgements

Jan Camenisch (IBM Research – Zurich),

Aris Gkoulalas-Divanis (IBM Research – Dublin)