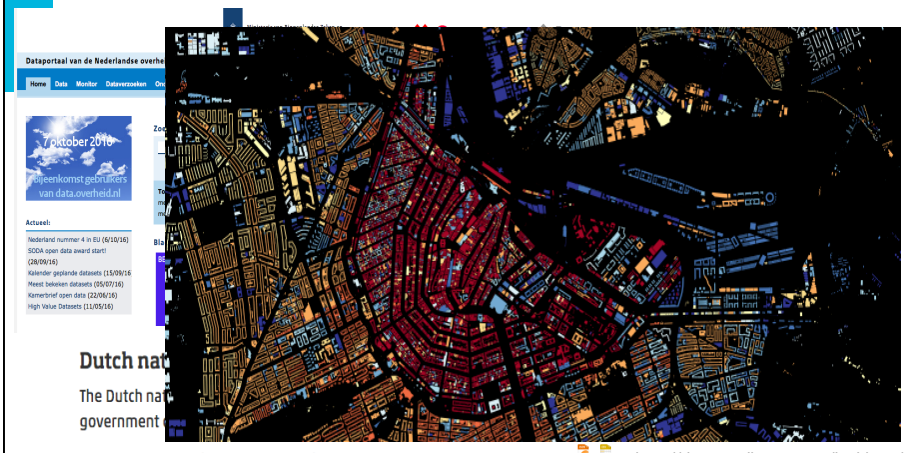
Software Analysis: Anonymity and Cryptography for Privacy

Winter School on Big Software on the Run:
Where software meets data

Dr. Zeki Erkin
z.erkin@tudelft.nl

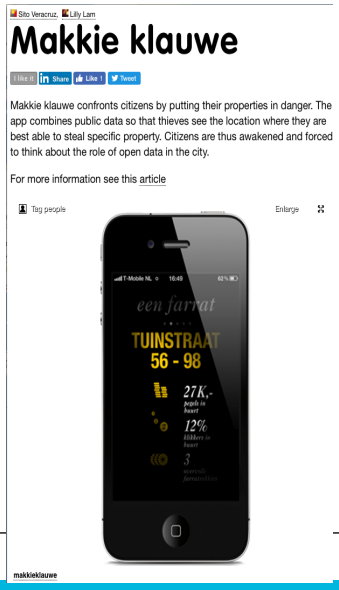
Cyber Security Group
Department of Intelligent Systems
Delft University of Technology

Open Data



Source: http://dev.citysdk.waag.org/buildings/

What can go wrong?



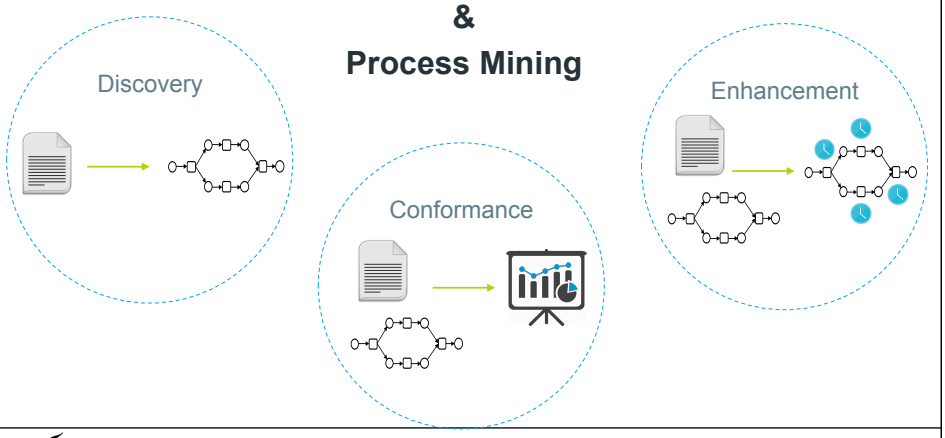
Github and email addresses…

Software Analysis via Logs

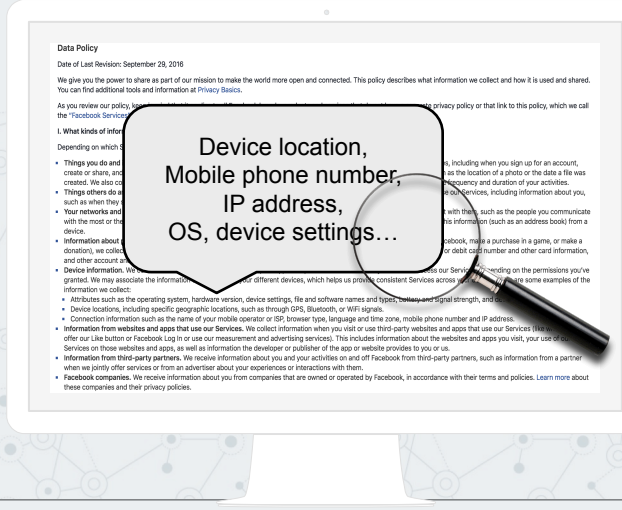
Model-based Analysis

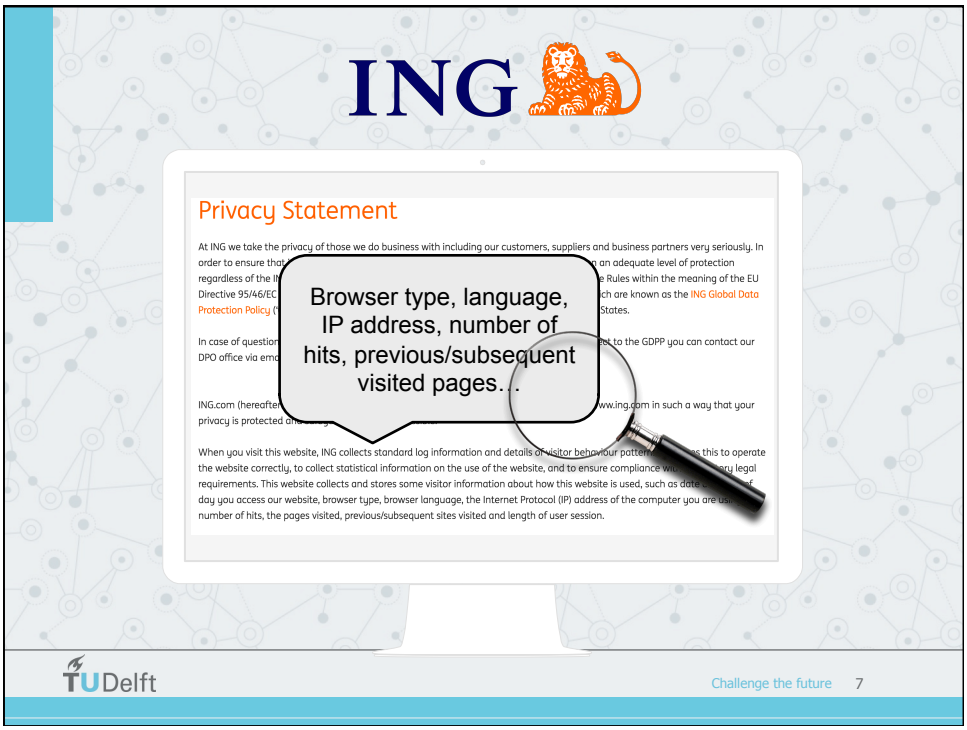
&

Process Mining



facebook





Privacy

- Case 1: Service provider is not trustworthy
  - process data for other purposes
  - sell data to third parties
- Case 2: Service provider is trustworthy
  - corporation take-over/bankruptcy
  - law and regulations prohibit storing sensitive data (medical data)
  - physical security/forgetful employees
  - competing service providers

Questions…

- From whom should we protect our data?
- How can we protect data?
- What are the best practices?
- How can we keep the balance between privacy and utility?

Solution

- Legislations and regulations
- Physical security
- Access control
- But not enough…
  - Anonymity
    - K-anonymity, L-diversity, T-closeness
    - Differential Privacy
  - Cryptography

Definitions and Terminology

Definition 1 (Privacy protection) Access to the published data should not enable the attacker to learn anything extra about any target victim compared to no access to the database, **even with the presence of any attacker's background knowledge obtained from other sources** [Dalenius in 1977].

Definition 2 (Privacy protection revised) Access to the published data should not enable the attacker to learn anything extra about any target victim compared to no access to the database, **given that the attacker has only a limited amount of background knowledge** [Dwork 2006].

Tables and Identifiers

- **Identifiers:** These are attributes, or a set there-of, that fully and non-ambiguously identify a person (also referred to as "victim") to some pieces of sensitive information in a certain table.
- **Quasi-Identifiers:** Represent a set of attributes used for linking with external information in order to uniquely identify individuals in a given anonymized table.
- **Sensitive Attributes:** These attributes contain values that are considered to be sensitive to the victim.
- **Non-sensitive Attributes:** These attributes are composed of column in the table that do not fall under any of the previously mentioned categories.

Linkage Attacks

- Linkage attacks try to link one individual to a record or to a value in a given table or to establish the presence of absence in the table itself.
  - Record linkage
  - Attribute linkage
  - Table linkage

Record Linkage

Job	Sex	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	35	Flu
Writer	Female	35	HIV
Dancer	Female	35	HIV
Dancer	Female	36	HIV

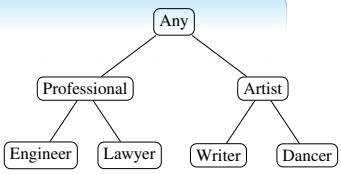
Name	Job	Sex	Age
Alice	Dancer	Female	36

K-Anonymity

Definition 3 (K-Anonymity) Reduce the granularity in such a way that for each *qid* in the table, there should be at least $k - 1$ other records with the same *qid*.

Job	Sex	Age	Disease
Professional	male	[35-40)	Hepatitis
Professional	male	[35-40)	Hepatitis
Professional	male	[35-40)	HIV
Artist	female	[35-40)	Flu
Artist	female	[35-40)	HIV
Artist	female	[35-40)	HIV
Artist	female	[35-40)	HIV

3-anonymous table



Job taxonomy tree

Attribute Linkage

- The goal is to determine which sensitive value belongs to the victim.
  - homogeneity attack
  - background knowledge attack

Job	Sex	Age	Disease
Professional	male	[35-40)	Hepatitis
Professional	male	[35-40)	Hepatitis
Professional	male	[35-40)	HIV
Artist	female	[35-40)	HIV
Artist	female	[35-40)	HIV
Artist	female	[35-40)	HIV

Homogeneity

Job	Sex	Age	Disease
Professional	male	[35-40)	Hepatitis
Professional	male	[35-40)	Hepatitis
Professional	male	[35-40)	Flu
Artist	female	[35-40)	Flu
Artist	female	[35-40)	HIV
Artist	female	[35-40)	HIV

Background knowledge

L-Diversity

Definition 4 (L-Diversity) A table is said to be l-diverse if every q*-block in the table contains at least l "well-represented" values for the sensitive attribute S.

Job	Sex	Age	Disease
Professional	male	[35-40)	Hepatitis
Professional	male	[35-40)	Hepatitis
Professional	male	[35-40)	HIV
Artist	female	[35-40)	HIV
Artist	female	[35-40)	HIV
Artist	female	[35-40)	Flu

2-Diverse patient table

Entropy L-diversity
Recursive (C,L)-Diversity

T-Closeness

1. Skewness attack: Due to the skewness of the data you could have partitions of the data.
2. Similarity attack: When the values in a q*-block are distinct, but semantically similar. For example, all the patients in a q*-block have some form of lung disease.

Definition 5 (T-Closeness) A table is said to achieve t-closeness if for every equivalence class (q*-block) in the table, the distribution of a sensitive values in the group is within t of the distribution of values in the the whole population.

Reading: t-Closeness: Privacy Beyond k-Anonymity and l-Diversity

Table Linkage

- In table linkage, the adversary tries to infer with a high enough probability that a certain individual is or is not present in a certain published sanitized table.

Age	Sex	Zipcode	Disease
[1,10]	m	[10001,20000]	gastric ulcer
[1,10]	m	[10001,20000]	dyspepsia
[1,10]	m	[10001,20000]	respiratory infection
[1,10]	m	[10001,20000]	respiratory infection
[11,20]	m	[20001,25000]	respiratory infection
[11,20]	m	[20001,25000]	respiratory infection
21	f	58000	flu
[26,30]	f	[35001,40000]	gastritis
[26,30]	f	[35001,40000]	pneumonia
56	f	33000	respiratory infection

General View

Privacy Model	Attack Model			
	Record Linkage	Attribute Linkage	Table Linkage	Probabilistic Attack
k-Anonymity	✓			
MultiR k-Anonymity	✓			
ℓ-Diversity	✓	✓		
Confidence Bounding		✓		
(α, k)-Anonymity	✓	✓		
(X, Y)-Privacy	✓	✓		
(k, e)-Anonymity		✓		
(ϵ, m)-Anonymity		✓		
Personalized Privacy		✓		
t-Closeness		✓		✓
δ-Presence			✓	
(c, t)-Isolation	✓			✓
ϵ-Differential Privacy			✓	✓
(d, γ)-Privacy			✓	✓
Distributional Privacy			✓	✓

Anonymization Operations

- Generalization and Suppression
- Anatomization and Permutation
- Perturbation
  - Additive noise
  - Data swapping
  - Synthetic data generation

Anatomization and Permutations

Age	Sex	Disease (sensitive)
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	Flu
38	Female	Flu
38	Female	Heart
38	Female	Heart

Original table

Age	Sex	GroupID
30	Male	1
30	Male	1
30	Male	1
32	Male	1
32	Male	1
32	Male	1
36	Female	2
38	Female	2
38	Female	2
38	Female	2

Quasi identifier table

Age	Sex	Disease (sensitive)
[30–35)	Male	Hepatitis
[30–35)	Male	Hepatitis
[30–35)	Male	HIV
[30–35)	Male	Hepatitis
[30–35)	Male	HIV
[30–35)	Male	HIV
[35–40)	Female	Flu
[35–40)	Female	Flu
[35–40)	Female	Heart
[35–40)	Female	Heart

Intermediate QID grouped table

GroupID	Disease (sensitive)	Count
1	Hepatitis	3
1	HIV	3
2	Flu	2
2	Heart	2

Sensitive table

Perturbation: Adding noise

- Data are distorted in a certain way that individual records cannot be recovered but aggregate distributions can be recovered.

$$\tilde{x} = x_i + r \text{ and } \tilde{y} = y_i + r \text{ for } 0 < i < N$$

$$r \in N(0, \sigma)$$

$$\sum_{i=0}^{N} \tilde{x} \cdot \tilde{y} = \sum_{i=0}^{N} (x_i + r) \cdot (y_i + r)$$

$$= \sum_{i=0}^{N} (x_i y_i + x_i r + y_i r + rr)$$

$$x_i r \approx 0$$

$$\approx \sum_{i=0}^{N} x_i y_i \qquad y_i r \approx 0$$

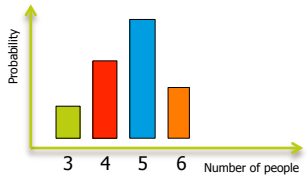
$$rr \approx 0$$

Perturbation: Challenges

- There is a trade-off between Privacy and Accuracy.
- Number of users(records) plays an important role.
- Security is not guaranteed. There is no tangible proof on either.
- The technique permits a limited number of operations to be performed.

Differential Privacy

- Normally the output is published, e.g. 5 people
- But now we output a number of outcomes with certain probabilities.



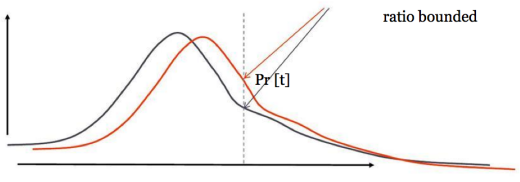
- And if you leave the population, the outcome does not change significantly

Differential Privacy

- Database: D
- Sanitizing algorithm: M

$$Pr(M(D_1) \in C) \leq e^{\epsilon} Pr(M(D_2) \in C)$$

Where
- D_1 and D_2 are any two neighbouring databases
- C in range(M)



$$e^{\epsilon} \sim (1 + \epsilon)$$

ε -Differential Privacy

- It is about the mechanism (algorithm that produces the outcome)
  - Unaffected by the auxiliary information
  - Independent of adversaries computational power
- Summary: Ensures that the removal or addition of a single database item does not (substantially) affect the outcome of any analysis.

$$\frac{1}{e^{\varepsilon}} \leq \frac{P(Response = r|X)}{P(Response = r|X^*)} \leq e^{\varepsilon}$$

Differential Privacy

- Techniques to achieve differential privacy
  - Input perturbation
  - **Output perturbation** (**Laplacian** and Exponential Mechanisms)
  - Perturbation of intermediate values
  - Sample and aggregate

Sensitivity of a function f

$$\Delta f = \max_{D_1, D_2} |f(D_1) - f(D_2)|$$

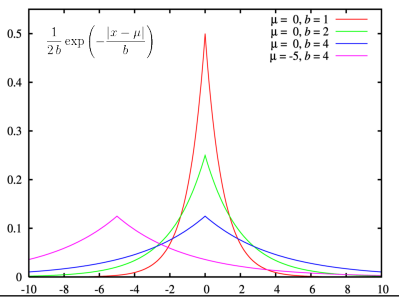
where D1 and D2 are two neighbouring data sets

- This measures how much one person changes the output
- Sensitivity is **1** for counting queries (number of people with a certain disease)

Laplacian Approach
Scaling noise with sensitivity

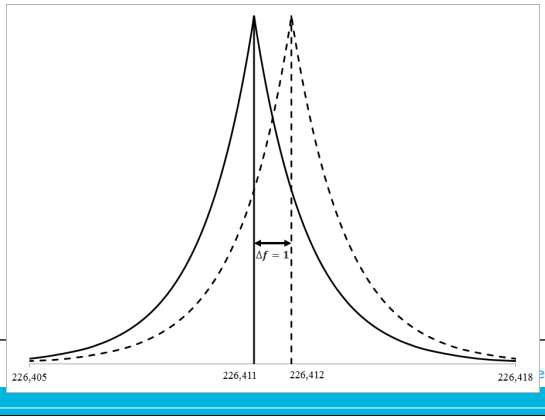
$$\Delta f = \max_{D_1, D_2} |f(D_1) - f(D_2)|$$

- To achieve ε -differential privacy, on query f, add scaled noise Lap(b) with b=Δf/ ε
- For sensitivity 1, add Lap(1/ ε)



What is achieved?

- Question: "How many people have more than $1 million in wealth?"
- Figure: Distribution of Query Response if the True Answer Contains, or Does Not Contain, Bill Gate



Properties

- **Post-processing:** the results of differentially private computations can be safely released because any post-processing computation will also be differentially private
- **Composition**: ε_1-DP mechanism followed by an ε_2-DP results in $(\varepsilon_1 + \varepsilon_2)$-differential privacy
- **Group privacy**: Composition queries (e.g. e vector of dimension d): Lap(dΔf/ ε): Sequential composibility is k ε -DP
- **Example**: Histogram queries with an output vector of size d
  - Could be Lap(dΔf/ ε)
  - But actually Lap(1/ ε)

Laplacian Approach

1. Select ε. The smaller the value, the greater the privacy
2. Compute the response A to the query using the original data
3. Compute the global sensitivity (Δf) for the query.
4. Generate a random value (noise) N from a Laplace distribution with mean = 0 and scale parameter $b = \Delta f/\varepsilon$.
5. Provide the user with response R=A+N

Example

- "The height of the average Lithuanian woman"

1. Select ε. Dwork suggests something between 0.01 to 0.1, or ln 2 or ln 3. Let's set ε=0.1
   - There are 2 queries: 1) total number of woman, and their total height, so for each query ε_q=0.05
2. Compute the Response to the Query Using the Original Data
3. Compute the Global Sensitivity ($\Delta \boldsymbol{f}$) for the Query
   - Total number of woman: $\Delta \boldsymbol{f}$=1
   - The sum of their height: $\Delta \boldsymbol{f}$=tallest persons' height

Example cont'd

4. Generate a Random Value (Noise) from a Laplace Distribution with Mean = 0 and Scale Parameter $b = \Delta f/\varepsilon$

Table 1—Response to Query on Average Height Over Database of Lithuanian Women
$\varepsilon_q = 0.05$

	True values	Δf	Laplace Noise		Noise Added Response	
			Low (0.01)	High (0.99)	Low	High
# of Lithuanian Women	1,603,014	1	−78	78	1,602,936	1,603,092
Total Height (inches)	105,798,924	99	−7,746	7,746	105,791,178	105,806,670
Average Height (inches)	66				65.99	66.01

Choice on ε

Table 7—The Probability that Laplace Noise Will Be Selected from Specified Ranges, for Varying Selections of ε $\Delta f = 1$

	0.001	**0.01**	**0.10**	**0.25**	**0.50**	***ln*(2)**	**1.00**	***ln*(3)**	**5.00**
±1	0.00	0.01	0.10	0.22	0.39	0.50	0.63	0.67	0.99
±2	0.00	0.02	0.18	0.39	0.63	0.75	0.86	0.89	1.00
±3	0.00	0.03	0.26	0.53	0.78	0.88	0.95	0.96	
±5	0.00	0.05	0.39	0.71	0.92	0.97	0.99	1.00	
±10	0.01	0.10	0.63	0.92	0.99	1.00	1.00		
±20	0.02	0.18	0.86	0.99	1.00				
±50	0.05	0.39	0.99	1.00					
±100	0.10	0.63	1.00						
±500	0.39	0.99							
±1000	0.63	1.00							
±5000	0.99								
±10000	1.00								

Important

- Selection of parameters Δf and ε is very important
  - Think about the skewness of the dataset: what if Bill Gates is in the list of people from Seattle
- Success depends on the function f
- Auxiliary information is very important
- Statements like "Turing is 2 inches taller than average Lithuanian woman" is challenging

Applications of DP

- Clever techniques do exist that achieve DP without adding worst case noise
- There are ε -DP variants of machine learning algorithms based on:
  - SVMs
  - Logistic regression

- K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. J. Mach. Learn. Res., 12:1069–1109, 2011
- R. Hall, A. Rinaldo, and L. Wasserman. Differential privacy for functions and functional data. J. Mach. Learn. Res., 14(1), 2013.
- G. Jagannathan, K. Pillaipakkamnatt, and R. N. Wright. A practical differentially private random decision tree classifier. In Proceedings of the 2009 IEEE International Conference on Data Mining Workshops, ICDMW '09, 2009.
- D. Kifer, A. D. Smith, and A. Thakurta. Private convex optimization for empirical risk minimization with applications to high-dimensional regression. Journal of Machine Learning Research - Proceedings Track, 23, 2012.

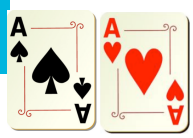
Cryptography

- Can we limit the information we leak?
  - Provable security
- Privacy Enhancing Technologies
- Privacy by Design
  - Multi-party computation
  - Cryptographic primitives
    - Homormorphic encryption

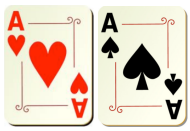
Love Game

Five-Card Trick (Bert den Boer, Eurocrypt 1989)

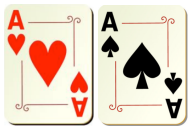
Alice Bob



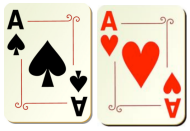
Yes



Yes

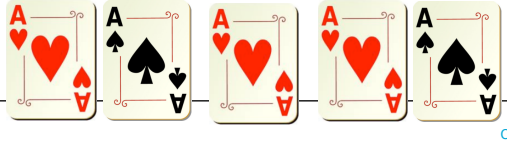


No



No

Match



No match

Love Game

- The game is actually an AND gate.

x	y	xy
0	0	0
1	0	0
0	1	0
1	1	1

$\cong$

Alice	*Bob*	*match?*
no	*no*	-
yes	*no*	-
no	*yes*	-
yes	*yes*	♡

Homomorphic Encryption

$$E_{pk}(m_1) \otimes E_{pk}(m_2) = E_{pk}(m_1 \oplus m_2)$$

- Paillier 1999"

$$E_{pk}(m_1) \times E_{pk}(m_2) = E_{pk}(m_1 + m_2)$$
$$\underbrace{E_{pk}(m) \times E_{pk}(m) \times \ldots E_{pk}(m)}_{c\ times} = E_{pk}(m)^c$$

An Example

Alice $[a], [b], pk$ Bob sk

$[\tilde{a}] := [a + r_1] = [a] \times [r_1]$

$[\tilde{b}] := [b + r_2] = [b] \times [r_2]$

$[\tilde{a}], [\tilde{b}]$ →

$\tilde{a} \cdot \tilde{b} = (a + r_1) \cdot (b + r_2)$

$= (ab + ar_2 + br_1 + r_1 r_2)$

$[a \cdot b] = [\tilde{a} \cdot \tilde{b}] \cdot [a]^{-r_2} \cdot [b]^{-r_1} \cdot [r_1 r_2]^{-1}$

← $[\tilde{a} \cdot \tilde{b}]$

43

Cryptographic Challenges

- Any function can be designed with secret inputs
- Design new systems based on efficiency
  - computational
  - communicational
- Disadvantage
  - Massive computation, multi-user, data expansion…etc.
- Security model
  - Honest-but-curious
  - Covert
  - Malicious

44

Summary

- Privacy in software analysis is a serious consideration for both individuals and business
- Privacy protections can be achieved by
  - Awareness
  - Law and regulations (EU Privacy Act)
  - Scientific/technological solutions
- There is no single solution but a group of them
  - anonymization, data perturbation, cryptographic protocol...etc
- None of them can provide full privacy. A combination of different approaches is needed
  - access control, physical security, privacy-preserving algorithms etc

Anonymisation is difficult…