

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

Dutch national data portal

The Dutch national government

- The Dutch national data portal is at <https://data.overheid.nl>

Publishers
Gemeente Amsterdam, Onderzoek, Informatie en Statistiek (162)

Winkels
Alle winkels in Amsterdam en omgeving zoals deze door het

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

What can go wrong?

The screenshot shows a news article titled "Makkie klauwe" by Lily Lam. The text states: "Makkie klauwe confronts citizens by putting their properties in danger. The app combines public data so that thieves see the location where they are best able to steal specific property. Citizens are thus awakened and forced to think about the role of open data in the city." Below the text is a smartphone displaying a real estate app with the following data: "TUINSTRAT 56 - 98", "27K+ people in house", "12% sold over 10 days", and "3 people in neighborhood".

Github and email addresses...

The screenshot shows a GitHub comment thread. A user named "futuretap" commented on Feb 18: "Please remove my data, too." Another user, "slang800", commented on Feb 26: "There are a few reasons why this is a bad idea...". Below the comments is a profile for "steipete" (PSPDFKit GmbH, Vienna, Austria, steipete@gmail.com, http://petersteinberger.com, joined Feb 27, 2009). To the right is a blue banner for "LEGAL ICT" with the text "Your legal services in IT law and...". Below the banner are categories: "Consumer rights", "Contracts", "Copyright", "General", "Net neutrality", "Patents", "Privacy", "Software". An article titled "Is it legal for GHTorrent to aggregate Github user data?" is also visible, dated 28 February 2016 by Arnoud Engelfriet. The article text includes: "Do users have a right to demand removal of their e-mail address from the GHTorrent data set? This is a frequent complaint for the project. GHTorrent aims at research on Github software projects, making metadata (such as user activity & profile information) easy to index and search. This includes e-mail addresses of users, allowing all kinds of links to be created. But how legal is that?"

Software Analysis via Logs

Model-based Analysis & Process Mining



facebook

Data Policy

Date of Last Revision: September 29, 2016

We give you the power to share as part of our mission to make the world more open and connected. This policy describes what information we collect and how it is used and shared. You can find additional tools and information at Privacy Basics.

As you review our policy, you may see references to our privacy policy or that link to this policy, which we call the "Facebook Service" or "Service".

What kinds of info?

- Depending on which Service you use, we collect information about you, including when you sign up for an account, as the location of a photo or the data a file was uploaded, the frequency and duration of your activities, and other information, including information about you, your contacts, and other people you interact with through the Service, such as the people you communicate with through the Service (such as an address book) from a contact on your phone, or when you use the Service to make a purchase in a game, or make a purchase on Facebook, or make a purchase on a website or other card information, or other information, including information about you, your contacts, and other people you interact with through the Service, depending on the permissions you've granted. We may associate the information we collect with different devices, which helps us provide consistent Services across devices. Here are some examples of the information we collect:
- Attributes such as the operating system, hardware version, device settings, file and software names and types, browser type and signal strength, and other information.
- Device locations, including specific geographic locations, such as through GPS, Bluetooth, or WiFi signals.
- Connection information such as the name of your mobile operator or ISP, browser type, language and time zone, mobile phone number and IP address.
- Information from websites and apps that use our Services. We collect information when you visit or use third-party websites and apps that use our Services (like our Like button or Facebook Login) or use our measurement and advertising services. This includes information about the websites and apps you visit, your use of our Services on those websites and apps, as well as information the developer or publisher of the app or website provides to you or us.
- Information from third-party partners. We receive information about you and your activities on and off Facebook from third-party partners, such as information from a partner when we jointly offer services or from an advertiser about your experiences or interactions with them.
- Facebook companies. We receive information about you from companies that are owned or operated by Facebook, in accordance with their terms and policies. Learn more about these companies and their privacy policies.

Device location,
Mobile phone number,
IP address,
OS, device settings...

ING 

Privacy Statement

At ING we take the privacy of those we do business with including our customers, suppliers and business partners very seriously. In order to ensure that we provide you with an adequate level of protection, we have implemented appropriate measures and procedures in accordance with the Rules within the meaning of the EU Directive 95/46/EC on the protection of personal data, which are known as the ING Global Data Protection Policy (the "Policy").

In case of question or complaint, you can contact our Data Protection Officer (DPO) office via email: dpo@ing.com. In the US, you can contact our Privacy Officer at privacy@ing.com. If you are in the EU, you can also contact our DPO via our website: www.ing.com in such a way that your privacy is protected and your data is not shared with third parties.

When you visit this website, ING collects standard log information and details of visitor behaviour patterns. We use this to operate the website correctly, to collect statistical information on the use of the website, and to ensure compliance with applicable legal requirements. This website collects and stores some visitor information about how this website is used, such as date and time of day you access our website, browser type, browser language, the Internet Protocol (IP) address of the computer you are using, the number of hits, the pages visited, previous/subsequent sites visited and length of user session.

Browser type, language, IP address, number of hits, previous/subsequent visited pages...

TU Delft Challenge the future 7

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

TU Delft Challenge the future 8

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?



Government



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].

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	✓			
Multi 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	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

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

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.

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

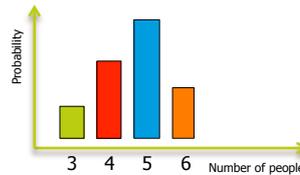
$$\begin{aligned} r &\in N(0, \sigma) \\ \sum_{i=0}^N \tilde{x} \cdot \tilde{y} &= \sum_{i=0}^N (x_i + r) \cdot (y_i + r) \\ &= \sum_{i=0}^N (x_i y_i + x_i r + y_i r + r r) \\ &\approx \sum_{i=0}^N x_i y_i \quad \begin{array}{l} x_i r \approx 0 \\ y_i r \approx 0 \\ r r \approx 0 \end{array} \end{aligned}$$

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

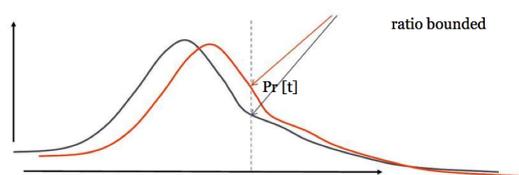
Differential Privacy

- 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 $\text{range}(M)$



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

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.

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

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

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

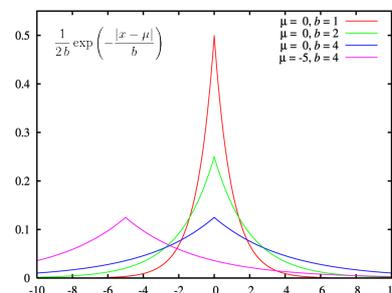
where D_1 and D_2 are two neighbouring data sets

- 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

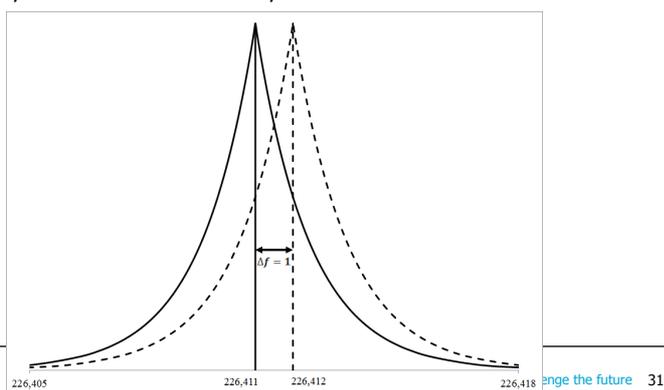
$$\Delta f = \max_{D_1, D_2} |f(D_1) - f(D_2)|$$

- To achieve ϵ -differential privacy, on query f , add scaled noise $\text{Lap}(b)$ with $b = \Delta f / \epsilon$
- For sensitivity 1, add $\text{Lap}(1/\epsilon)$



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 $(\epsilon_1 + \epsilon_2)$ -differential privacy
- **Group privacy:** Composition queries (e.g. \mathbf{e} vector of dimension d):
 $\text{Lap}(d\Delta f / \epsilon)$: Sequential composability is $k\epsilon$ -DP
- **Example:** Histogram queries with an output vector of size d
 - Could be $\text{Lap}(d\Delta f / \epsilon)$
 - But actually $\text{Lap}(1 / \epsilon)$

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 $b = \Delta f / \epsilon$.
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 $\epsilon = 0.1$
 - There are 2 queries: 1) total number of woman, and their total height, so for each query $\epsilon_q = 0.05$
 2. Compute the Response to the Query Using the Original Data
 3. Compute the Global Sensitivity (Δf) for the Query
 - Total number of woman: $\Delta f = 1$
 - The sum of their height: $\Delta f = \text{tallest persons' height}$

Example cont'd

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

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

	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 $\epsilon \Delta f = 1$

	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. Pillaijakkammatt, 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

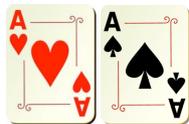
- Can we limit the information we leak?
 - Provable security
- Privacy Enhancing Technologies
- Privacy by Design
 - Multi-party computation
 - Cryptographic primitives
 - Homomorphic encryption

Love Game

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



Yes



No



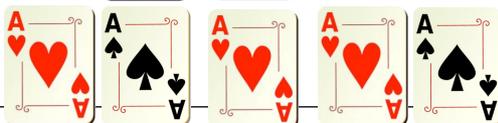
Yes



No



Match



No match

Love Game

- The game is actually an AND gate.

x	y	xy		<i>Alice</i>	<i>Bob</i>	<i>match?</i>
0	0	0	\cong	no	no	-
1	0	0		yes	no	-
0	1	0		no	yes	-
1	1	1		yes	yes	♥

Homomorphic Encryption

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

- Paillier 1999"

$$E_{pk}(m_1) \times E_{pk}(m_2) = E_{pk}(m_1 + m_2)$$

$$\underbrace{E_{pk}(m) \times E_{pk}(m) \times \dots \times E_{pk}(m)}_{c \text{ times}} = E_{pk}(m)^c$$

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...

