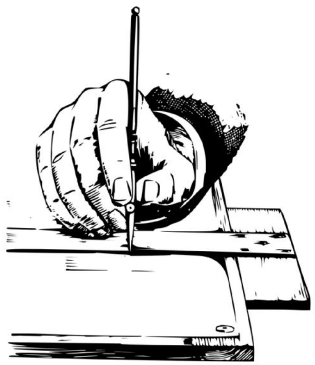
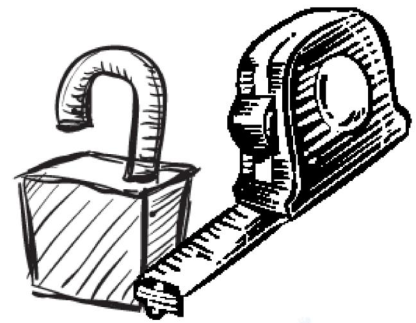
Systematic Privacy by Design Engineering





Carmela Troncoso

27th June 2017

institute imdea software

Privacy by Design – Let's have it!

Information and Privacy Commissioner of Ontario

PbD

Privacy by Design

Privacy by Design 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

Cavoukian et al. (2010)

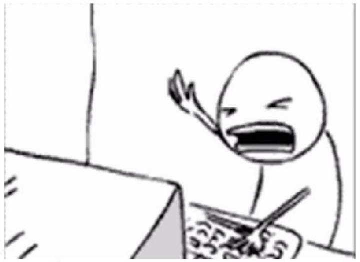
Article 25 European General Data Protection Regulation

GDPR

EU General Data Protection Regulation

“the controller shall [...] implement appropriate technical and organisational measures […] which are designed to implement data-protection principles[...] in order to meet the requirements of this Regulation and protect the rights of data subjects.”

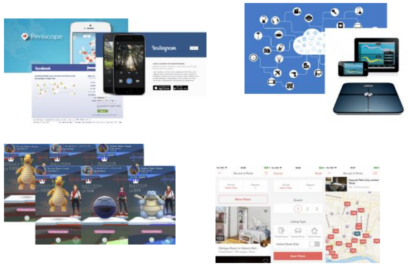
Actually... “Data Protection by design and by default”



BUT HOW ???????????

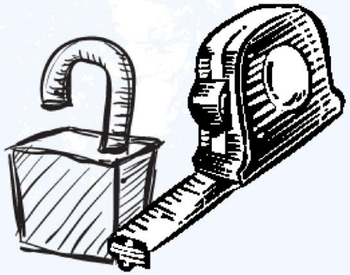
https://www.ipc.on.ca/images/resources/7foundationalprinciples.pdf
http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

HIGH PRIVACY

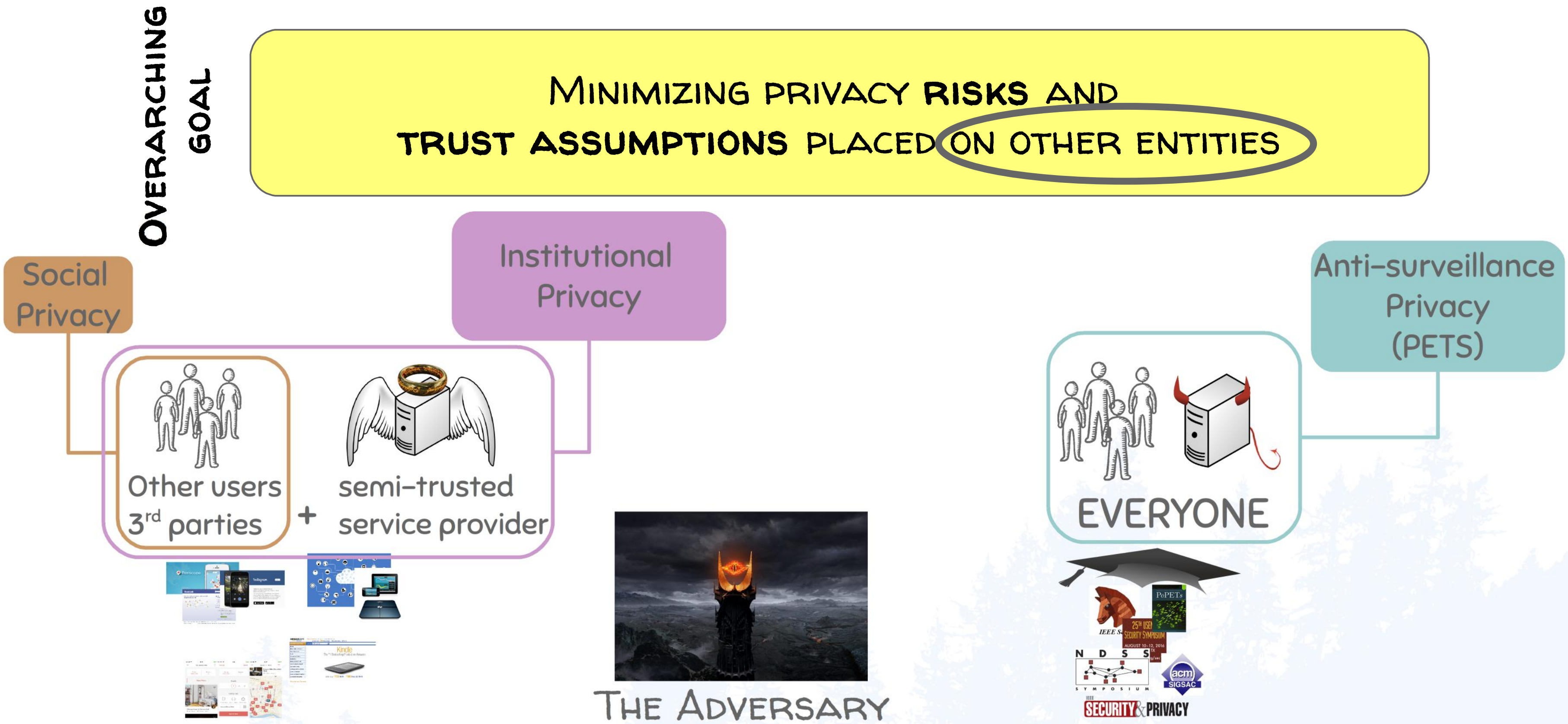


PART I:
Reasoning about Privacy when designing systems

PART II:
Evaluating Privacy in Privacy-Preserving systems



Privacy by Design Strategies



Seda Gurses, Carmela Troncoso, Claudia Diaz. Engineering Privacy by Design.Computers, Privacy & Data Protection. 2011
Seda Gurses, Carmela Troncoso, Claudia Diaz. Engineering Privacy by Design Reloaded. Amsterdam Privacy Conference. 2015
Seda Gurses and Claudia Diaz. "Two tales of privacy in online social networks." IEEE Security & Privacy Magazine. 2013

Privacy by Design Strategies

Overarching goal

Minimizing privacy **risks** and **trust assumptions** placed on other entities

Strategies

Minimize Collection	Minimize Disclosure	Minimize Linkability
Minimize Centralization	Minimize Replication	Minimize Retention

Great! but... how do we use these strategies?

We make explicit the activities and reasoning in **privacy engineering design** process

Seda Gurses, Carmela Troncoso, Claudia Diaz. Engineering Privacy by Design.Computers, Privacy & Data Protection. 2011
Seda Gurses, Carmela Troncoso, Claudia Diaz. Engineering Privacy by Design Reloaded. Amsterdam Privacy Conference. 2015

Case study: Electronic Toll Pricing

Motivation: European Electronic Toll Service (EETS)

Toll collection on European Roads trough On Board Equipment

Two approaches: <u>Satellite Technology</u> / DSRC

Starting assumptions

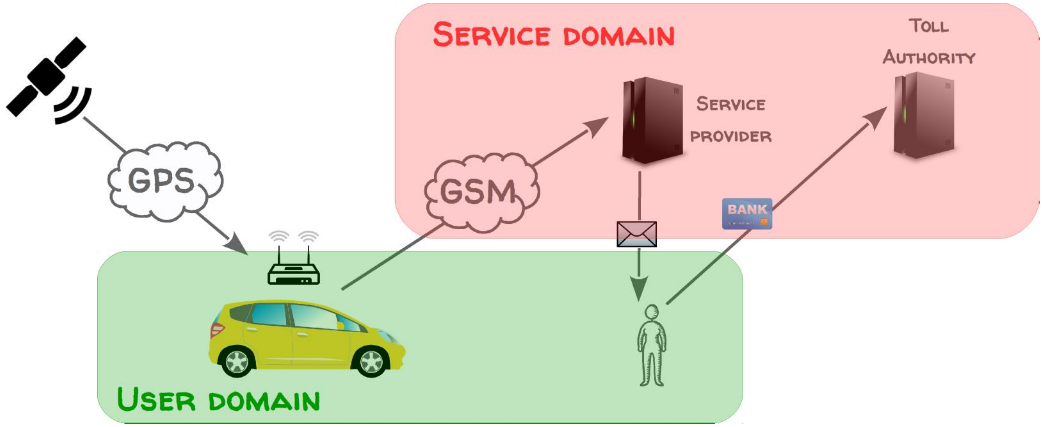
1) Well defined functionality
   - Charge depending on driving

2) Security, privacy & service integrity requirements
   - Users location should be private
   - No cheating clients

3) Initial reference system

Commission Decision of 6 October 2009 on the definition of the European Electronic Toll Service and its technical elements
http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32009D0750

Case study: Electronic Toll Pricing



Activity 1: Classify Entities in domains

User domain: components under the control of the user, eg, user devices

Service domain: components outside the control of the user, eg, backend system at provider

Activity 2: Identify necessary data for providing the service

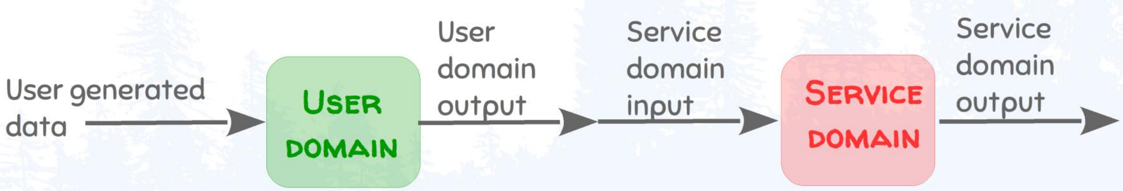
Location data – compute bill

Billing data – charge user

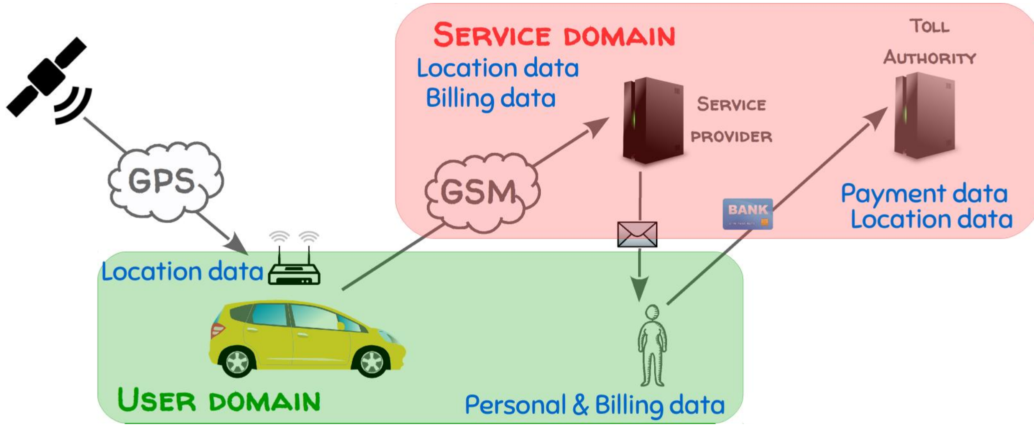
Personal data – send bill

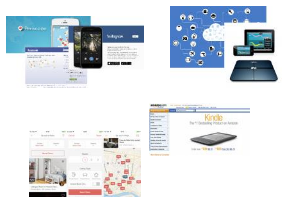
Payment data – perform payment

Activity 3: Distribute data in architecture



Case study: Electronic Toll Pricing





Activity 1: Classify Entities in domains

User domain: components under the control of the user, eg, user devices

Service domain: components outside the control of the user, eg, backend system at provider

Activity 2: Identify necessary data for providing the service

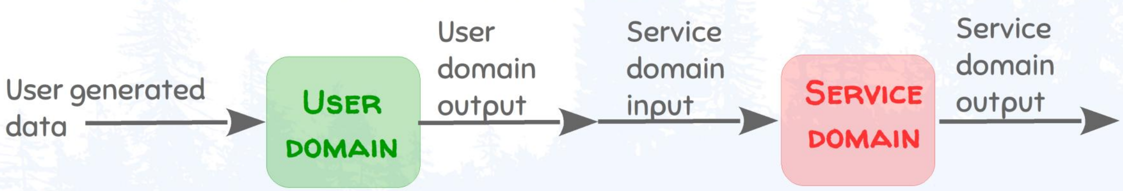
Location data – compute bill

Billing data – charge user

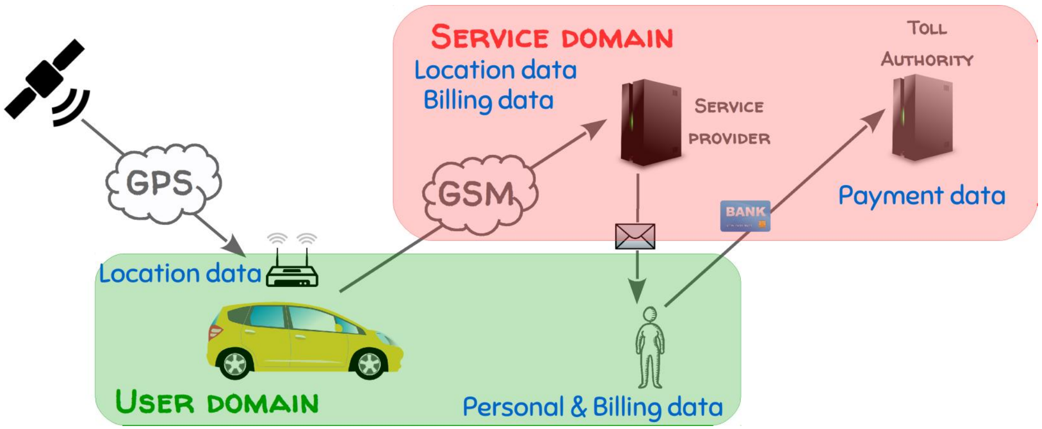
Personal data – send bill

Payment data – perform payment

Activity 3: Distribute data in architecture



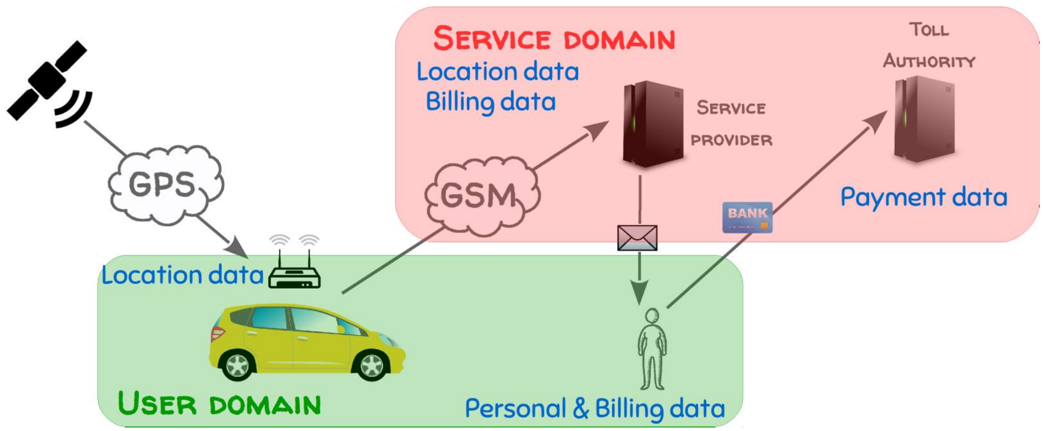
Case study: Electronic Toll Pricing



Trust Service to keep privacy of location data

Risk of privacy breach

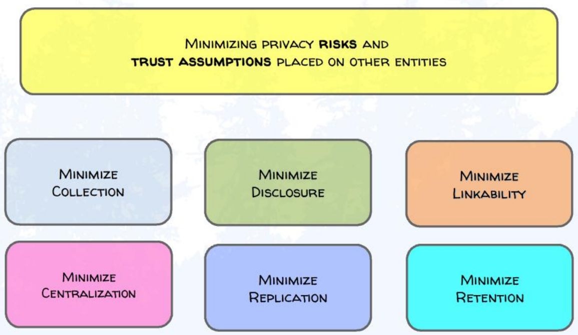
Case study: Electronic Toll Pricing



Location is not needed,
only the amount to bill!

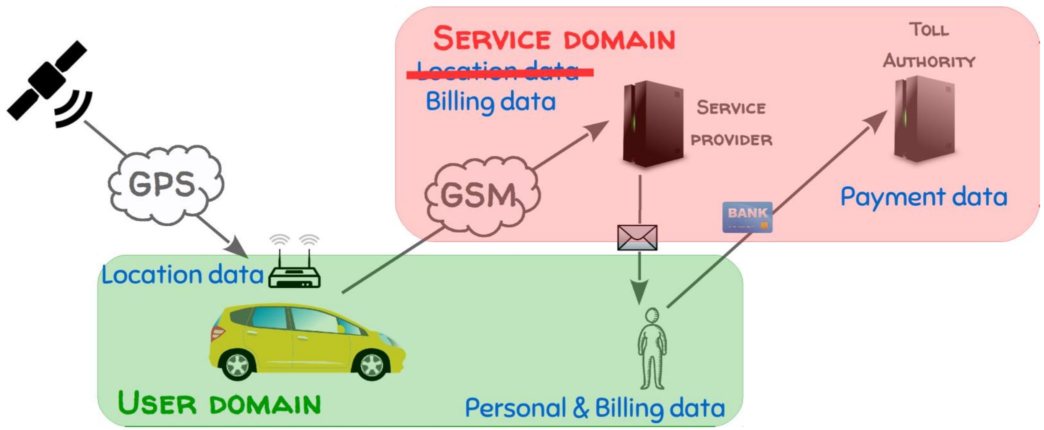
Activity 4: Select technological solutions following →

- not sending the data (local computations)
- encrypting the data
- advanced privacy-preserving protocols
- obfuscate the data
- anonymize the data



J. Balasch, A. Rial, C. Troncoso, B. Preneel, I. Verbauwhede, C. Geuens. PrETP "Privacy-Preserving Electronic Toll Pricing" USENIX Security Symposium 2010
C. Troncoso, G. Danezis, E. Kosta, J. Balasch, B. Preneel. "PriPAYD. Privacy-Friendly Pay-As-You-Drive Insurance" IEEE TDSC 2011

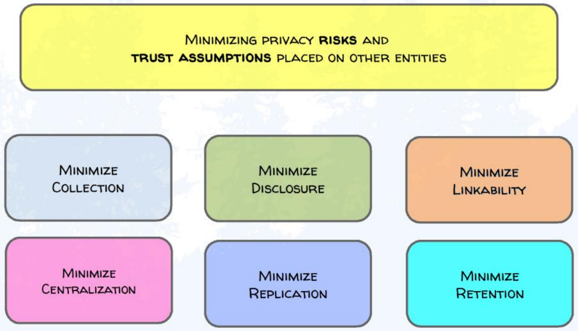
Case study: Electronic Toll Pricing



Location is not needed, only the amount to bill!

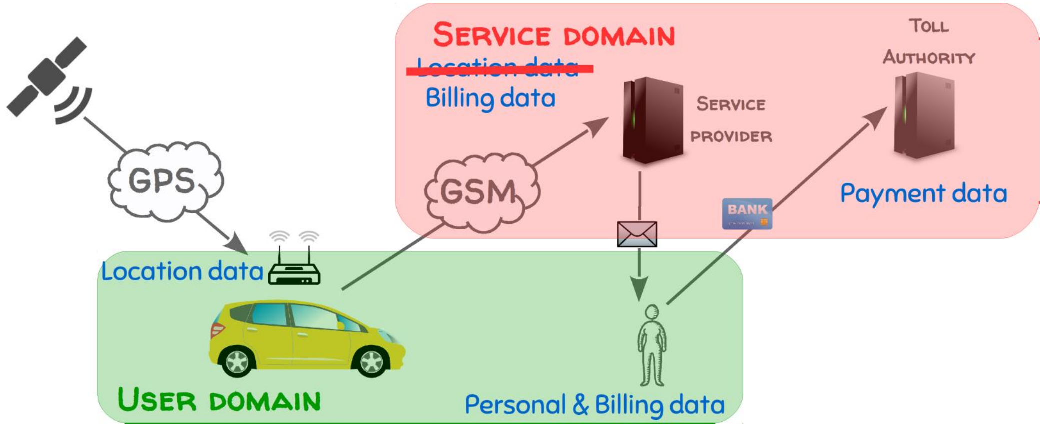
Activity 4: Select technological solutions following →

- not sending the data (local computations)
- encrypting the data
- advanced privacy-preserving protocols
- obfuscate the data
- anonymize the data



J. Balasch, A. Rial, C. Troncoso, B. Preneel, I. Verbauwhede, C. Geuens. PrETP "Privacy-Preserving Electronic Toll Pricing" USENIX Security Symposium 2010
C. Troncoso, G. Danezis, E. Kosta, J. Balasch, B. Preneel. "PriPAYD. Privacy-Friendly Pay-As-You-Drive Insurance" IEEE TDSC 2011

Case study: Electronic Toll Pricing

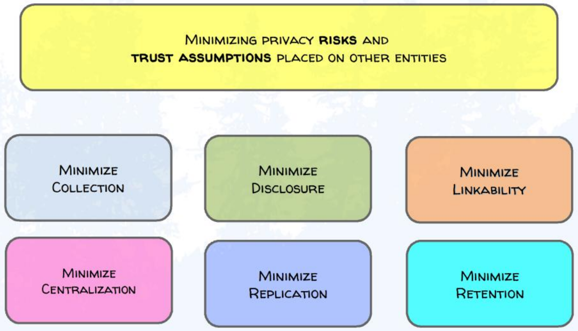


Location is not needed, only the amount to bill!

Service integrity?

Activity 4: Select technological solutions following →

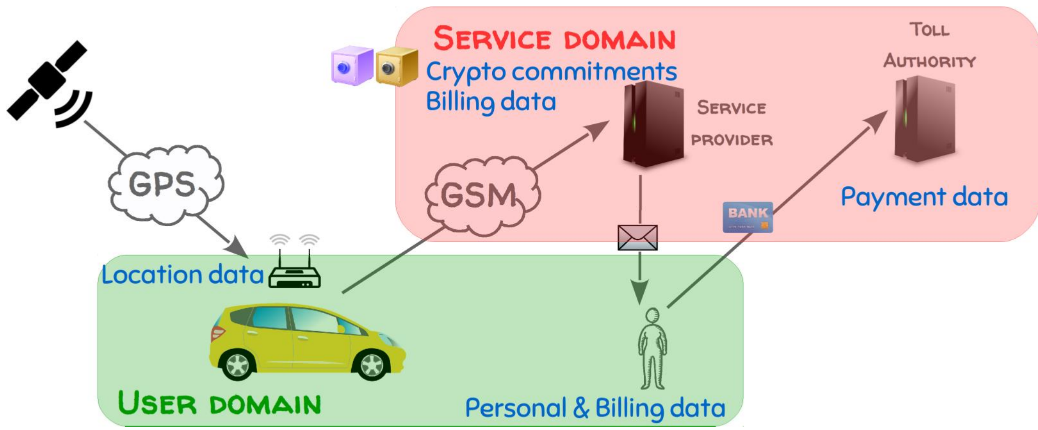
- not sending the data (local computations)
- encrypting the data
- advanced privacy-preserving protocols
- obfuscate the data
- anonymize the data



J. Balasch, A. Rial, C. Troncoso, B. Preneel, I. Verbauwhede, C. Geuens. PrETP "Privacy-Preserving Electronic Toll Pricing" USENIX Security Symposium 2010

C. Troncoso, G. Danezis, E. Kosta, J. Balasch, B. Preneel. "PriPAYD. Privacy-Friendly Pay-As-You-Drive Insurance" IEEE TDSC 2011

Case study: Electronic Toll Pricing



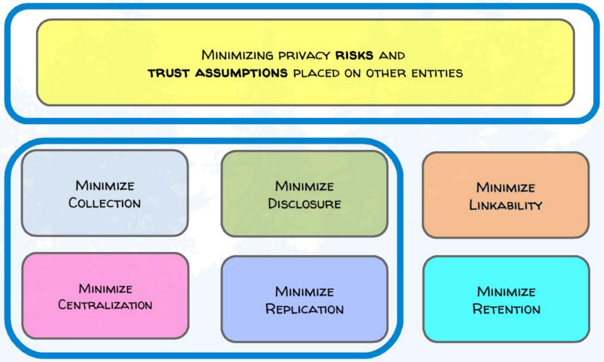
Location is not needed, only the amount to bill!

Service integrity

Requires knowledge of PETs
Privacy ENABLING Technologies

Activity 4: Select technological solutions following →

- not sending the data (local computations)
- encrypting the data
- advanced privacy-preserving protocols
- obfuscate the data
- anonymize the data

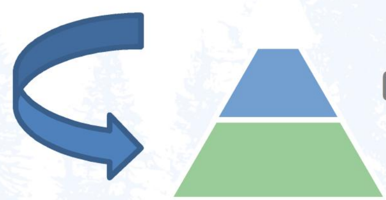


J. Balasch, A. Rial, C. Troncoso, B. Preneel, I. Verbauwhede, C. Geuens. PrETP "Privacy-Preserving Electronic Toll Pricing" USENIX Security Symposium 2010
C. Troncoso, G. Danezis, E. Kosta, J. Balasch, B. Preneel. "PriPAYD. Privacy-Friendly Pay-As-You-Drive Insurance" IEEE TDSC 2011

Privacy by design Engineering:
A change in the way we reason about systems

The Usual approach

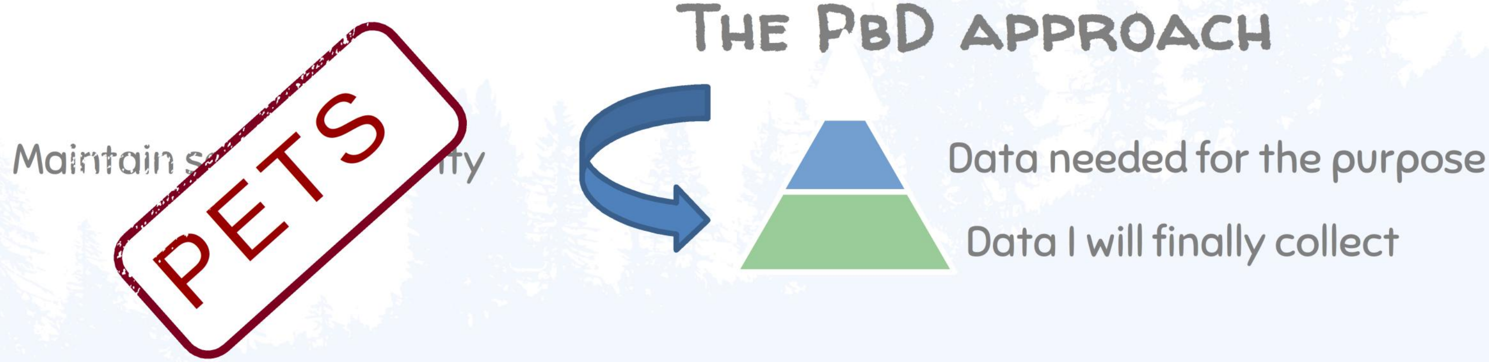
The PbD approach



Privacy by design Engineering:
A change in the way we reason about systems

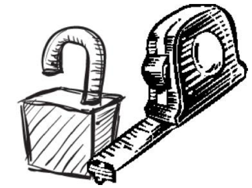
The Usual approach

The PbD approach



PART I:
Reasoning about Privacy when designing systems

PART II:
Evaluating Privacy in Privacy-Preserving systems



PRIVACY-PRESERVING SOLUTIONS

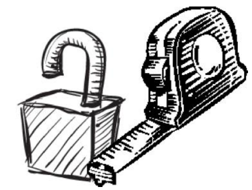
CRYPTO-BASED VS ANONYMIZATION/OBFUSCATION

WELL ESTABLISHED DESIGN AND EVALUATION METHODS

- Private searches
- Private billing
- Private comparison
- Private sharing
- Private statistics computation
- Private electronic cash
- Private genomic computations
- …

PART I:
Reasoning about
Privacy when designing
systems

PART II:
Evaluating Privacy in
Privacy-Preserving
systems



PRIVACY-PRESERVING SOLUTIONS

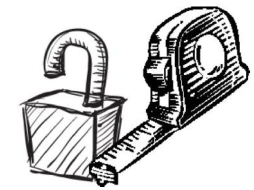
CRYPTO-BASED VS ANONYMIZATION/OBFUSCATION

WELL ESTABLISHED DESIGN AND EVALUATION METHODS

but expensive and require expertise

PART I:
Reasoning about Privacy when designing systems

PART II:
Evaluating Privacy in Privacy-Preserving systems



PRIVACY-PRESERVING SOLUTIONS

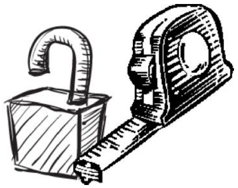
CRYPTO-BASED VS ANONYMIZATION/OBFUSCATION

cheap but...

DIFFICULT TO DESIGN / EVALUATE

PART I:
Reasoning about Privacy when designing systems

PART II:
Evaluating Privacy in Privacy-Preserving systems



PRIVACY-PRESERVING SOLUTIONS

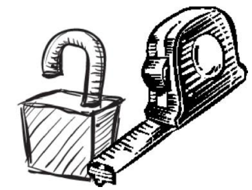
CRYPTO-BASED VS ANONYMIZATION/OBFUSCATION

cheap but...

DIFFICULT TO DESIGN / EVALUATE

PART I:
Reasoning about Privacy when designing systems

PART II:
Evaluating Privacy in Privacy-Preserving systems



PRIVACY-PRESERVING SOLUTIONS

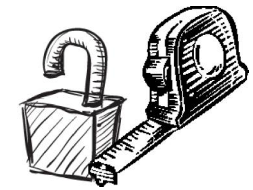
CRYPTO-BASED VS ANONYMIZATION/OBFUSCATION

cheap but…

DIFFICULT TO DESIGN / EVALUATE

PART I:
Reasoning about Privacy when designing systems

PART II:
Evaluating Privacy in Privacy-Preserving systems



PRIVACY-PRESERVING SOLUTIONS

CRYPTO-BASED VS ANONYMIZATION/OBFUSCATION

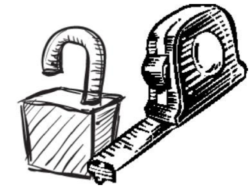
cheap but...

DIFFICULT TO DESIGN / EVALUATE

The adversary knows!

PART I:
Reasoning about Privacy when designing systems

PART II:
Evaluating Privacy in Privacy-Preserving systems



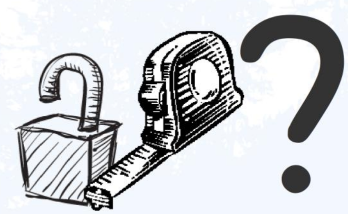
PRIVACY-PRESERVING SOLUTIONS

CRYPTO-BASED VS ANONYMIZATION/OBFUSCATION

cheap but...

DIFFICULT TO DESIGN / EVALUATE

The adversary knows!



We need technical objectives – PRIVACY GOALS

Anonymity: decoupling identity and action

Pseudonymity: pseudonymous as ID (personal data!)

Unlinkability: hiding link between actions

Unobservability: hiding the very existence of actions

Plausible deniability: not possible to prove a link between identity and action

"obfuscation": not possible to recover a real item from a noisy item

Why is it so difficult to achieve them?

Let's take one example: Anonymity

Art. 29 WP's opinion on anonymization techniques:

3 criteria to decide a dataset is non-anonymous (pseudonymous):

1) is it still possible to single out an individual

2) is it still possible to link two records within a dataset (or between two datasets)

3) can information be inferred concerning an individual?

http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Let's take one example: Anonymity

1) is it still possible to single out an individual

"the median size of the individual's anonymity set in the U.S. working population is 1, 21 and 34,980, for locations known at the granularity of a census block, census track and county respectively"

On the Anonymity of Home/Work Location Pairs

Philippe Golle and Kurt Partridge

Palo Alto Research Center
{pgolle, kurt}@parc.com

Abstract. Many applications benefit from user
cation data raises privacy concerns. Anonymizatio

location

Unique in the Crowd: The privacy bounds of human mobility

Yves-Alexandre de Montjoye[1,2], César A. Hidalgo[1,3,4], Michel Verleysen[2] & Vincent D. Blondel[2,5]

[1]Massachusetts Institute of Technology, Media Lab, 20 Ames Street, Cambridge, MA 02139 USA, [2]Université catholique de Louvain, Institute for Information and Communication Technologies, Electronics and Applied Mathematics, Avenue Georges Lemaître 4, B-1348 Louvain-la-Neuve, Belgium, [3]Harvard University, Center for International Development, 79 JFK Street, Cambridge, MA 02138, USA, [4]Instituto de Sistemas Complejos de Valparaíso, Paseo 21 de Mayo, Valparaíso, Chile, [5]Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

We study fifteen months of human mobility data for one and a half million individuals and find that human mobility traces are highly unique. In fact, in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals. We coarsen the data spatially and temporally to find a

Let's take one example: Anonymity

1) is it still possible to single out an individual

On the Anonymity of Home/Work Location Pairs

Philippe Golle and Kurt Partridge

Palo Alto Research Center
{pgolle, kurt}@parc.com

Abstract. Many applications benefit from user
cation data raises privacy concerns. Anonymizatio

location

Unique in the Crowd: The privacy bounds of human mobility

Yves-Alexandre de Montjoye[1,2], César A. Hidalgo[1,3,4], Michel Verleysen[2] & Vincent D. Blondel[2,5]

[1]Massachusetts Institute of Technology, Media Lab, 20 Ames Street, Cambridge, MA 02139 USA, [2]Université catholique de Louvain, Institute for Information and Communication Technologies, Electronics and Applied Mathematics, Avenue Georges Lemaître 4, B-1348 Louvain-la-Neuve, Belgium, [3]Harvard University, Center for International Development, 79 JFK Street, Cambridge, MA 02138, USA, [4]Instituto de Sistemas Complejos de Valparaíso, Paseo 21 de Mayo, Valparaíso, Chile, [5]Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

We study fifteen months of human mobility data for one and a half million individuals and find that human mobility traces are highly unique. In fact, in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals. We coarsen the data spatially and temporally to find a

"if the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals." [15 montsh, 1.5M people]

Let's take one example: Anonymity

1) is it still possible to single out an individual

On the Anonymity of Home/Work Location Pairs

Philippe Golle and Kurt Partridge

Palo Alto Research Center
{pgolle, kurt}@parc.com

Abstract. Many applications benefit from user cation data raises privacy concerns. Anonymizatio

location

Unique in the Crowd: The privacy bounds of human mobility

Yves-Alexandre de Montjoye[1,2], César A. Hidalgo[1,3,4], Michel Verleysen[2] & Vincent D. Blondel[2,5]

[1]Massachusetts Institute of Technology, Media Lab, 20 Ames Street, Cambridge, MA 02139 USA, [2]Université catholique de Louvain, Institute for Information and Communication Technologies, Electronics and Applied Mathematics, Avenue Georges Lemaître 4, B-1348 Louvain-la-Neuve, Belgium, [3]Harvard University, Center for International Development, 79 JFK Street, Cambridge, MA 02138, USA, [4]Instituto de Sistemas Complejos de Valparaíso, Paseo 21 de Mayo, Valparaíso, Chile, [5]Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

We study fifteen months of human mobility data for one and a half million individuals and find that human mobility traces are highly unique. In fact, in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals. We coarsen the data spatially and temporally to find a

How Unique is Your Browser?
a report on the Panopticlick experiment

Peter Eckersley
Senior Staff Technologist
Electronic Frontier Foundatio
pde@eff.org

83.6% had completely unique fingerprints
(entropy: 18.1 bits, or more)

94.2% of "typical desktop browsers" were unique
(entropy: 18.8 bits, or more)

web browser

Let's take one example: Anonymity

1) is it still possible to single out an individual

On the Anonymity of Home/Work Location Pairs

Philippe Golle and Kurt Partridge

Palo Alto Research Center
{pgolle, kurt}@parc.com

Abstract. Many applications benefit from user
cation data raises privacy concerns. Anonymizatio

location

Unique in the Crowd: The privacy bounds of human mobility

Yves-Alexandre de Montjoye[1,2], César A. Hidalgo[1,3,4], Michel Verleysen[2] & Vincent D. Blondel[2,5]

[1]Massachusetts Institute of Technology, Media Lab, 20 Ames Street, Cambridge, MA 02139 USA, [2]Université catholique de Louvain, Institute for Information and Communication Technologies, Electronics and Applied Mathematics, Avenue Georges Lemaître 4, B-1348 Louvain-la-Neuve, Belgium, [3]Harvard University, Center for International Development, 79 JFK Street, Cambridge, MA 02138, USA, [4]Instituto de Sistemas Complejos de Valparaíso, Paseo 21 de Mayo, Valparaíso, Chile, [5]Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

We study fifteen months of human mobility data for one and a half million individuals and find that human mobility traces are highly unique. In fact, in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals. We coarsen the data spatially and temporally to find a

How Unique is Your Browser?
a report on the Panopticlick experiment

Peter Eckersley
Senior Staff Technologist
Electronic Frontier Foundation
pde@eff.org

web browser

L. Sweeney, Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.

Simple Demographics Often Identify People Uniquely

Latanya Sweeney
Carnegie Mellon University
latanya@andrew.cmu.edu

"It was found that 87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth}"

Let's take one example: **Anonymity**

2) Link two records within a dataset (or datasets)

De-anonymizing Social Networks

Arvind Narayanan and Vitaly Shmatikov
The University of Texas at Austin

Abstract

Operators of online social networks are increasingly sharing potentially sensitive information about users and their relationships with advertisers, application developers, and data-mining researchers. Privacy is typically protected by anonymization, i.e., removing names, addresses, etc.

We present a framework for analyzing privacy and anonymity in social networks and develop a new re-identification algorithm targeting anonymized social-network graphs. To demonstrate its effectiveness on real-

associated with individual nodes are suppressed. Such suppression is often misinterpreted as removal of "personally identifiable information" (PII), even though PII may include much more than names and identifiers (see the discussion in Appendix B). For example, the EU privacy directive defines "personal data" as "any information relating to an identified or identifiable natural person [...]; an identifiable person is one who can be identified, directly or indirectly, in particular by
one or more fac
mental, economi

take two graphs representing social networks and map the nodes to each other based on the *graph structure alone* —no usernames, no nothing

Netflix Prize, Kaggle contest

social graphs

An Automated Social Graph De-anonymization Technique

Kumar Sharad
University of Cambridge, UK
kumar.sharad@cl.cam.ac.uk

George Danezis
University College London, UK
g.danezis@ucl.ac.uk

ABSTRACT

We present a generic and automated approach to re-identifying nodes in anonymized social networks which enables novel anonymization techniques to be quickly evaluated. It uses machine learning (decision forests) to matching pairs of nodes in disparate anonymized sub-graphs. The technique uncovers artefacts and in-

Social network graphs in particular are high dimensional and feature rich data sets, and it is extremely hard to preserve their anonymity. Thus, any anonymization scheme has to be evaluated in detail, including those with a sound theoretical basis [11]. Techniques have been proposed to resist de-anonymization [8, 17, 22], however, Dwork and Naor have shown [7] that preserving privacy of

Let's take one example: Anonymity

2) Link two records within a dataset (or datasets)

De-anonymizing Social Networks

Arvind Narayanan and Vitaly Shmatikov
The University of Texas at Austin

Abstract

Operators of online social networks are increasingly sharing potentially sensitive information about users and their relationships with advertisers, application developers, and data-mining researchers. Privacy is typically protected by anonymization, i.e., removing names, addresses, etc.

We present a framework for analyzing privacy and anonymity in social networks and develop a new re-identification algorithm targeting anonymized social-network graphs. To demonstrate its effectiveness on real-

associated with individual nodes are suppressed. Such suppression is often misinterpreted as removal of "personally identifiable information" (PII), even though PII may include much more than names and identifiers (see the discussion in Appendix B). For example, the EU privacy directive defines "personal data" as "any information relating to an identified or identifiable natural person [...]; an identifiable person is one who can be identified, directly or indirectly, in particular by
one or more fac
mental, economi

social graphs

An Automated Social Graph De-anonymization Technique

Kumar Sharad
University of Cambridge, UK
kumar.sharad@cl.cam.ac.uk

George Danezis
University College London, UK
g.danezis@ucl.ac.uk

ABSTRACT

We present a generic and automated approach to re-identifying nodes in anonymized social networks which enables novel anonymization techniques to be quickly evaluated. It uses machine learning (decision forests) to matching pairs of nodes in disparate anonymized sub-graphs. The technique uncovers artefacts and in-

Social network graphs in particular are high dimensional and feature rich data sets, and it is extremely hard to preserve their anonymity. Thus, any anonymization scheme has to be evaluated in detail, including those with a sound theoretical basis [11]. Techniques have been proposed to resist de-anonymization [8, 17, 22], however, Dwork and Naor have shown [7] that preserving privacy of

DE GRUYTER OPEN   Proceedings on Privacy Enhancing Technologies ; 2016 (3):155–171

Rebekah Overdorf* and Rachel Greenstadt

Blogs, Twitter Feeds, and Reddit Comments: Cross-domain Authorship Attribution

Abstract: Stylometry is a form of authorship attribution that relies on the linguistic information to attribute

curity by serving as a verification or identification tool for digital text across the Internet.

As social media and micro-blogging sites increase in popularity, so does the need to identify the authors of these types of text. The accuracy with which stylometry can identify anonymous and pseudonymous authors has direct security implications. It can be used for verification of a person's claimed identity, or to identify the author of an anonymous threat should a suspect set be

Doppelgänger Finder: Taking Stylometry To The Underground

Sadia Afroz*, Aylin Caliskan-Islam†, Ariel Stolerman†, Rachel Greenstadt† and Damon McCoy‡
*University of California, Berkeley   †Drexel University   ‡George Mason University

Abstract—**Stylometry is a method for identifying anonymous authors of anonymous texts by analyzing their writing style. While stylometric methods have produced impressive results in previous experiments, we wanted to explore their performance on a challenging dataset of particular interest to the security research community. Analysis of underground forums can provide key information about who controls a given bot network or sells a service, and the size and scope of the cybercrime**

Other information gleaned from underground forums is providing security researchers, law enforcement, and policy makers valuable information on how the market is segmented and specialized, the social dynamics of the community, and potential bottlenecks that are vulnerable to interventions. These advances have been accomplished primarily through

Link messages from same person with different pseudonyms

stylometry

Let's take one example: Anonymity

2) Link two records within a dataset (or datasets)

De-anonymizing Social Networks

Arvind Narayanan and Vitaly Shmatikov
The University of Texas at Austin

Abstract

Operators of online social networks are increasingly sharing potentially sensitive information about users and their relationships with advertisers, application developers, and data-mining researchers. Privacy is typically protected by anonymization, i.e., removing names, addresses, etc.

We present a framework for analyzing privacy and anonymity in social networks and develop a new re-identification algorithm targeting anonymized social-network graphs. To demonstrate its effectiveness on real-

associated with individual nodes are suppressed. Such suppression is often misinterpreted as removal of "personally identifiable information" (PII), even though PII may include much more than names and identifiers (see the discussion in Appendix B). For example, the EU privacy directive defines "personal data" as "any information relating to an identified or identifiable natural person […]; an identifiable person is one who can be identified, directly or indirectly, in particular by one or more fac mental, economi

social graphs

An Automated Social Graph De-anonymization Technique

Kumar Sharad
University of Cambridge, UK
kumar.sharad@cl.cam.ac.uk

George Danezis
University College London, UK
g.danezis@ucl.ac.uk

ABSTRACT

We present a generic and automated approach to re-identifying nodes in anonymized social networks which enables novel anonymization techniques to be quickly evaluated. It uses machine learning (decision forests) to matching pairs of nodes in disparate anonymized sub-graphs. The technique uncovers artefacts and in-

Social network graphs in particular are high dimensional and feature rich data sets, and it is extremely hard to preserve their anonymity. Thus, any anonymization scheme has to be evaluated in detail, including those with a sound theoretical basis [11]. Techniques have been proposed to resist de-anonymization [8, 17, 22], however, Dwork and Naor have shown [7] that preserving privacy of

Authorship attribution also works across domains!!

DE GRUYTER OPEN Proceedings on Privacy Enhancing Technologies ; 2016 (3):155–171

Rebekah Overdorf* and Rachel Greenstadt

Blogs, Twitter Feeds, and Reddit Comments: Cross-domain Authorship Attribution

Abstract: Stylometry is a form of authorship attribution that relies on the linguistic information to attribute

curity by serving as a verification or identification tool for digital text across the Internet.

As social media and micro-blogging sites increase in popularity, so does the need to identify the authors of these types of text. The accuracy with which stylometry can identify anonymous and pseudonymous authors has direct security implications. It can be used for verification of a person's claimed identity, or to identify the

Link messages from same person with different pseudonyms

Doppelgänger Finder: Taking Stylometry To The Underground

Sadia Afroz*, Aylin Caliskan-Islam†, Ariel Stolerman†, Rachel Greenstadt† and Damon McCoy‡
*University of California, Berkeley †Drexel University ‡George Mason University

***Abstract*—Stylometry is a method for identifying anonymous authors of anonymous texts by analyzing their writing style. While stylometric methods have produced impressive results in previous experiments, we wanted to explore their performance on a challenging dataset of particular interest to the security research community. Analysis of underground forums can provide key information about who controls a given bot network**

Other information gleaned from underground forums is providing security researchers, law enforcement, and policy makers valuable information on how the market is segmented and specialized, the social dynamics of the community, and potential bottlenecks that are vulnerable to interventions. These advances have been accomplished primarily through

stylometry

“Anti-surveillance PETs” technical goals

privacy properties: **Anonymity**

3) infer information about an individual

Inference Attacks on Location Tracks

John Krumm

Microsoft Research
One Microsoft Way
Redmond, WA, USA
jckrumm@microsoft.com

Abstract. Although the privacy threats and countermeasures associated with location data are well known, there has not been a thorough experiment to assess the effectiveness of either. We examine location data gathered from volunteer subjects to quantify how well four different algorithms can identify

“Based on GPS tracks from, we identify the latitude and longitude of their homes. From these locations, we used a free Web service to do a reverse “white pages” lookup, which takes a latitude and longitude coordinate as input and gives an address and name. [172 individuals]”

Let's take one example: Anonymity

3) infer information about an individual

Inference Attacks on Location Tracks

John Krumm

Microsoft Research
One Microsoft Way
Redmond, WA, USA
jckrumm@microsoft.com

Abstract. Although the privacy threats and countermeasures associated with location data are well known, there has not been a thorough experiment to assess the effectiveness of either. We examine location data gathered from volunteer subjects to quantify how well four different algorithms can identify

"We investigate the subtle cues to user identity that may be exploited in attacks on the privacy of users in web search query logs. We study the application of simple classifiers to map a sequence of queries into the gender, age, and location of the user issuing the queries."

"I Know What You Did Last Summer" — Query Logs and User Privacy

Rosie Jones Ravi Kumar Bo Pang Andrew Tomkins
Yahoo! Research, 701 First Ave, Sunnyvale, CA 94089.
{jonesr,ravikumar,bopang,atomkins}@yahoo-inc.com

ABSTRACT

We investigate the subtle cues to user identity that may be exploited in attacks on the privacy of users in web search query logs. We study the application of simple classifiers to map a sequence of queries into the gender, age, and location of the user issuing the queries. We then show how these classifiers may be carefully combined at multiple granularities to map a sequence of queries into a

bilities; this is the goal of this paper. We initiate the study of subtle cues to user identity that exist as vulnerabilities in web search query logs, which may be exploited in attacks on the privacy of users.

Privacy attack models. We begin with a characterization of two key forms of attack against which a query log privacy scheme must be resilient. The first is a *trace attack*, in which an attacker studies a privacy-enhanced version of a sequence of searches (*trace*) made

Let's take one example: **Anonymity**

Wishful thinking!

this cannot happen in general!

Data anonymization is a **weak privacy mechanism**

Impossible to sanitize without severely damaging usefulness

Removing PII is not enough! – Any aspect could lead to re-identification

Art. 29 WP's opinion :

Risk of de-anonymization? Probabilistic Analysis

$Pr[\text{identity} \rightarrow \text{action} \mid \text{observation}]$

Privacy evaluation is a **Probabilistic analysis**

systematic reasoning to evaluate a mechanism

Anonymity – Pr[identity → action | observation]

Unlinkability – Pr[action A ↔ action B | observation]

Obfuscation – Pr[real action | observed noisy action]

1) Model the privacy-preserving mechanism as a probabilistic transformation

2) Determine what the adversary will see

data

metadata

...

Privacy evaluation is a **Probabilistic analysis**

systematic reasoning to evaluate a mechanism

Anonymity – Pr[identity → action | observation]

Unlinkability – Pr[action A ↔ action B | observation]

Obfuscation – Pr[real action | observed noisy action]

1) Model the privacy-preserving mechanism as a probabilistic transformation

If it is not probabilistic, it is not secure

2) Determine what the adversary will see

3) "Invert" the mechanism as the adversary would do

the adversary knows!!!

4) Compute probability after "inversion"

5) Measure… mean error, entropy (any Flavour), Diff. Privacy

"Inversion"? What do you mean?

1) Analytical mechanism inversion

Given the description of the system, develop the mathematical expressions that effectively invert the system:

Pr[obs | real data, PET] → Pr[Real data | obs, PET]

Not always possible – May require aprox. or sampling

2) Machine learning (data driven)

Train a classifier to break the mechanisms!

Only possible if enough data (though data can be created)

Must take Inversion into account!! Systematic design!!!

That's another talk.....

Take aways

Realizing Privacy by design is non-trivial

PART I:
Reasoning about Privacy when designing systems

Explicit privacy engineering activities

Fully fledged methodology?

Requirements? Evaluation?

Training on PETS (Universities are there!)

Understanding & Implementation

PART II:
Evaluating Privacy in Privacy-Preserving systems

Systematic reasoning for privacy evaluation

Assumption's dependency

No known generic methods

More training!

THANKS!

Any questions?

carmela.troncoso@imdea.org
https://software.imdea.org/~carmela.troncoso/
(these slides will be there soon)

From the 1st of November
Assistant Professor at

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Template: http://www.brainybetty.com/
Figures: SlidesCarnival