Privacy Technologies

Claudia Diaz
K.U.Leuven COSIC
Outline

- Perspectives of privacy
- Privacy models
  - Trust-based privacy (data protection)
  - Hard privacy (PETs)
- Anonymous authentication
- Protection against traffic analysis
- Identification + data minimization
- Database privacy
- Conclusions
Perspectives on privacy

- Popular definitions:
  - “The right to be let alone”
  - “Informational self-determination”
  - “The freedom from unreasonable constraints on the construction of one's own identity”

- Solove:
  - identifies 16 privacy threats relating to information collection, processing and dissemination, and invasion

- Technical privacy properties:
  - Anonymity, Pseudonymity, Unlinkability, Unobservability, Plausible deniability (OTR), Location privacy…
Data protection

- Data collected for specific and legitimate purposes
- Proportional: adequate, relevant and not excessive (data minimization)
- With the subject’s awareness and consent
- Data subject’s right to access, correct, delete her data
- Data security
  - Integrity, confidentiality of the data
- Identified or identifiable person -- does not apply to anonymous data
Privacy as Security

- Privacy as *informational self-determination*
- Gaining control over one’s informational environment
- Giving out less information
- Minimizing the need to trust others to behave according to our best interests

- These are the goals of **Privacy Enhancing Technologies**
Privacy models
Trust-based privacy

• System model
  • Data subject provides her data
  • Data controller responsible (trusted) for its protection
    • One or several data processors

• Threat model
  • External parties, errors, malicious insider
Trust-based privacy

- Controller/processors: main “users” of security technologies
- Policies, access control, audits (liability)
Trust-based privacy

- Data subject has already lost control of her data
  - In practice, very difficult for data subject to verify how her data is collected and processed
Trust-based privacy

- Data subject has already lost control of her data
  - In practice, very difficult for data subject to verify how her data is collected and processed
  - Need to trust data controllers (honesty, competence) and hope for the best
Darling admits Revenue loss of 25 million personal records

Lost: Two discs, 25 million accounts

By John Oates • Get more from this author

Posted in Government, 20th November 2007 18:22 GMT

UK Identity Crisis  Alistair Darling told the House of Commons this afternoon that a police investigation has been launched into how Her Majesty's Revenue and Customs has lost child benefit records relating to 25 million people.

Records for 25 million people, relating to child benefit payments for 7.25 million families, were sent using the HMRC's own postal system, called grid, but never arrived.

The Chancellor, flanked by PM Gordon Brown, told the House that the National Audit Office requested information which was first sent to them in March, in breach of HMRC procedures, and then returned to HMRC.
Clarkson stung after bank prank

TV presenter Jeremy Clarkson has lost money after publishing his bank details in his newspaper column.

The Top Gear host revealed his account numbers after rubbing off the fuss over the loss of 25 million people's personal details on two computer discs.

He wanted to prove the story was a fuss about nothing.

But Clarkson admitted he was "wrong" after he discovered a reader had used the details to create a £500 direct debit to the charity Diabetes UK.

Clarkson published details of his Barclays account in the Sun newspaper, including his account number and sort code. He even told people how to find out his address.

"All you'll be able to do with them is put money into my account. Not take it out. Honestly, I've never known such a palaver about nothing," he told readers.

But he was proved wrong, as the 47-year-old wrote in his Sunday Times column.
Problems of trust-based privacy

- Data minimization (proportionality) often ignored
- Informed consent?
- Trust assumptions may not be realistic
  - Incompetence
  - Malicious insiders
  - Incentives?
  - Purpose (function creep)
  - Cost of securing the data
- How can you check that your data is not being abused?
- Weak enforcement, low penalties
Problems of trust-based privacy

- Technologically enforced?
  - Like security, privacy must be technologically supported
  - Privacy/security needs cannot just be satisfied with good intentions.
  - Laws are necessary but not sufficient to protect privacy/security.
  - Technology must provide assurances where possible
    - Examples: legal interception, data retention
Other problems

- What others reveal about us
- consent?
Hard Privacy (PETs)

- System model
  - Subject provides as little data as possible
  - Reduce as much as possible the need to “trust” other entities
- Threat model
  - Strategic adversary with certain resources motivated to breach privacy (similar to security systems)
  - Adversarial environment: communication provider, data holder
Hard Privacy (PETs)

- Subject is an active security “user”
- Goal (data protection): data minimization
Two main approaches

- **Anonymity**
  - Service provider can observe access to the service
  - Cannot observe the identity of the user

- **Oblivious Transfer (OT) / Private Information Retrieval (PIR)**
  - Service provider can identify user
  - Cannot observe details of the access to the service
    - Which records were accessed
    - Which search keywords were used
    - Which content was downloaded
    - ...

- All parties have assurance that the other participants in the protocol are cannot cheat
PETs to achieve anonymity
Authentication

- Entity authentication often first step of a transaction

- Makes sense in an organizational environment (government, military, even commercial)
  - …but what if there is no closed group?
  - The **Identity Management** concept

- Possible solutions:
  - Private authentication: hide against 3\textsuperscript{rd} parties (Just Fast Keying)
  - Anonymous credentials: protect against everybody
Idea behind credentials

- Many transactions involve attribute certificates
  - ID docs: state certifies name, birth dates, address
  - Letter reference: employer certifies salary
  - Club membership: club certifies some status

- Do you want to show all attributes for each transaction?

- Credential: token certifying attributes
  - Prover proves to the Verifier that she holds a credential with certain properties certified by the Issuer
Properties

- Cryptographic protocols between <Issuer, Prover, Verifier>
  - Prover can prove that he holds a credential with certain attributes
  - or any expression on them (simple arithmetic, boolean) (e.g. salary > 30,000 and contract = permanent)

- Unforgeability and Privacy
- Verifier gains no more information: One party proves to another that a statement is true, without revealing anything other than the veracity of the statement.
- Secure even if Issuer and Verifier collude (single/multiple show)
- Security: cryptographic (Hard Privacy)
## PKI vs Anonymous Credentials

<table>
<thead>
<tr>
<th>PKI</th>
<th>Anonymous credentials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signed by a trusted issuer</td>
<td>Signed by a trusted issuer</td>
</tr>
<tr>
<td>Certification of attributes</td>
<td>Certification of attributes</td>
</tr>
<tr>
<td>Authentication (secret key)</td>
<td>Authentication (secret key)</td>
</tr>
<tr>
<td>Double-signing detection</td>
<td>Double-signing detection</td>
</tr>
<tr>
<td>No data minimization</td>
<td>Data minimization</td>
</tr>
<tr>
<td>Users are identifiable</td>
<td>Users are anonymous</td>
</tr>
<tr>
<td>Users can be tracked</td>
<td>Users are unlinkable in different contexts</td>
</tr>
<tr>
<td>(Signature linkable to other contexts where PK is used)</td>
<td></td>
</tr>
</tbody>
</table>
Types of anonymous credentials

• Brands:
  • “Minimal disclosure tokens”
  • One-show
  • Credentica – uProve (Microsoft, Card Space)

• Camenish-Lysyanskaya
  • Multi-show (detect misbehaviour)
  • Less efficient
  • Idemix (IBM) - Free source? … the patents war

Future identity cards and passports?
Anonymous e-cash

- Secure and private payments
  - Cannot forge money or payments
  - with the anonymity of cash
  - Not just cash: cinema tickets

- Anonymous credentials can provide this
  - The bank certifies I have one euro
  - Payment: prover shows the credential, verifier accepts it
  - Verifier goes to the bank to deposit the coin

- Security properties:
  - Unforgeability
  - Privacy (for payer)
  - Double spending prevention!
Example application: e-petitions

- Formal requests addressed to an authority and signed by numerous individuals
- Typically citizens provide
  - Unique identifier (name, national ID number)
  - Signature
- Verification:
  - Validating that the signatures correspond to the identifiers
  - Discarding multiple/invalid signatures
- Benefits of going electronic:
  - Many resources are needed in order to physically collect the signatures
  - Manual signature verification is a costly and tedious process
The straightforward e-petition implementation

- Have users sign the petitions with their e-ID
  1. Select petition
  2. Sign using the e-ID (2-factor authentication)
  3. Check that the petition has not yet been signed with that e-ID
  4. Count (or discard) the signature

- Privacy risks
  - Leak sensitive information on political beliefs, religious inclinations, etc.
  - Through unique identifiers, petition signatures can be linked to other data
e-petition requirements

- Basic requirements
  - Authentication: citizen is who claims to be (i.e., no impersonation)
  - Required attributes: citizen is entitled to sign (e.g., age $\geq 18$ and nationality $\in$ EU)
  - Uniqueness: citizens sign a petition only once
  - Correctness: all valid signatures are counted

- Privacy requirements
  - Citizen unlinkable to petition (i.e., not possible to identify who are the signers)
Citizen Petition server Credential show + Petition signing

credential issuing protocol

Credential issuer Petition server

Count signature and store transcript

Yes

Multiple signing?

No
Properties

- Only citizens entitled to sign can do so
  - Possession of e-ID + knowledge of PIN
  - Attribute verification (e.g., age, locality)
  - One credential per citizen

- Citizens can sign only once (multiple signing is detectable so that repeated signatures can be deleted)

- Collusion of credential issuer and e-Petition server does not reveal the identity of a signer

- Need for anonymous communication channel to preserve privacy properties
Protection against traffic analysis
Communication infrastructure

- Applications assume that the **communication** channels are secured / maintain privacy properties
  - Example: previous protocols are useless if the adversary can link transactions based on traffic data (e.g., IP address)
- Private channels
- Data confidentiality and integrity: same as traditional security
- Confidentiality of identities (**anonymity**) and relations (**unlinkability**):
  - Cryptographically: credential protocols
  - Network: protection against traffic analysis
  - The infrastructure is **shared** by individuals, business, government, military, etc: privacy threats affect all
Anonymous communications

- Anonymity / unlinkability **not** provided by default by the communication infrastructure
- **Traffic** data (origin, destination, time, volume): side channel information
  - Less volume than content: coarser, but highly valuable information
  - Formats that are easy to process for machines
  - Can be used to select targets for more intensive surveillance
  - Hard to conceal
- Adversarial:
  - **Third party** with access to the communication channels
  - **Recipient**: adversarial or trusted (subject can authenticate over the anonymous channel)
Anonymous communications: abstract model

- Objective: hide the identity of the sender (or receiver, or both)
- Make the bit patterns of inputs and outputs different (bitwise unlinkability)
- Destroy the timing characteristics (traffic analysis resistance)
Basic Anonymity Properties

- 3rd party anonymity
  - Alice and Bob trust each other but do not want other parties to learn that they are communicating

- Sender anonymity
  - Alice sends to Bob, and Bob cannot trace Alice’s identity

- Receiver Anonymity
  - Bob can contact Alice, without knowing her identity.

- Bi-directional Anonymity
  - Alice and Bob communicate without knowing each other’s identities.
Systems for anonymous communications

- Theoretical / Research
  - Mix networks (1981)
  - DC-networks (1985)
  - ISDN mixes (1992)
  - Onion Routing (1996)
  - Crowds (1998)

- Real world systems
  - Single proxy (90s): anon.penet.fi, Anonymizer, SafeWeb
  - Remailers: Cipherpunk Type 0, Type 1, Mixmaster (1994), Mixminion (2003)
Attacks against anonymity systems

- Traffic Analysis: against vanilla or hardened systems
  - Extract information out of patterns of traffic (no content)
- Many adversary models are possible and realistic
- Hard to protect
  - Traffic correlation / confirmation
  - Long-term intersection attacks
  - Predecessor attack (random routing)
  - Sybil
Steganography and covert communications

- **Encryption**: hide data content
- **Anonymity/unlinkability**: hide identities / relations
- **Unobservability**: hide existence
- **Communications**:
  - Hide the fact that there is any communications
  - Embed a communication within another
  - **Covert channels**: hide secrets within public information
- **Storage**:
  - Hide the existence of files
  - Under coercion can deny there are any files to decrypt
Identification + data minimization
Oblivious Transfer (OT)

- A inputs two information items, B inputs the index of one of A’s items
- B learns his chosen item, A learns nothing
  - A does not learn which item B has chosen;
  - B does not learn the value of the item that he did not choose
- Generalizes M instead of 2, etc.
- Example: retrieving location-based content
Buying digital content

- Identify customer, but conceal which information item is retrieved
- Pre-paid system
Private Search

- Alice stores documents
- Bob wants to retrieve documents matching some keywords
- Properties:
  - Bob gets documents containing the keywords
  - Alice does not learn Bob’s keywords
  - Alice does not learn the results of the search
Electronic Toll Pricing

- Differentiated payment for mobility: Congestion pricing
  - Users will pay depending on their use of the car and roads

- European Electronic Toll Service (EETS) Decision (Oct 2009)
  - Defines EETS architecture and interfaces
  - Within three years for vehicles above 3.5 tons, all other vehicles within five years.
EETS straightforward implementation
Privacy for Electronic Toll Pricing

- Privacy issues?
  - Pay *as you drive*
  - Fine grained GPS data allows for inferences
    - Medical issues (visits to a Cancer specialized clinic)
    - Political affiliation (visits to the headquarters of a political party)
    - Industry espionage (visits to other companies)

- What data is necessary?
  - Final fee that the user must pay to the provider/government
  - This is the actual purpose of the whole system – and not collecting everyone’s detailed location data
    - Enormous **reduction of risk and cost** by eliminating the need to store all the raw data

- Legal / service integrity issues
  - Actors must not be able to cheat
  - Actors must be held liable when misusing the system
Privacy-Friendly Electronic Toll Pricing

- No personal data leaves the domain of the user
Enforcement

- OBU in hands of the user
  - Incentive to cheat (paying less)
  - Even if the box is tamper-resistant, the input is easy to spoof
- We need to:
  - Detect vehicles with inactive OBUs
  - Detect vehicles reporting false location data
  - Detect vehicles using incorrect road prices
  - Detect vehicles reporting false final fees
- Combination of law + technology
Non-Interactive Commitment Schemes

<table>
<thead>
<tr>
<th></th>
<th>00u00 – 07u00</th>
<th>22u00 – 00u00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highway</td>
<td>p₁</td>
<td>p₂</td>
</tr>
<tr>
<td>Primary</td>
<td>p₃</td>
<td>p₄</td>
</tr>
<tr>
<td>Residential</td>
<td>p₅, p₆, ...</td>
<td>p₇</td>
</tr>
</tbody>
</table>

Where you at…?

HIDING PROPERTY

BINDING PROPERTY

Slide credit: Carmela Troncoso

10/03/2011
Homomorphic commitments

- The content of the vaults can be added up without being known.
How does it work?

OBU

Does not reveal information about the trajectory

Cannot be changed

Toll Service Provider

01·GBB·1 + Pos + Time + auth

Confirmation

Toll Charger

License Plate Reader

01·GBB·1 Pos + Time

50 Claudia Diaz (K.U.Leuven)

Slide credit: Carmela Troncoso 10/03/2011
What can we prove?

- OBU was active
  - A commitment with the committed location and time must be available

- OBU used correct prices
  - Prices in the table signed by Toll Service Provider
  - Check correct pricing upon commitment opening

- OBU was at reported location
  - Compare photo location with committed location

- OBU made correct operations
  - Homomorphic commitments: prices in the “vaults” can be added to verify that they correspond to the reported final fee without being opened
Performance Analysis

- OBU platform: NXP ARM7 microcontroller (32 bit) / SW
- TSP platform: commodity computer (Inter Core2Duo)

OBU timings and average speed tolerance for a 1-hour journey

<table>
<thead>
<tr>
<th>Security Operation</th>
<th>Medium (1024 bit)</th>
<th>High (1536 bit)</th>
<th>Very High (2048 bit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Map-Matching</td>
<td>839.11 s</td>
<td></td>
<td></td>
</tr>
<tr>
<td>One segment</td>
<td>7.88 s</td>
<td>22.13 s</td>
<td>47.79 s</td>
</tr>
<tr>
<td>Max. Speed</td>
<td>350 km/h</td>
<td>124 km/h</td>
<td>57 km/h</td>
</tr>
</tbody>
</table>

TSP capacity tolerance assuming an average of 1500 km/month/vehicle

<table>
<thead>
<tr>
<th>Security Commit</th>
<th>Medium (1024 bit)</th>
<th>High (1536 bit)</th>
<th>Very High (2048 bit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5 Km</td>
<td>82 000</td>
<td>29 000</td>
<td>14 000</td>
</tr>
<tr>
<td>1 Km</td>
<td>164 000</td>
<td>58 000</td>
<td>29 000</td>
</tr>
<tr>
<td>2 Km</td>
<td>329 000</td>
<td>117 000</td>
<td>58 000</td>
</tr>
</tbody>
</table>

- Communication overhead
  - Sending full GPS: 2.05Mbytes for 1500Km
  - PrETP:
    - 1.5Kbytes per segment
    - ~2Mbytes a month
    - Eventual 50Kbytes to open a commitment
Data anonymization

- Anonymized data can be very useful, for example, for research purposes
  - Incidence of diseases: medical research
  - Social network structures: epidemiology, sociology
  - Optimization of services (e.g., transport or computer infrastructures)
- Measure the risk of **re-identification** of anonymized data:
  - Records in an anonymized database
    - Medical data
    - Internet searches (AOL case)
    - Movie ratings (Netflix)
- Note: data protection does not apply to anonymized data
  - Often, we hear unsubstantiated claims of “anonymization”
K-anonymity

- Removing obvious identifiers (e.g., name) is not enough:
- “The triple (date of birth, gender, zip code) suffices to uniquely identify at least 87% of US citizens in publicly available databases (1990 U.S. Census summary data).” [Swe]
- Sets of attributes constitute Quasi Identifiers (Qis)

<table>
<thead>
<tr>
<th>Hospital Patient Data</th>
<th>Vote Registration Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DOB</strong></td>
<td><strong>Sex</strong></td>
</tr>
<tr>
<td>1/21/76</td>
<td>Male</td>
</tr>
<tr>
<td>4/13/86</td>
<td>Female</td>
</tr>
<tr>
<td>2/28/76</td>
<td>Male</td>
</tr>
<tr>
<td>1/21/76</td>
<td>Male</td>
</tr>
<tr>
<td>4/13/86</td>
<td>Female</td>
</tr>
</tbody>
</table>
K-anonymity

- Use suppression and generalization to ensure that each record in a database is indistinguishable from k-1 other records

- Example:

  **Release Table**

<table>
<thead>
<tr>
<th>Race</th>
<th>Birth</th>
<th>Gender</th>
<th>ZIP</th>
<th>Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>1965</td>
<td>m</td>
<td>0214*</td>
<td>short breath</td>
</tr>
<tr>
<td>Black</td>
<td>1965</td>
<td>m</td>
<td>0214*</td>
<td>chest pain</td>
</tr>
<tr>
<td>Black</td>
<td>1965</td>
<td>f</td>
<td>0213*</td>
<td>hypertension</td>
</tr>
<tr>
<td>Black</td>
<td>1964</td>
<td>f</td>
<td>0213*</td>
<td>hypertension</td>
</tr>
<tr>
<td>Black</td>
<td>1964</td>
<td>f</td>
<td>0213*</td>
<td>obesity</td>
</tr>
<tr>
<td>White</td>
<td>1964</td>
<td>m</td>
<td>0213*</td>
<td>chest pain</td>
</tr>
<tr>
<td>White</td>
<td>1964</td>
<td>m</td>
<td>0213*</td>
<td>obesity</td>
</tr>
<tr>
<td>White</td>
<td>1964</td>
<td>m</td>
<td>0213*</td>
<td>short breath</td>
</tr>
<tr>
<td>White</td>
<td>1967</td>
<td>m</td>
<td>0213*</td>
<td>chest pain</td>
</tr>
<tr>
<td>White</td>
<td>1967</td>
<td>m</td>
<td>0213*</td>
<td>chest pain</td>
</tr>
</tbody>
</table>

  **External Data Source**

<table>
<thead>
<tr>
<th>Name</th>
<th>Birth</th>
<th>Gender</th>
<th>ZIP</th>
<th>Race</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andre</td>
<td>1964</td>
<td>m</td>
<td>02135</td>
<td>White</td>
</tr>
<tr>
<td>Beth</td>
<td>1964</td>
<td>f</td>
<td>55410</td>
<td>Black</td>
</tr>
<tr>
<td>Carol</td>
<td>1964</td>
<td>f</td>
<td>90210</td>
<td>White</td>
</tr>
<tr>
<td>Dan</td>
<td>1967</td>
<td>m</td>
<td>02174</td>
<td>White</td>
</tr>
<tr>
<td>Ellen</td>
<td>1968</td>
<td>f</td>
<td>02237</td>
<td>White</td>
</tr>
</tbody>
</table>

Figure 2 Example of k-anonymity, where k=2 and QI=\{Race, Birth, Gender, ZIP\}
Differential privacy

- k-anonymity
  - Privacy guarantees are 'uncertain'
  - l-diversity, t-closeness, background information?
- Statistical disclosure control. (Dalenius ‘77)
  - “Access to the DB should not allow to learn anything more about an individual than if it had not been accessed”
- Differential Privacy. (Dwork ‘06)
  - Provides a general impossibility result showing that a formalization of Dalenius’ goal along the lines of semantic security cannot be achieved.
  - “The inclusion of an individual’s record should not make much of a difference to the inference”
  - The risk of a privacy breach is not increased by participating in the database
  - Privacy “budget”: DB stops answering queries when the privacy budget is consumed
  - Property holds for arbitrary adversarial background information
Off-The-Record (OTR) security

- Examples: Briefing a journalist, talking on the phone to your lawyer or friends.
- Still want Authenticity, Confidentiality and Integrity.
- **Plausible Deniability** (not non-repudiation): no one can prove you said something.
- **Forward secrecy**: once the communication is securely over, I cannot decrypt it any more (ephemeral keys)
  - Minimize consequences of security breach
  - Compulsion

Privacy challenges

- Privacy requirements and privacy by design
  - Privacy protection needed at all layers
- Finding robust and secure mechanisms
  - Proposed techniques keep on getting broken
  - Secure implementation is even harder
- Usability issues: ease of use, performance
- Economic incentives: tradeoffs privacy/cost (overhead, usability)
- Awareness and transparency
New challenging scenarios

- Location privacy
  - Real time
  - Space-Time relation
  - Device fingerprinting
- Ubiquitous environments
  - Principle of data maximization
  - Constrained devices
  - Securing the physical link
- Social networks: tension with data sharing
  - Ongoing development of SNS plug-in for content confidentiality
- Cloud computing: outsourcing of storage/computations
Conclusions

- Privacy is not “opposed” to security, but rather a security property
- PETs can reconcile aggressive data minimization and service integrity guarantees
- Compliance is a strong driver
- Trust-based privacy is the state of the art
  - Hidden costs of securing the data silos
- Hard Privacy solutions:
  - Active research
  - Poor deployment
Thanks!

http://homes.esat.kuleuven.be/~cdiaz/