Privacy Technologies

Claudia Diaz

K.U.Leuven COSIC

Outline

- Privacy models
  - Data Protection Technologies
  - Privacy Enhancing Technologies
- Anonymity technologies
- Protection against traffic analysis
- Identification +data minimization
- Database privacy
- Other PETs
- Conclusions

Perspectives on privacy

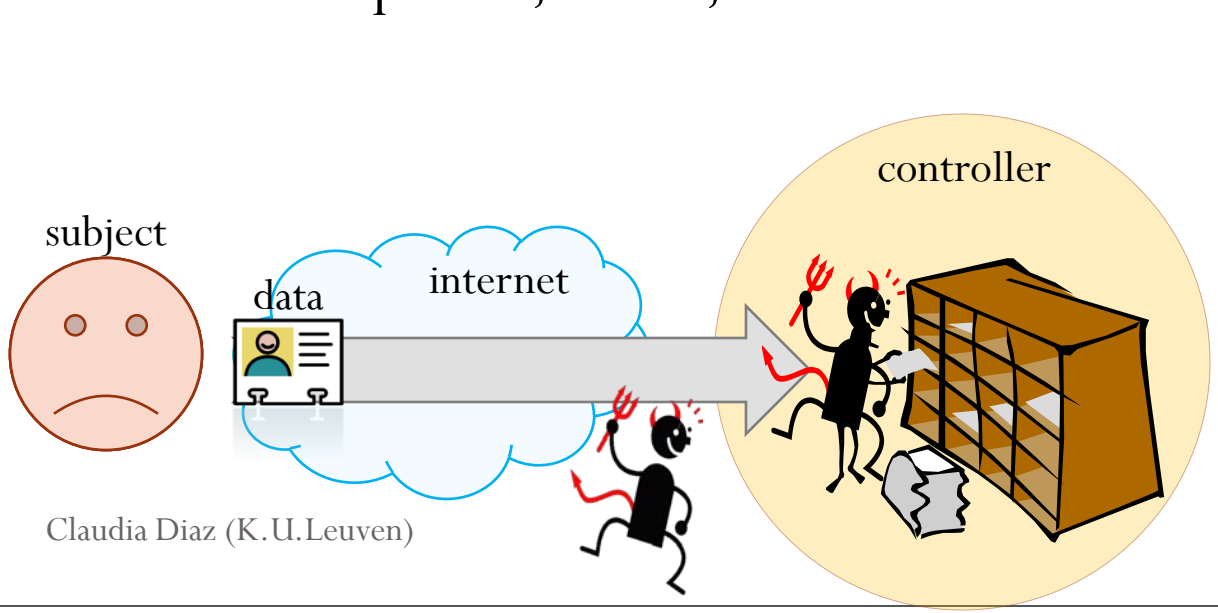
- Popular definitions:
  - "The right to be let alone"
  - "Informational self-determination"
  - "The freedom from unreasonable constraints on the construction of one's own identity"
- Solove:
  - identifies 16 privacy threats relating to information collection, processing, dissemination, and invasion
- Technical privacy properties:
  - Anonymity, Pseudonymity, Unlinkability, Unobservability, Plausible deniability (OTR), Location privacy…

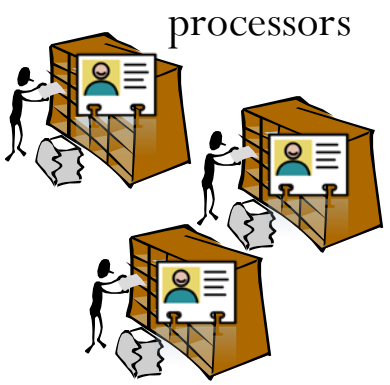
Data protection

- 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

Data protection technologies

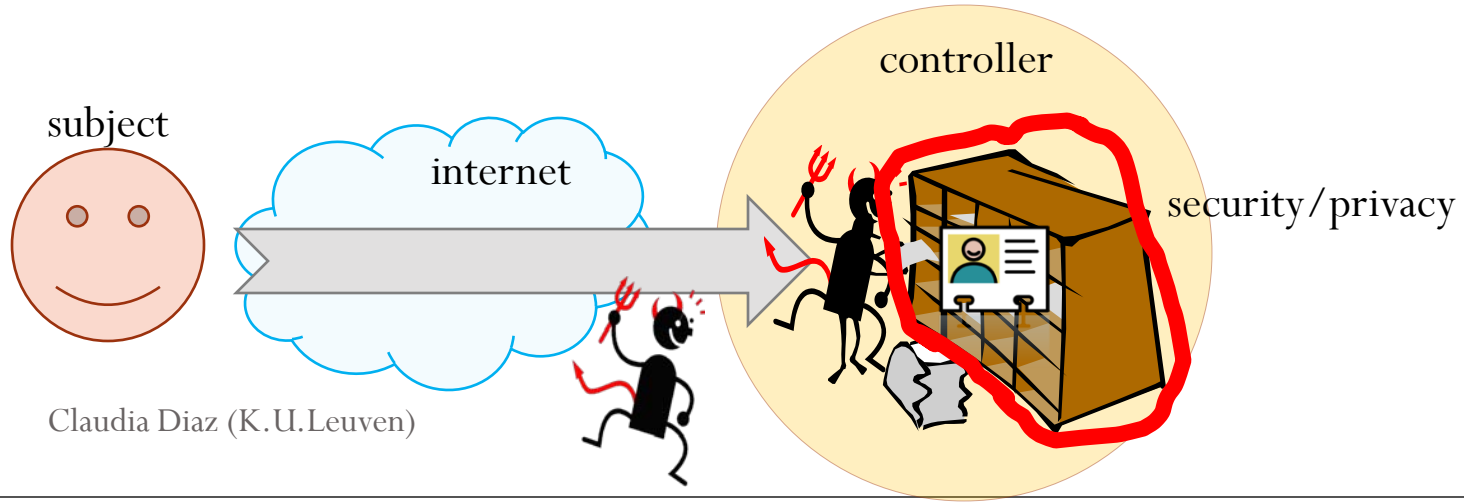
- System model
  - Data subject provides her data
  - Data controller responsible (trusted) for its protection
    - One or several data processors
- Threat model
  - External parties, errors, malicious insider





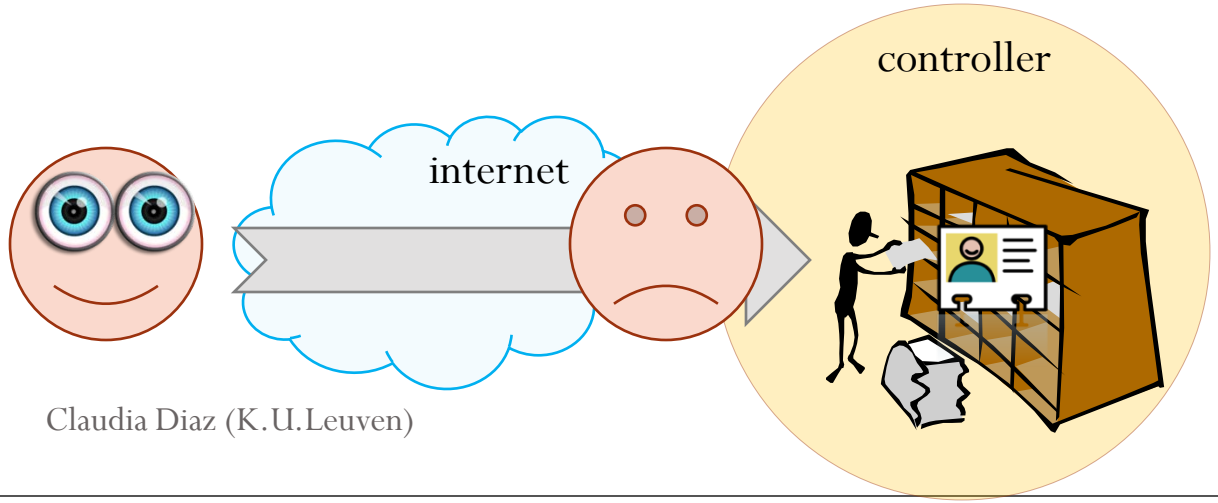
Data protection technologies

- Controller/processors: main "users" of security technologies
- Policies, access control, audits (liability)



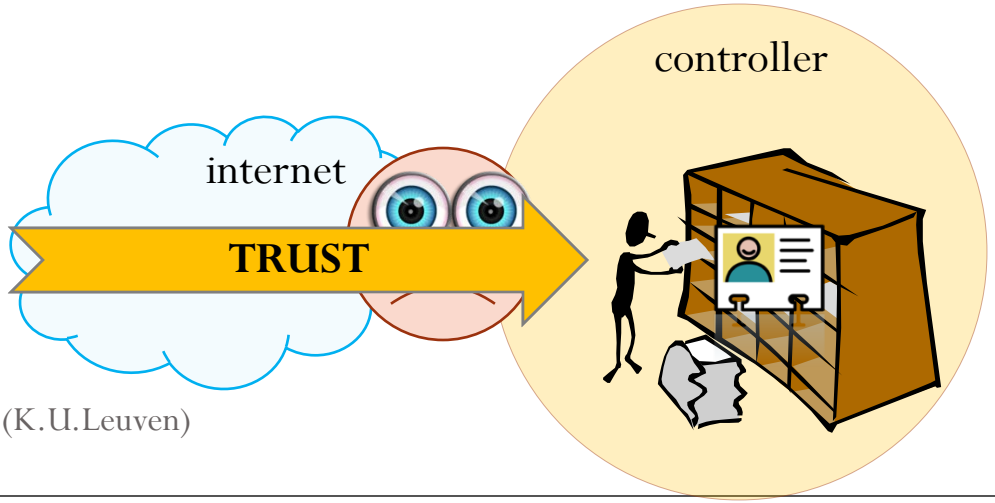
Data protection technologies

- 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



Data protection technologies

- 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



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

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
    - Example: legal interception interface abuse
- How can you check that your data is not being abused?
- Weak enforcement, low penalties
- No protection for "anonymous" data

the Web's Cutting Edge, Anonymity in Name Only

By EMILY STEEL and JULIA ANGWIN

(Please see Corrections & Amplifications item below)

You may not know a company called [x+1] Inc., but it may well know a lot about you.

One Smart Cookie ›

New York ad company [x+1] made predictions about users based on just one click on a website. Read more about the users, see the code transmitted and review the companies' assumptions.

View Interactive

From a single click on a web site, [x+1] correctly identified Carrie Isaac as a young Colorado Springs parent who lives on about $50,000 a year, shops at Wal-Mart and rents kids' videos. The company deduced that Paul Boulifard, a Nashville architect, is childless, likes to travel and buys used cars. And [x+1] determined that Thomas Burney, a Colorado building contractor, is a skier with a college degree and looks like he has good credit.

The company didn't get every detail correct. But its ability to make snap assessments of individuals is accurate enough that Capital One Financial Corp. uses [x+1]'s calculations to instantly decide which credit cards to show first-time visitors to its website.

In short: Websites are gaining the ability to decide whether or not you'd be a good customer, before you tell them a single thing about yourself.

The Web's New Gold Mine: Your Secrets

A Journal investigation finds that one of the fastest-growing businesses on the Internet is the business of spying on consumers. First in a series.

By JULIA ANGWIN

Hidden inside Ashley Hayes-Beaty's computer, a tiny file helps gather personal details about her, all to be put up for sale for a tenth of a penny.

The file consists of a single code— 4c812db292272995e5416a323e79bd37—that secretly identifies her as a 26-year-old female in Nashville, Tenn.

When a user like you logs onto the internet ...

Next

The code knows that her favorite movies include "The Princess Bride," "50 First Dates" and "10 Things I Hate About You." It knows she enjoys the "Sex and the City" series. It knows she browses entertainment news and likes to take quizzes.

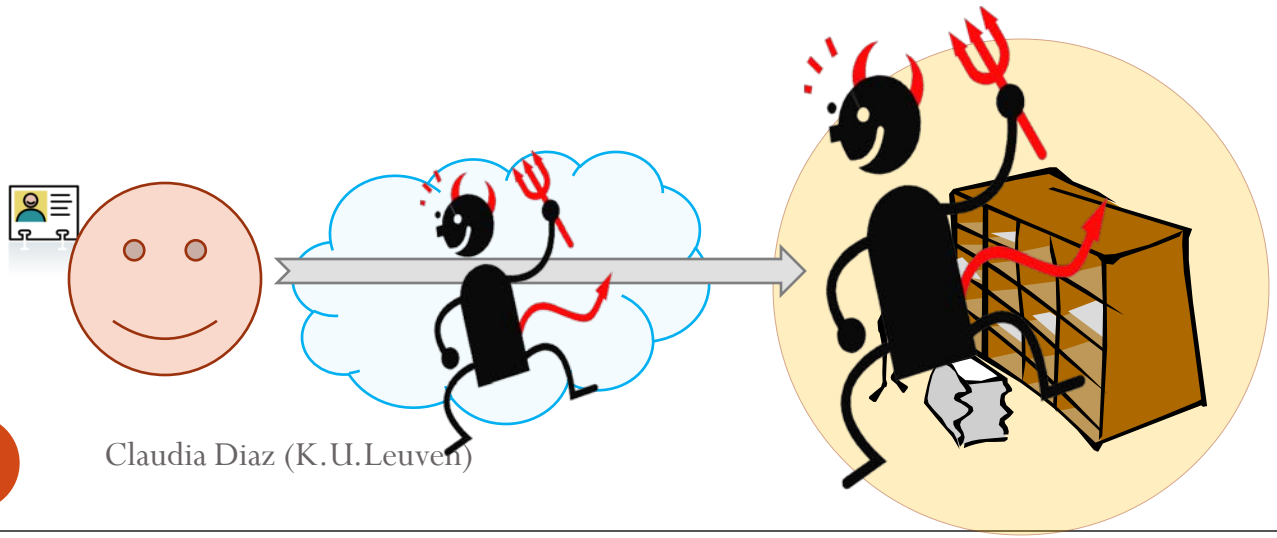
"Well, I like to think I have some mystery left to me, but apparently not!" Ms. Hayes-Beaty said when told what that snippet of code reveals about her. "The profile is eerily correct."

In an interview with WSJ's Alan Murray, WPP CEO Sir Martin Sorrell conceded that advertisers must do better to inform customers about the tracking and mapping of online behavior. On the U.S. economy, he characterized the last 6-7 months as "America Bites Back" but wonders how long the recovery will last.

Ms. Hayes-Beaty is being monitored by Lotame Solutions Inc., a New York company that uses sophisticated software called a "beacon" to capture what people are typing on a website—their comments on movies, say, or their interest in parenting and pregnancy. Lotame packages that data into profiles about individuals, without determining a person's name, and sells the profiles to companies seeking customers. Ms. Hayes-Beaty's tastes can be sold wholesale (a batch of movie lovers is $1 per thousand) or customized (26-year-old Southern fans of "50 First Dates").

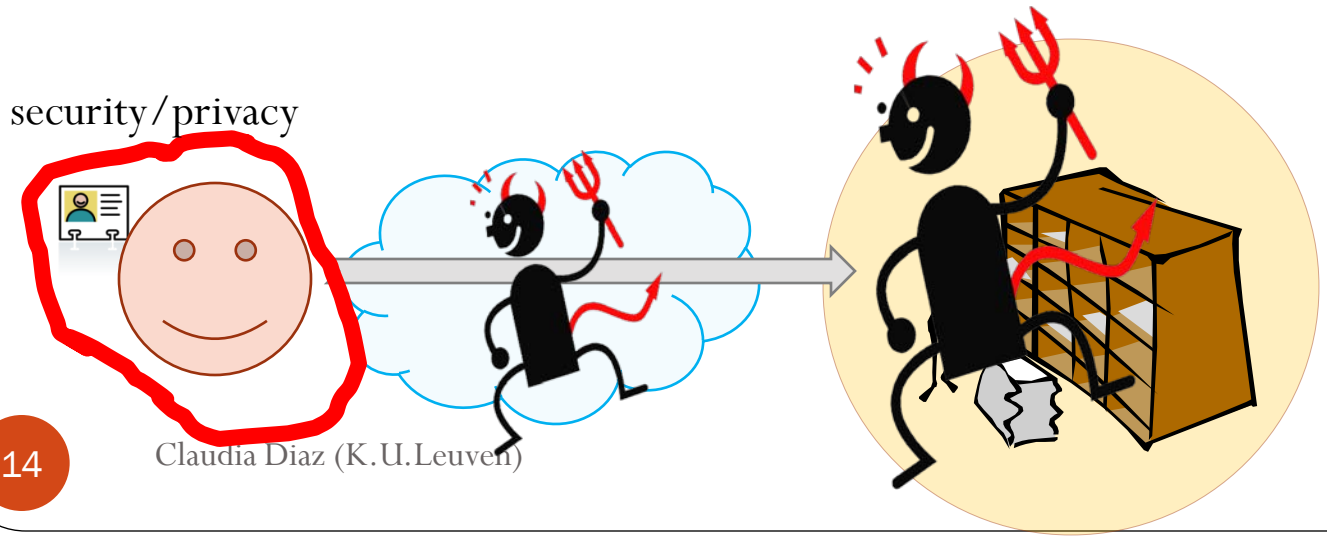
Privacy Enhancing Technologies

- 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



Privacy Enhancing Technologies

- 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 3rd 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

PKI

Signed by a trusted issuer

Certification of attributes

Authentication (secret key)

Double-signing detection

No data minimization

Users are identifiable

Users can be tracked
(Signature linkable to other contexts where PK is used)

Anonymous credentials

Signed by a trusted issuer

Certification of attributes

Authentication (secret key)

Double-signing detection

Data minimization

Users are anonymous

Users are unlinkable in different contexts

Types of anonymous credentials

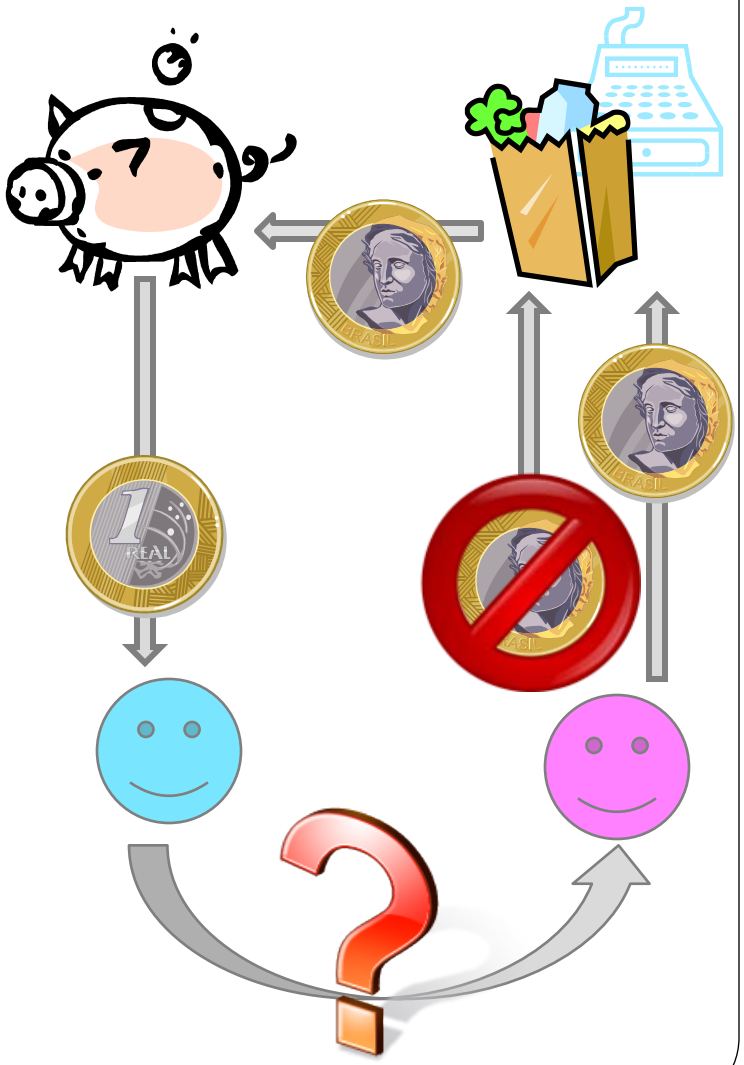
- Brands:
  - "Minimal disclosure tokens"
  - One-show
  - Credentica – uProve (Microsoft, Card Space)

- 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 or transport 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

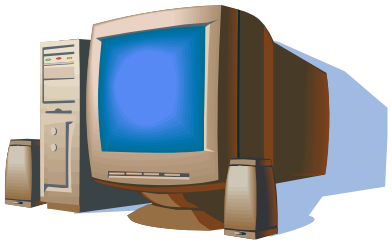
The straightforward e-petition implementation

- 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 ≥ 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)

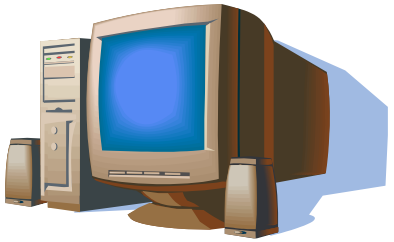
credential issuing protocol



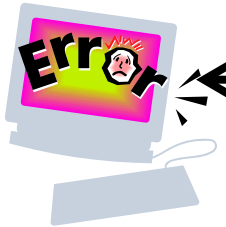
Credential issuer

Citizen

Credential show
+
Petition signing



Petition server



Multiple signing?

Yes

No

Count signature and store transcript

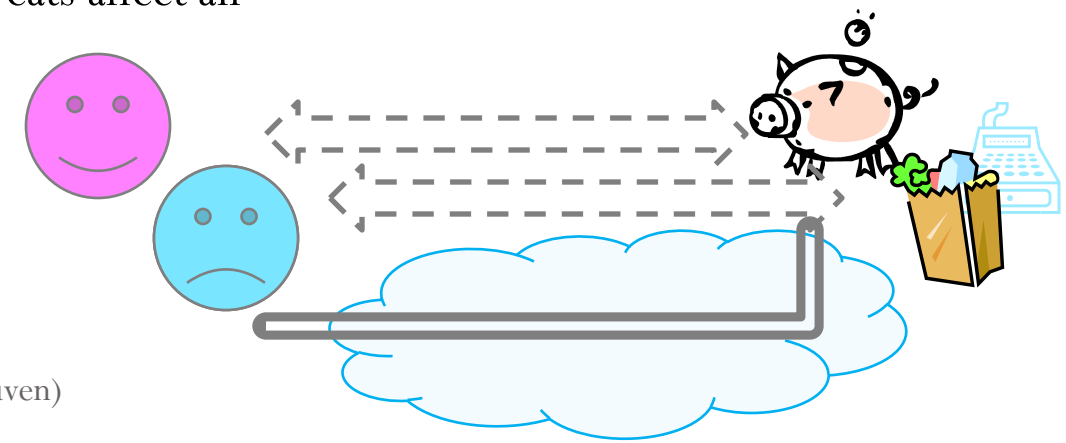
Properties

- 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
  - Hard to conceal
  - Can be used to select targets for more intensive surveillance
  - "Traffic analysis, not cryptanalysis, is the backbone of communications intelligence"
- 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

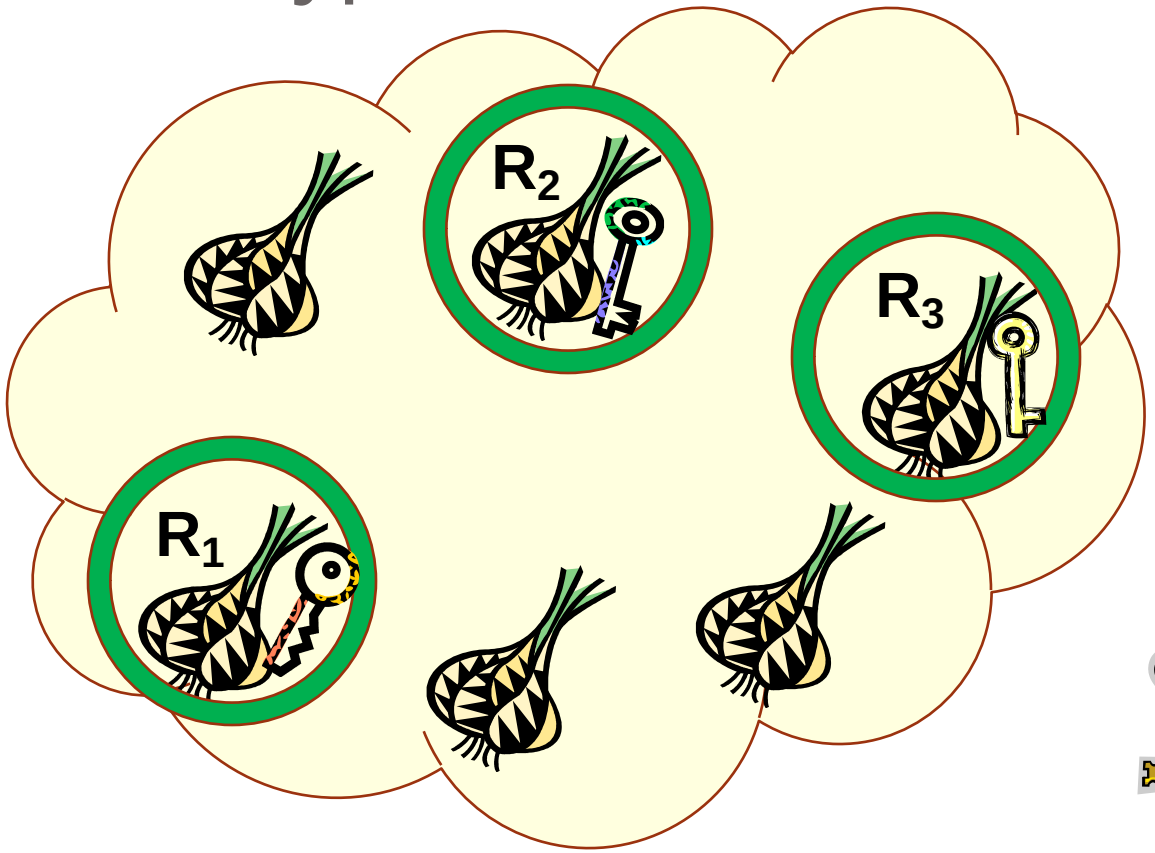
- 3rd party anonymity
  - Alice and Bob trust each other but do not want other parties to learn that they are communicating
- Sender anonymity
  - Alice sends to Bob, and Bob cannot trace Alice's identity
- Receiver Anonymity
  - Bob can contact Alice, without knowing her identity.
- Bi-directional Anonymity
  - Alice and Bob communicate without knowing each other's identities.

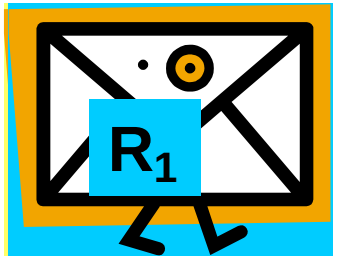
Slide credit: George Danezis

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)
  - Low-latency communication: Freedom Network (1999-2001), JAP (2000), Tor (2005)

Onion encryption

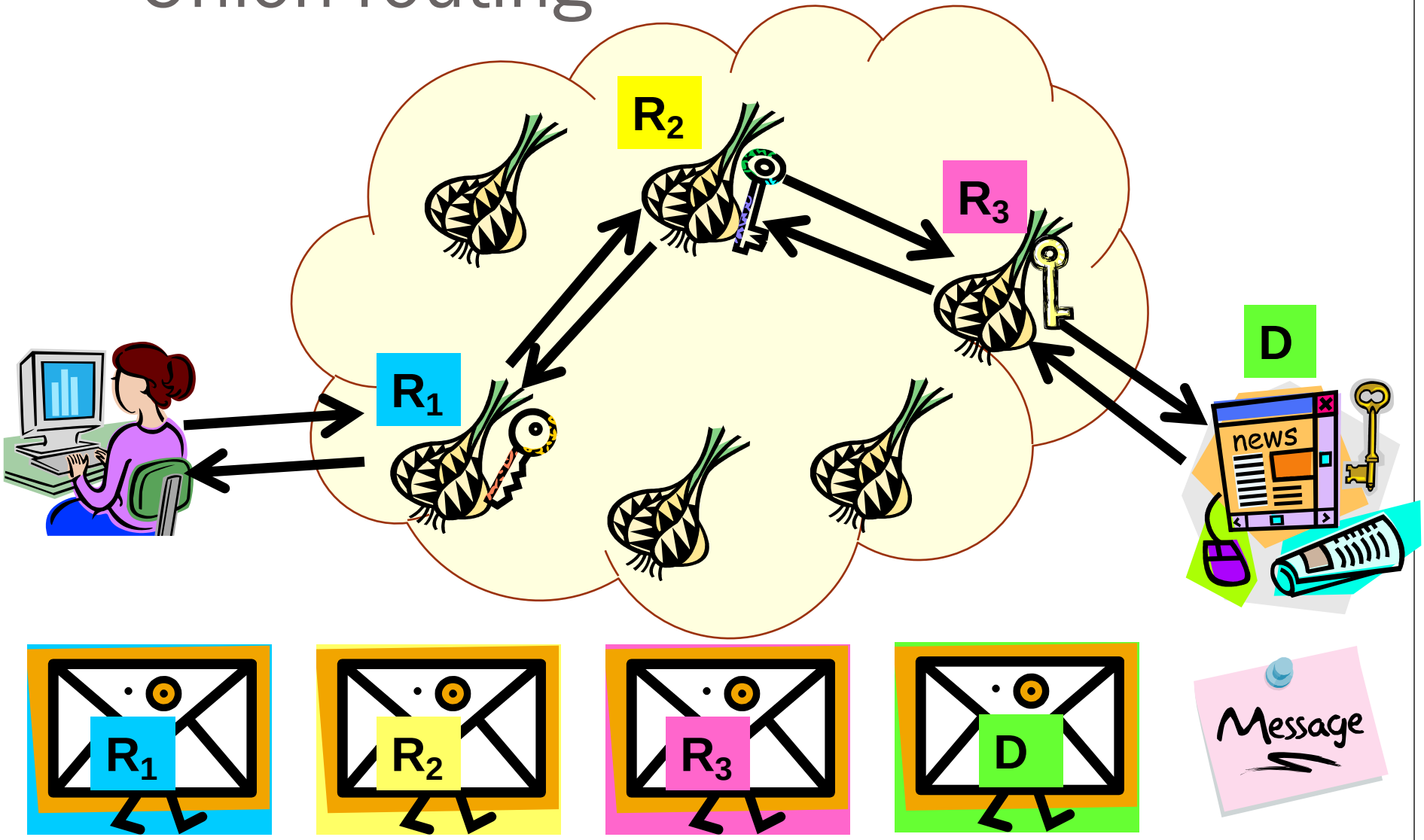




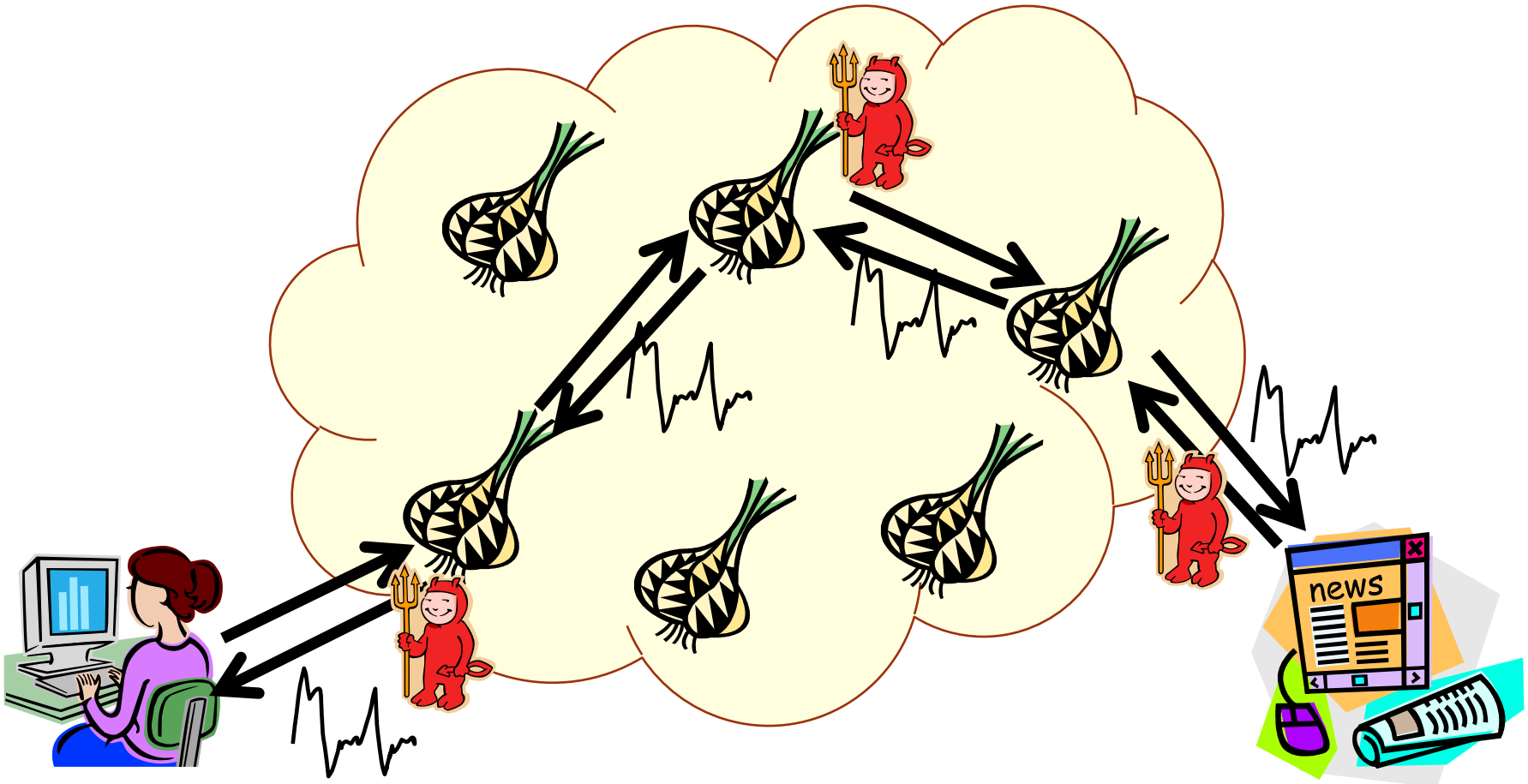
D

Claudia Diaz (K.U.Leuven)

Onion routing

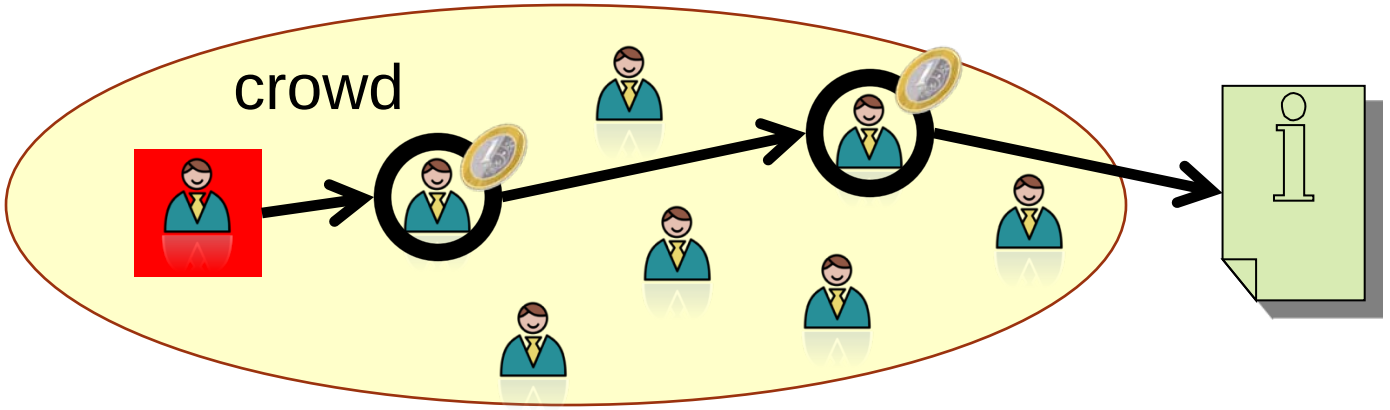


TOR – adversary model



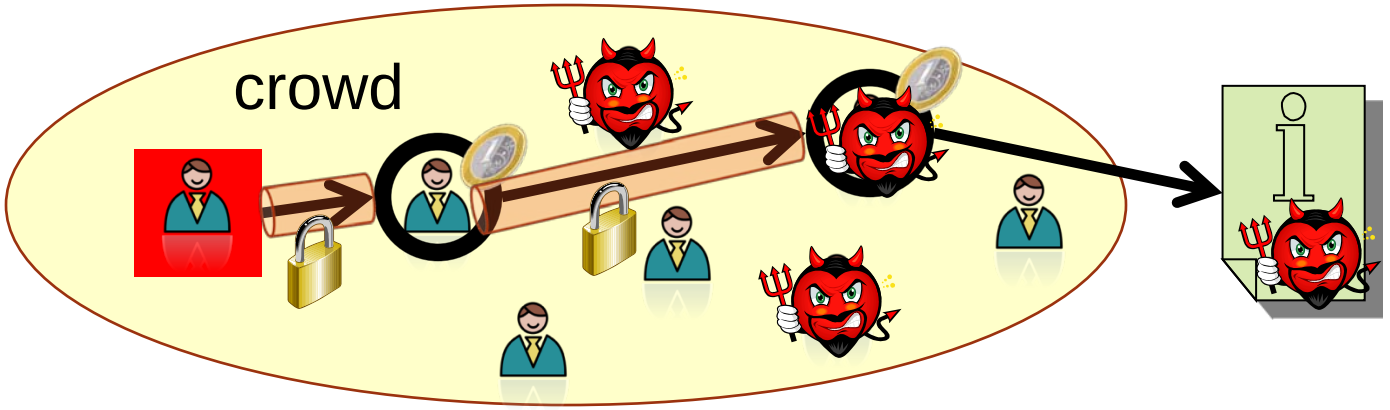
Crowds (Reiter, Rubin 1998)

- Anonymity for web browsing
- Group of users form a "crowd"
- Initiator chooses a random member of the crowd and forwards the web request to her
- The recipient of the request flips a biased coin and forwards the request to another member with probability *p* and to the end server with probability *1-p*
- A tunnel is established between the initiator and the exit crowd member (static paths)



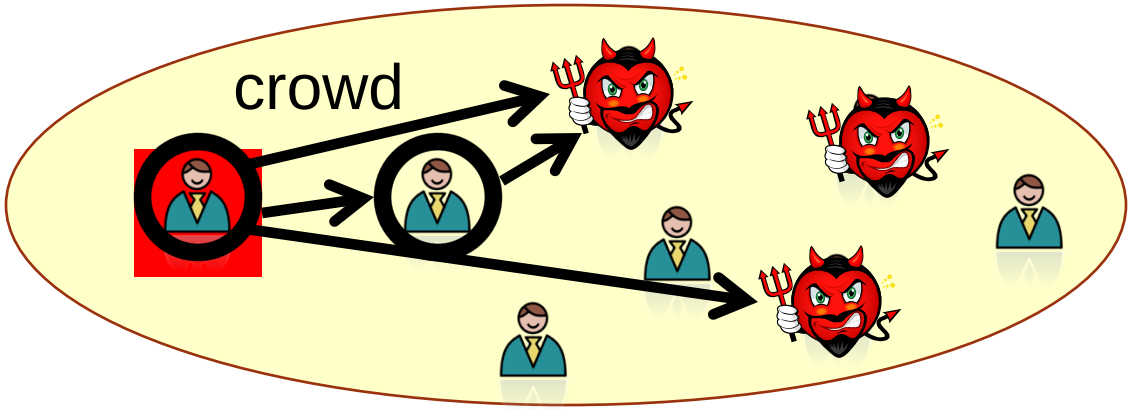
Crowds (Reiter, Rubin 1998)

- Communication between members is encrypted with symmetric keys
  - BUT: all members can see the request in clear
- Adversary model:
  - Assumed adversary cannot control all links
  - Instead, the adversary controls a subset of the crowd and/or the end server
- Probability that predecessor/exit is the initiator or just a forwarder
  - We can measure initiator anonymity as a function of the fraction of corrupted nodes and the probability of forwarding



Crowds (Reiter, Rubin 1998)

- Predecessor attacks
  - If initiator repeatedly accesses the same resource over different sessions, it will appear as predecessor of the first adversarial member more often than other crowd members
  - Anonymity degrades with
    - Amount of linkable requests made in different sessions
    - Size of the crowd
  - These attacks are applicable to all P2P anonymity systems



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

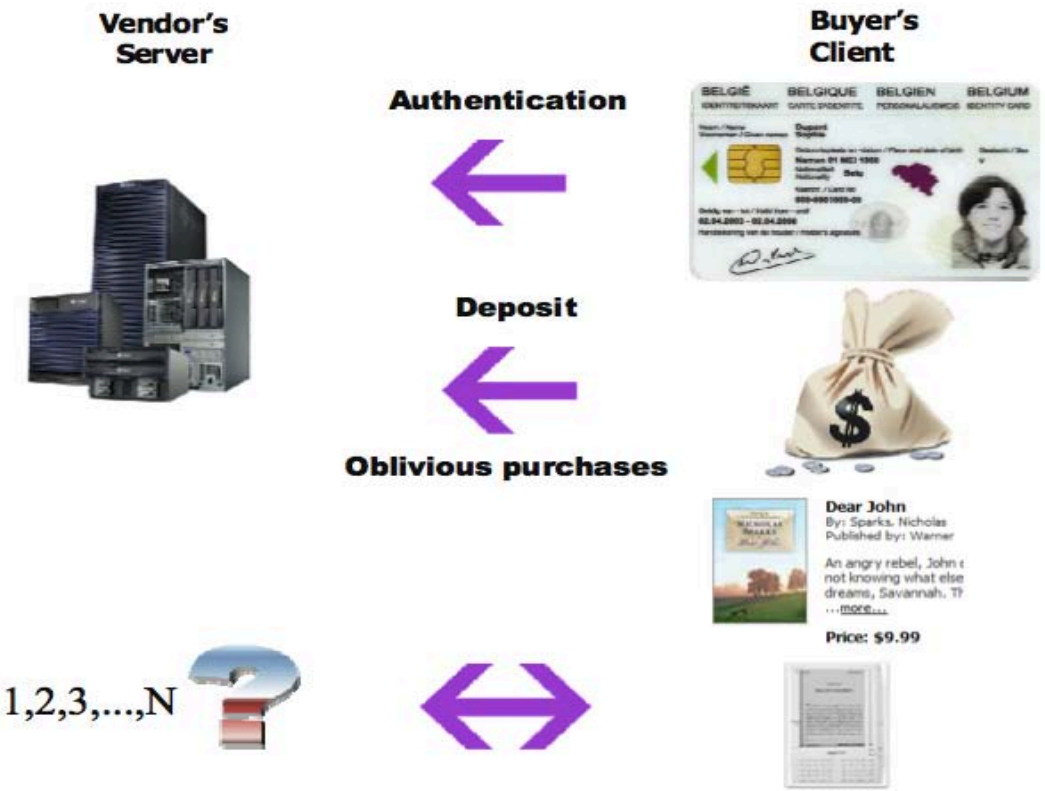
Identification + data minimization

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

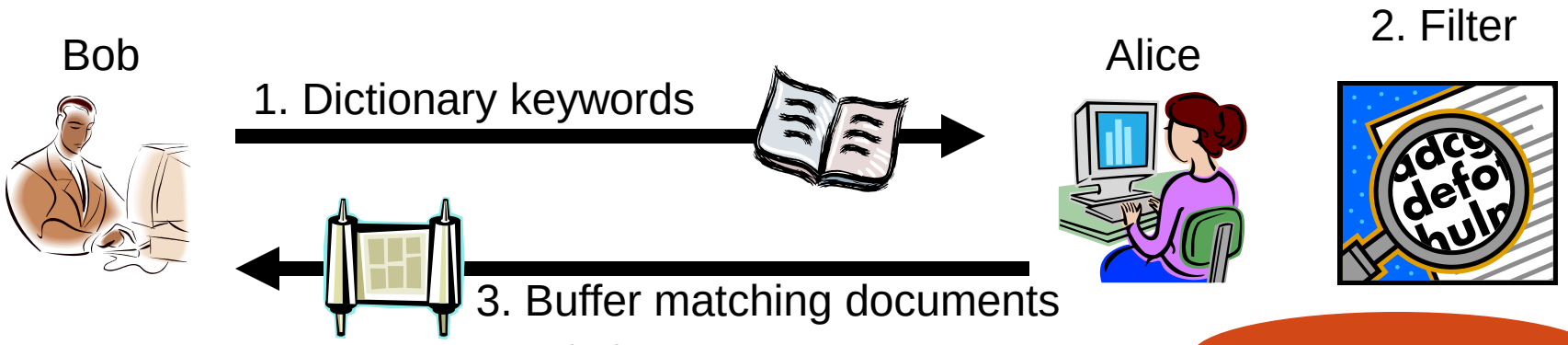
Buying digital content

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



Private Search

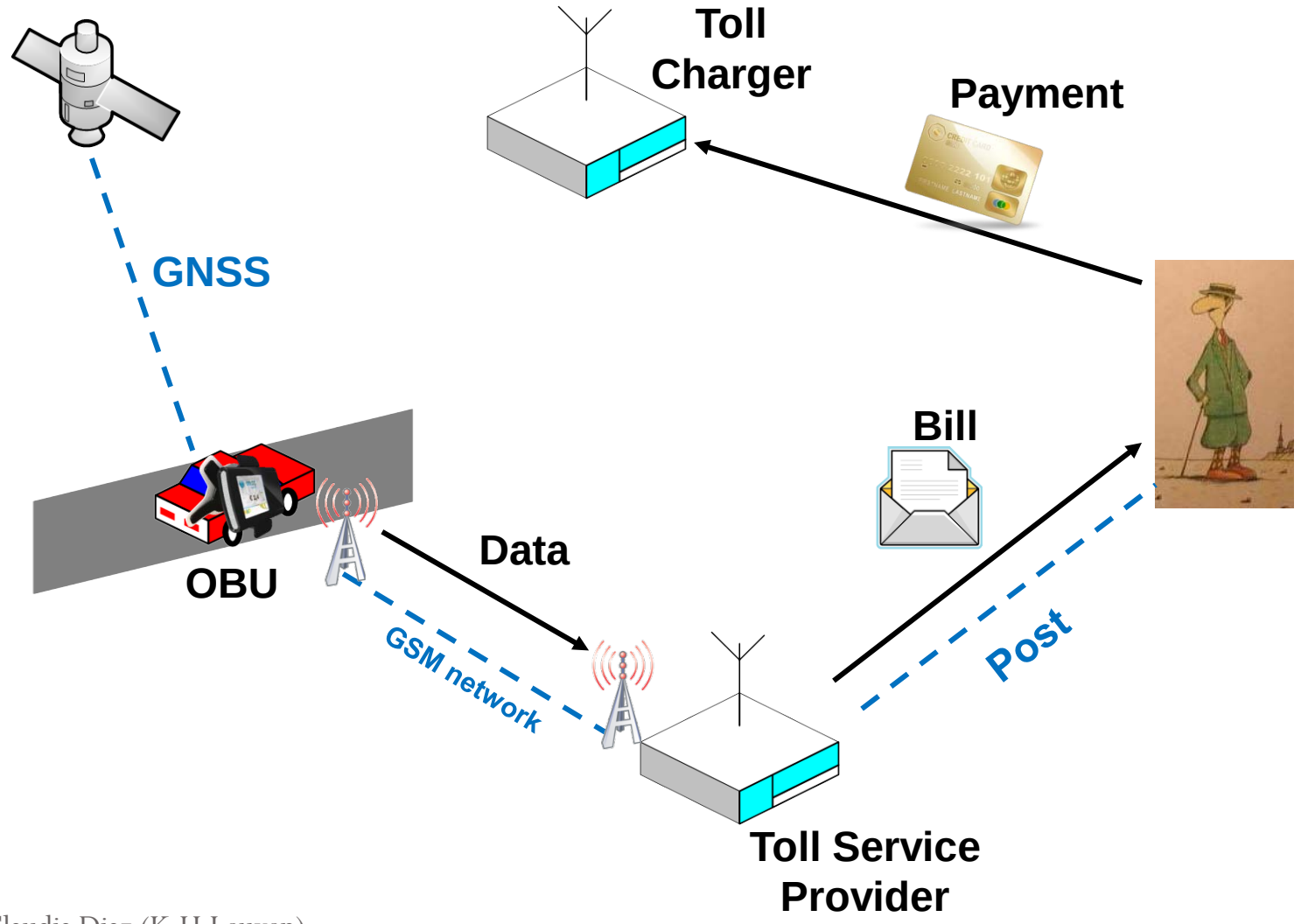
- Alice stores documents
- Bob wants to retrieve documents matching some keywords
- Properties:
  - Bob gets documents containing the keywords
  - Alice does not learn Bob's keywords
  - Alice does not learn the results of the search



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
  - Member states should implement it in the coming years

EETS straightforward implementation



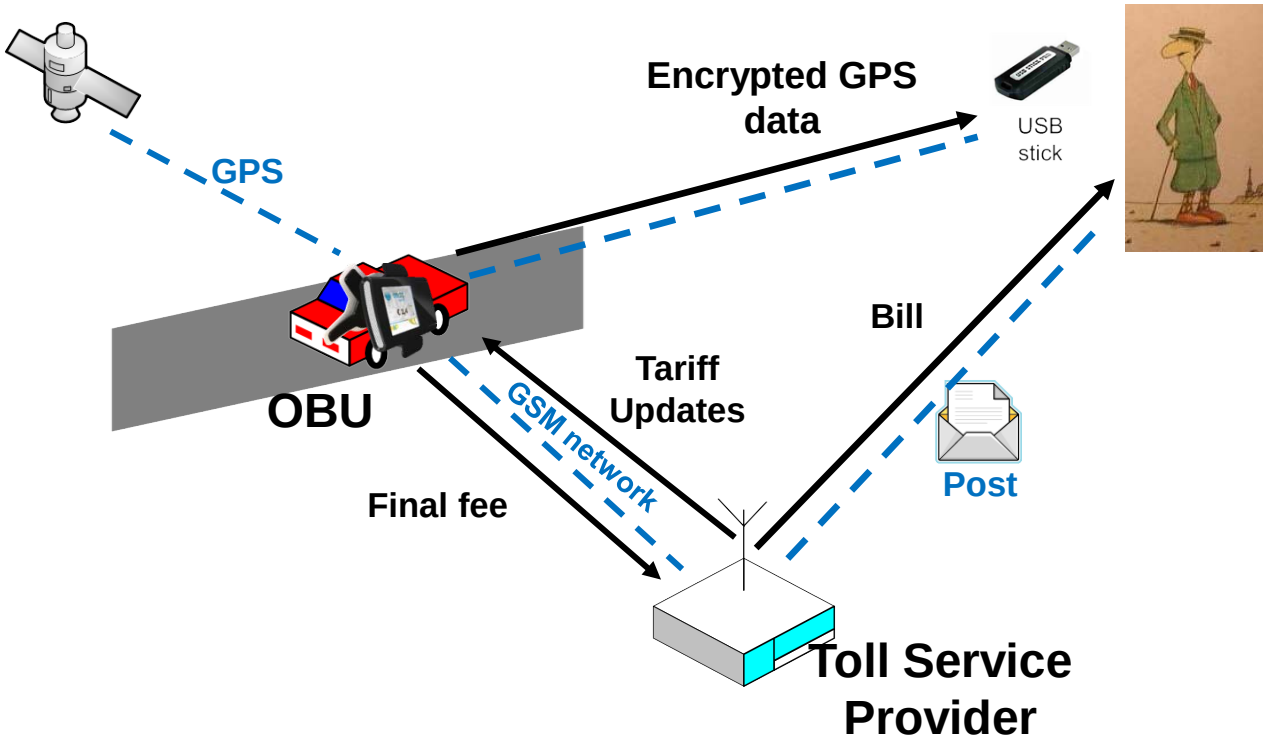
Slide credit: Carmela Troncoso

Privacy for Electronic Toll Pricing

- Privacy issues?
  - Pay *as you drive*
  - Fine grained GPS data allows for inferences
- 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



Slide credit: Carmela Troncoso

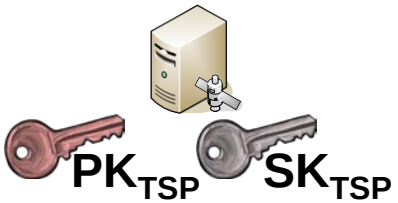
Enforcement

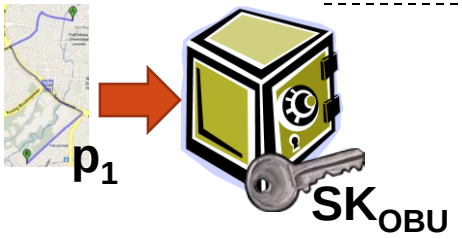
- OBU in hands of the user
  - Incentive to cheat (paying less)
  - Even if the box is tamper-resistant, the input is easy to spoof
- We need to:
  - Detect vehicles with inactive OBUs
  - Detect vehicles reporting false location data
  - Detect vehicles using incorrect road prices
  - Detect vehicles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

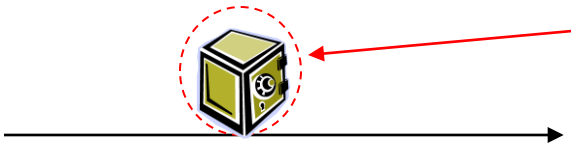
Slide credit: Carmela Troncoso

Non-Interactive Commitment Schemes

	00u00 – 07u00	· · · · · · · · ·	22u00 – 00u00
Highway	p_1	· · · · · · · · ·	p_2
Primary	p_3	· · · · · · · · ·	p_4
· · · · · · · · ·	· · · · · · · · ·	· · · · · · · · ·	· · · · · · · · ·
Residential	p_{n-1}	· · · · · · · · ·	p_n

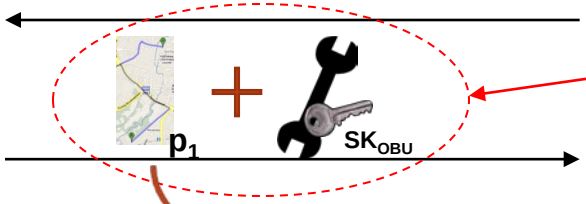




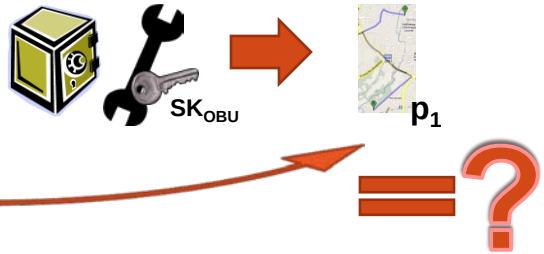


HIDING PROPERTY

Where you at….?



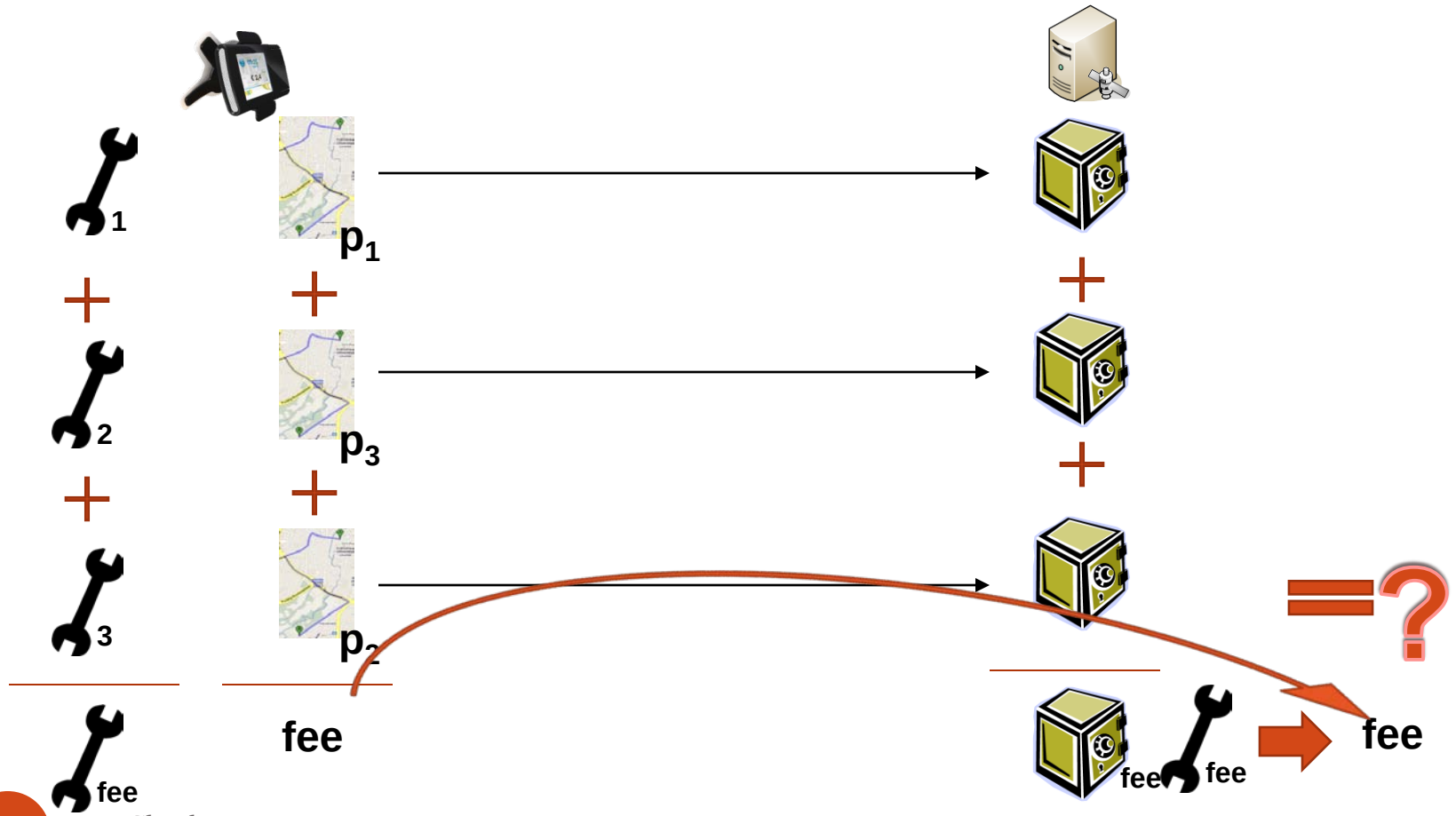
BINDING PROPERTY



Slide credit: Carmela Troncoso

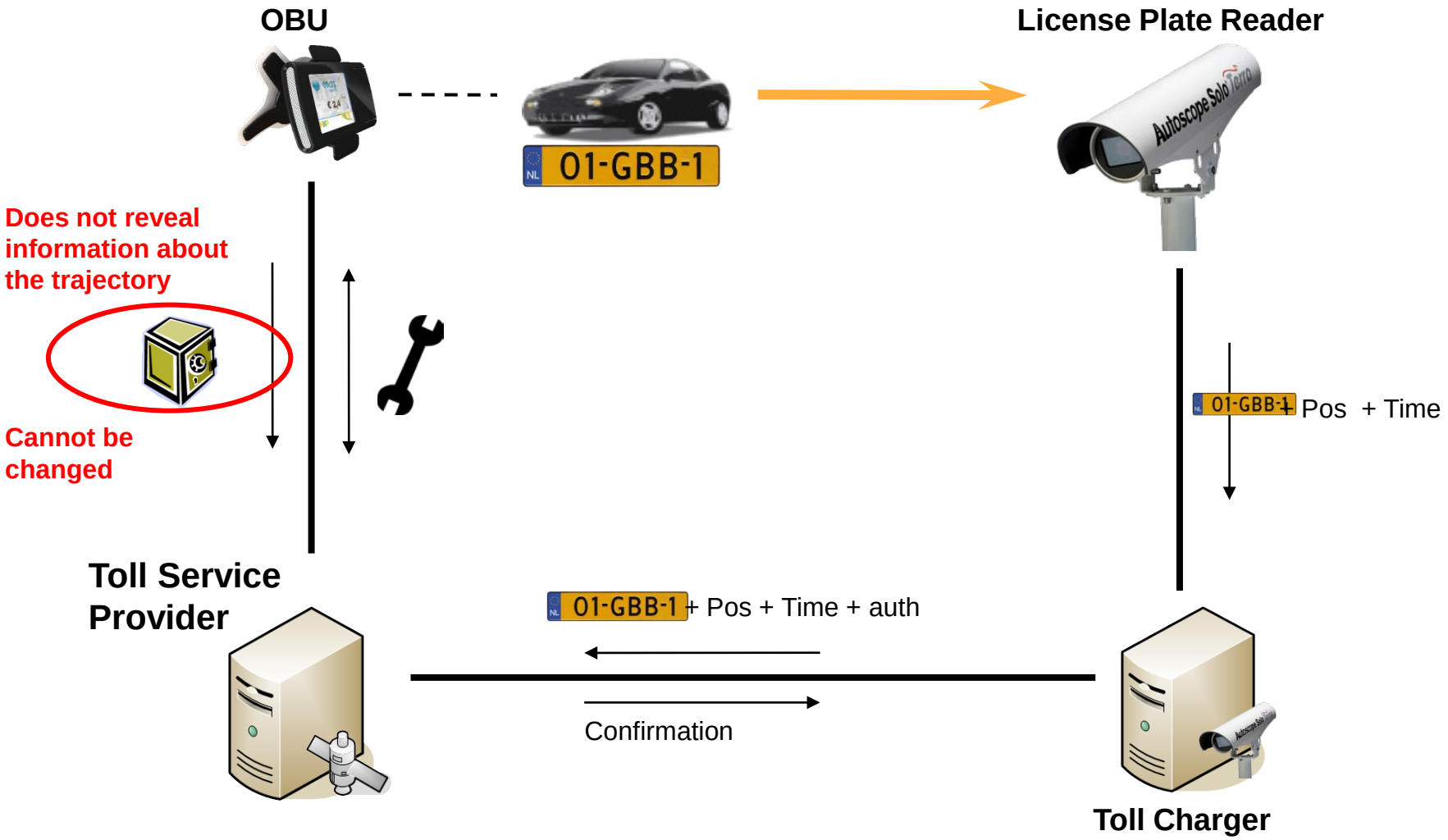
Homomorphic commitments

- The content of the vaults can be added up without being known



Slide credit: Carmela Troncoso

How does it work?



Slide credit: Carmela Troncoso

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

Slide credit: Carmela Troncoso

Privacy in databases

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
    - 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)

Hospital Patient Data

DOB	Sex	Zipcode	Disease
1/21/76	Male	53715	Heart Disease
4/13/86	Female	53715	Hepatitis
2/28/76	Male	53703	Brochitis
1/21/76	Male	53703	Broken Arm
4/13/86	Female	53706	Flu
2/28/76	Female	53706	Hang Nail

Vote Registration Data

Name	DOB	Sex	Zipcode
Andre	1/21/76	Male	53715
Beth	1/10/81	Female	55410
Carol	10/1/44	Female	90210
Dan	2/21/84	Male	02174
Ellen	4/19/72	Female	02237

K-anonymity

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

Release Table

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

External Data Source

Name	Birth	Gender	ZIP	Race
Andre	1964	m	02135	White
Beth	1964	f	55410	Black
Carol	1964	f	90210	White
Dan	1967	m	02174	White
Ellen	1968	f	02237	White

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

Other technologies

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

State of the art: OTR plug-in for Instant Messaging (IM).

Fake transactions to protect against profiling

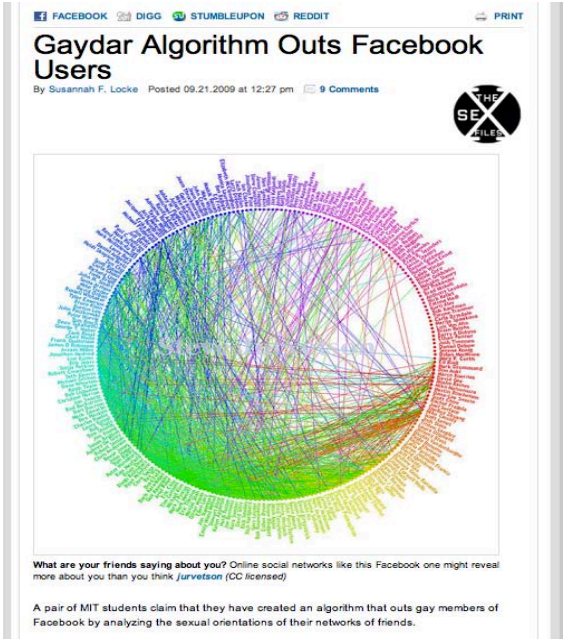
- TackMeNot: obfuscation-based private web search
- Generation of dummy queries to:
  - Obfuscate search profile (interests)
  - Provide query deniability
- Issues
  - Difficulty of generating plausible/indistinguishable dummy queries
  - Difficulty of concealing profiles with moderate amounts of dummy queries
  - Assuming that the profile is successfully obfuscated, privacy concerns remain: profiles will be used even if they are inaccurate

Location privacy

- Smart phones becoming ubiquitous, development of a variety of location-based services
- Location data can be highly sensitive: possible to infer movements, relationships, status, lifestyle, . . . not just location!
- "Anonymous" location traces easy to re-identify
- Mix zones: based on mixes used in anonymous communications
  - Long-term linkability of traces?
- Cloaking regions: based on k-anonymity in databases
  - Problems with how the ideas have been translated from databases to location based services
  - Intersection of various requests?

Privacy in social networks

- Tension with data sharing
- Variety of possible privacy breaches
  - Content more widely available than intended
    - Misconfiguration of privacy settings
    - Changes of settings by the provider
    - Info disseminated by others
  - Large scale SNS providers have access to rich information on millions of people: implications?
  - Profiling and inferences
  - Not just about user-generated content, but also interaction information
  - Censorship, filter bubble, privatization of the social space



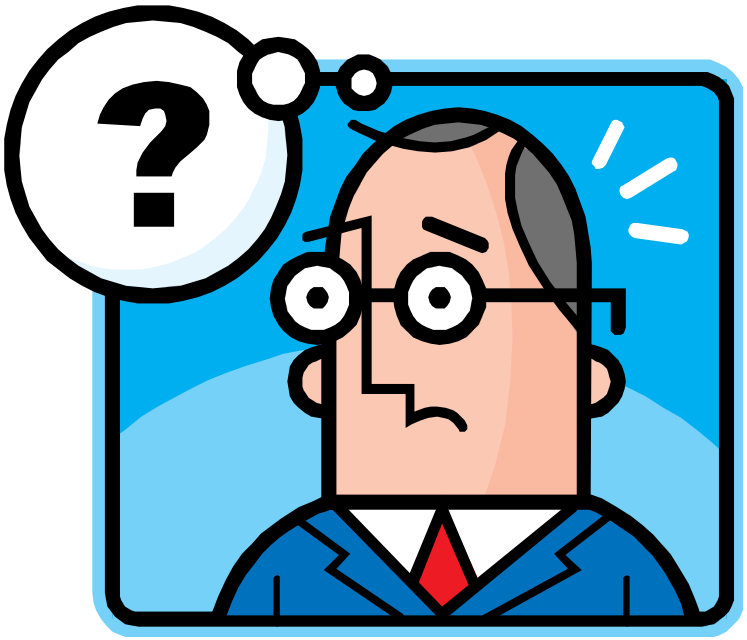
Privacy challenges

- 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

Conclusions

- PETs can reconcile aggressive data minimization and service integrity guarantees
- Compliance is a strong driver
- Data protection technologies
  - Hidden costs of securing large databases
- Privacy Enhancing Technologies
  - Active research, lots of proposed solutions
  - Poor deployment

Thanks !



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