Privacy Technologies

Claudia Diaz

K.U.Leuven COSIC

Acks slides: George Danezis and Carmela Troncoso

Outline

- Privacy and security
- Approaches to privacy
  - Trust-based privacy (data protection)
  - Hard privacy (PETs)
- Overview Privacy Technologies
- Open problems, challenges, conclusions

Caricature of the debate: Security or Privacy

- "Privacy" important but. . .
  - …what about abuse and accountability?
  - …difficulties for Law Enforcement?
  - …copyright or libel
  - (…what does a good, honest person have to hide anyway?
- Established wisdom:
  - Need for a balance…
  - Result: Surveillance by design → no privacy (often).

Privacy *is* Security

- Shared infrastructure:
  - Telecommunications, operating systems, search engines, on-line shops, software, . . .
  - Denying security to some (by building in surveillance), means denying it to all
- Company secrets may be leaked by…
  - Looking at certain patents, search queries
  - Phone calls and movements of an employee (or CEO)
  - Using a cloud to store or process information
  - Employees using social networks
- Same for governments / the police (national security)

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

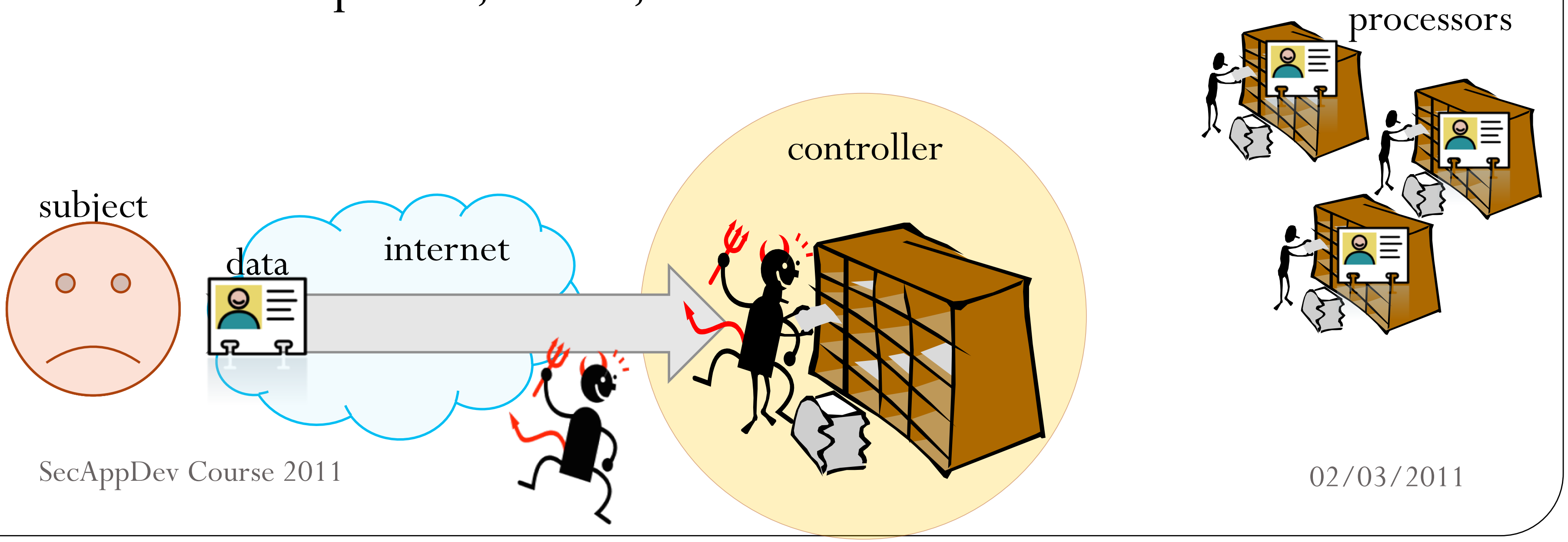
- These are the goals of **Privacy Enhancing Technologies**

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

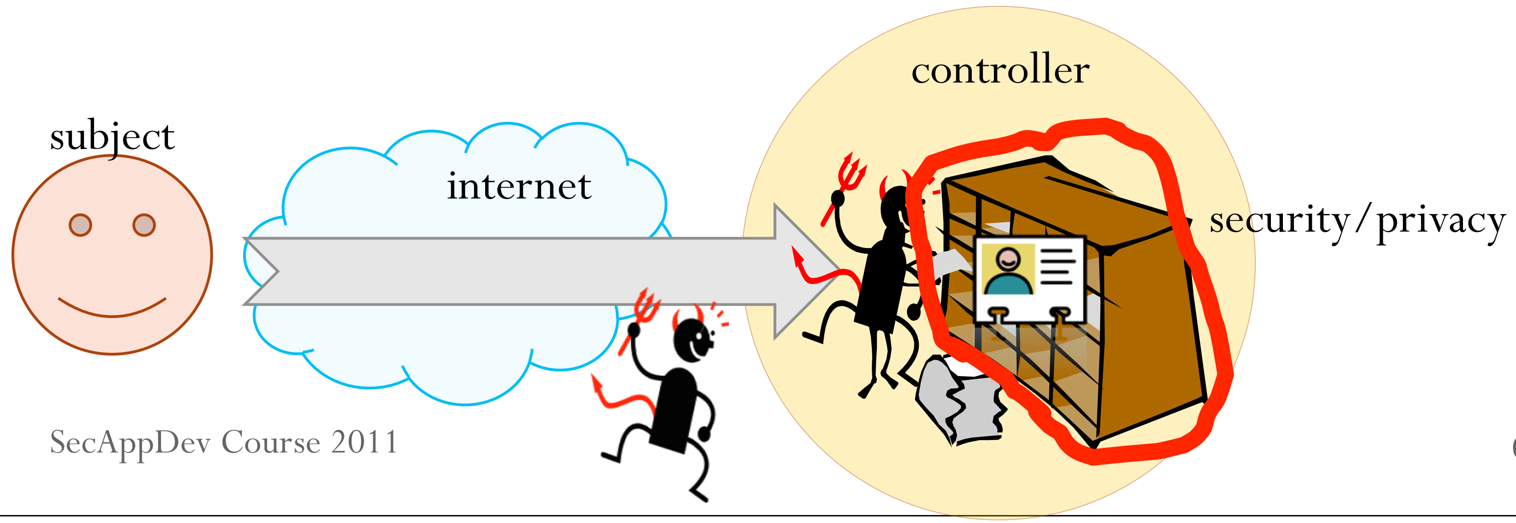
Trust-based privacy

- 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



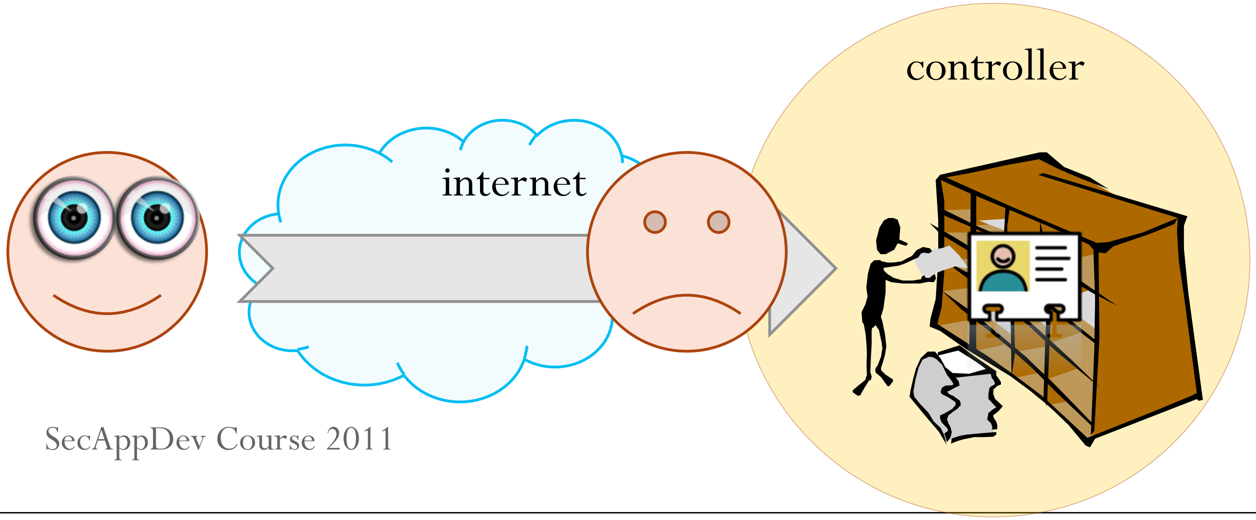
Trust-based privacy

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



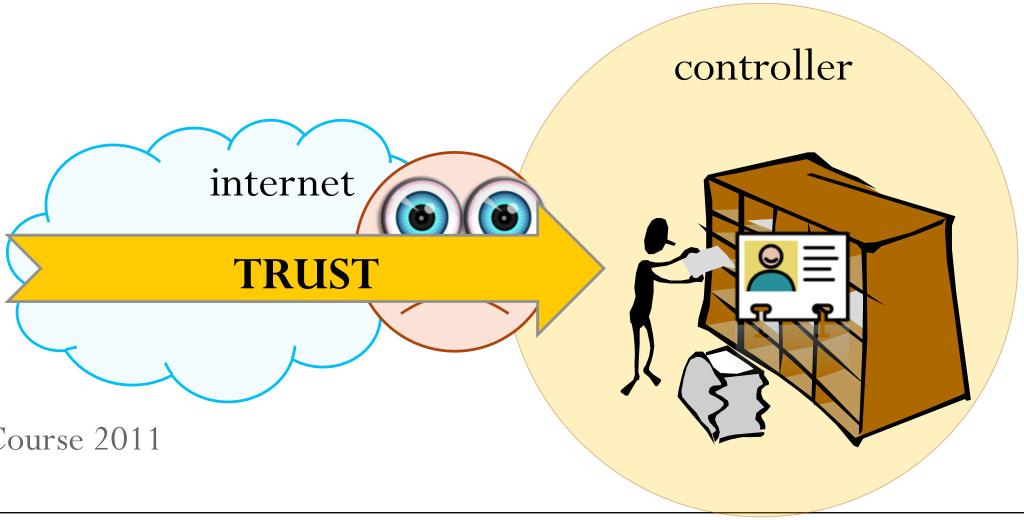
Trust-based privacy

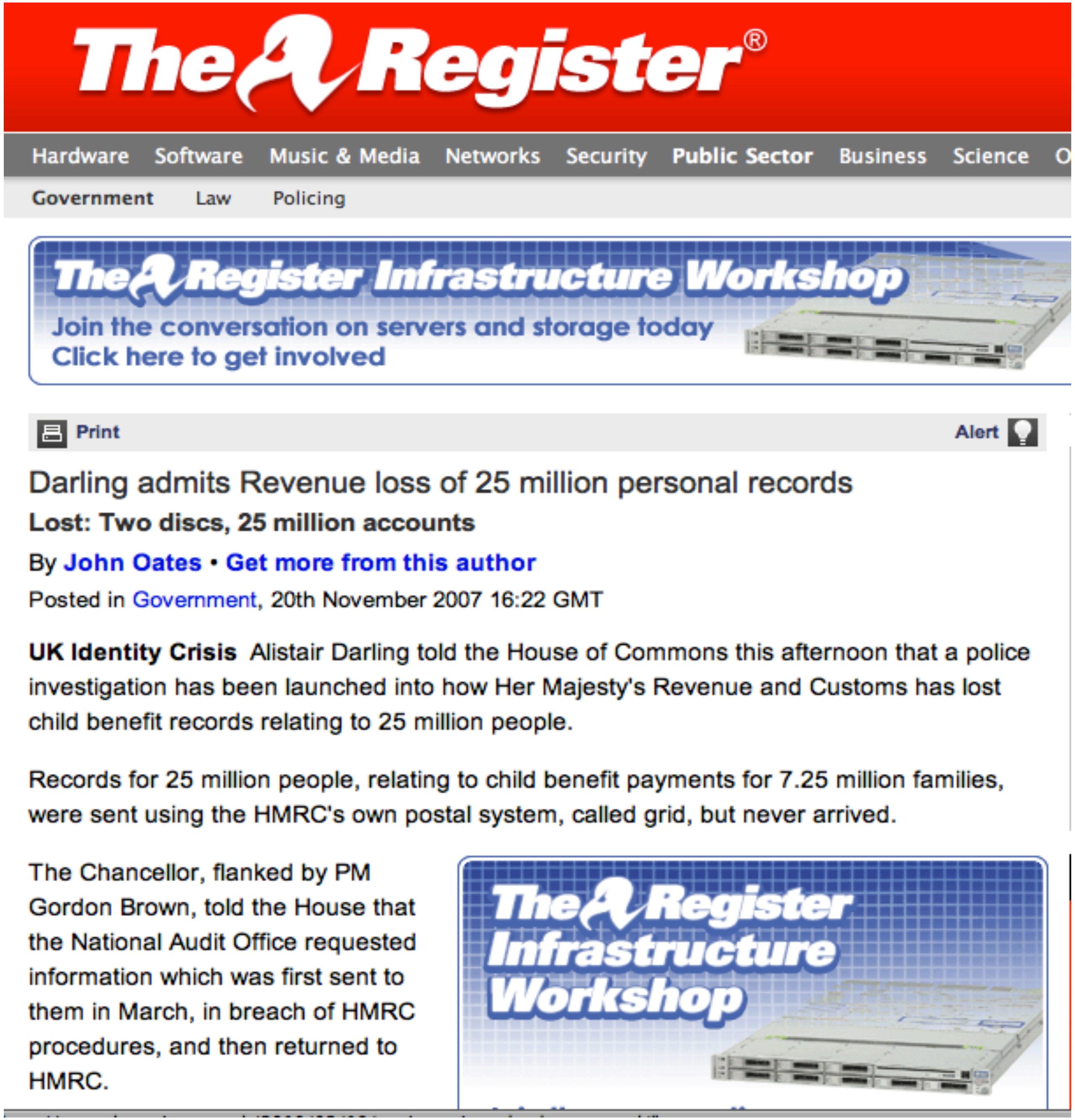
- Data subject has already lost control of her data
  - In practice, very difficult for data subject to verify how her data is collected and processed



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

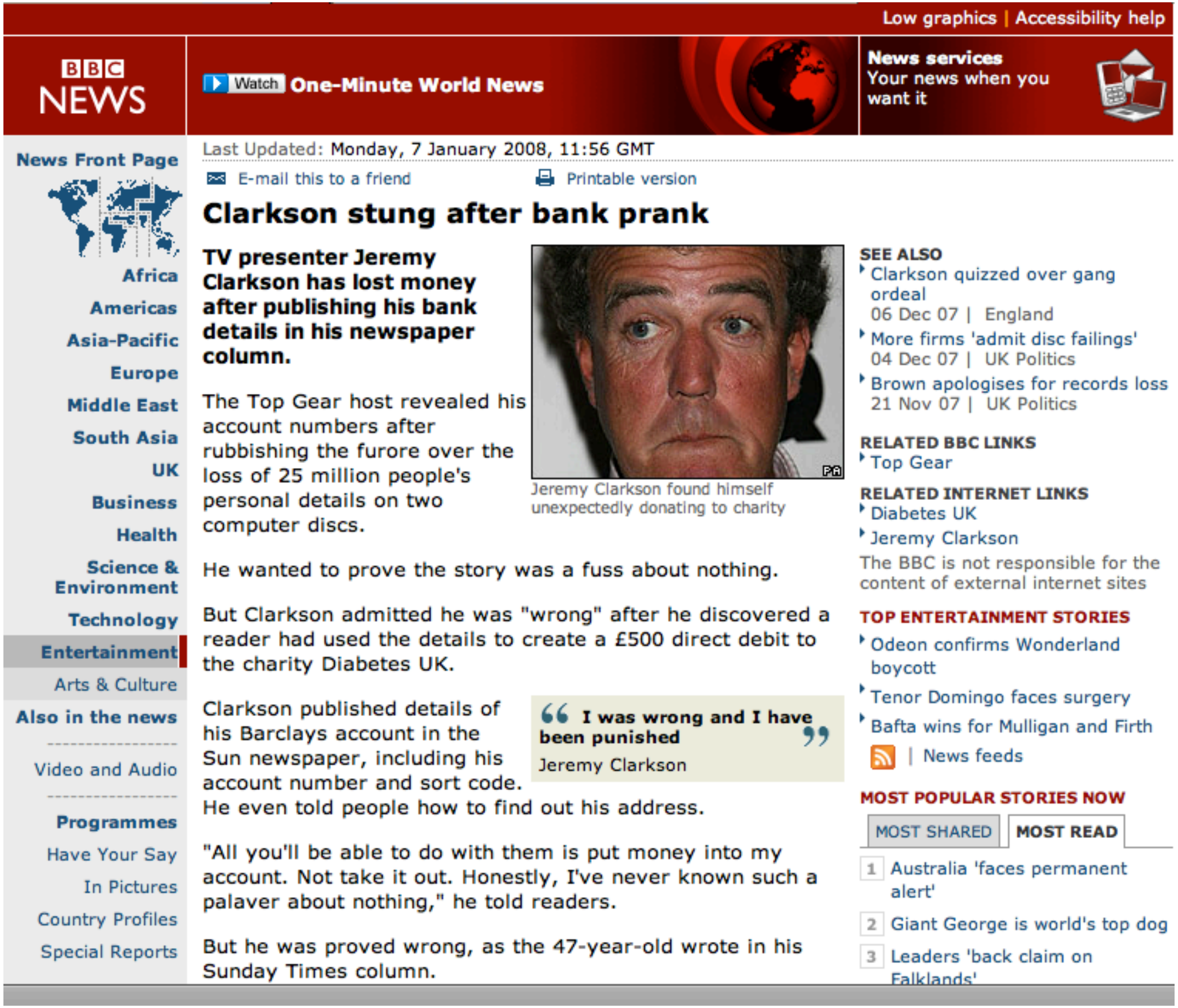
By John Oates • Get more from this author

Posted in Government, 20th November 2007 16:22 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.



Low graphics | Accessibility help
BBC NEWS
Watch One-Minute World News
News services
Your news when you want it

Last Updated: Monday, 7 January 2008, 11:56 GMT

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 rubbishing the furore over the loss of 25 million people's personal details on two computer discs.

Jeremy Clarkson found himself unexpectedly donating to charity

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.

> **I was wrong and I have been punished**
> Jeremy Clarkson

"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.

SEE ALSO
- Clarkson quizzed over gang ordeal
  06 Dec 07 | England
- More firms 'admit disc failings'
  04 Dec 07 | UK Politics
- Brown apologises for records loss
  21 Nov 07 | UK Politics

RELATED BBC LINKS
- Top Gear

RELATED INTERNET LINKS
- Diabetes UK
- Jeremy Clarkson

The BBC is not responsible for the content of external internet sites

TOP ENTERTAINMENT STORIES
- Odeon confirms Wonderland boycott
- Tenor Domingo faces surgery
- Bafta wins for Mulligan and Firth

| News feeds

MOST POPULAR STORIES NOW

1 Australia 'faces permanent alert'
2 Giant George is world's top dog
3 Leaders 'back claim on Falklands'

Problems of trust-based privacy

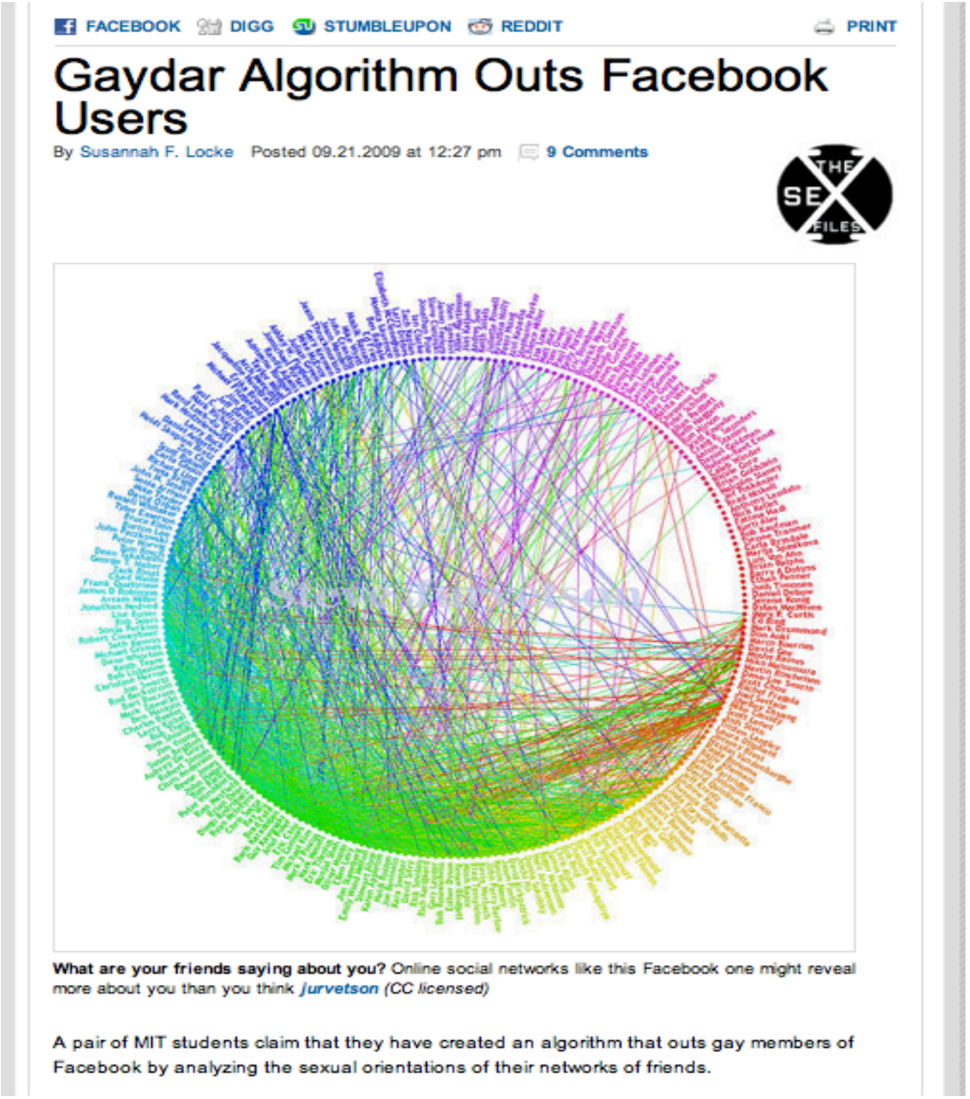
- Data minimization (proportionality) often ignored
- Informed consent?
- Trust assumptions may not be realistic
  - Incompetence
  - Malicious insiders
  - Incentives?
  - Purpose (function creep)
  - Cost of securing the data
- How can you check that your data is not being abused?
- Weak enforcement,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



Gaydar Algorithm Outs Facebook Users

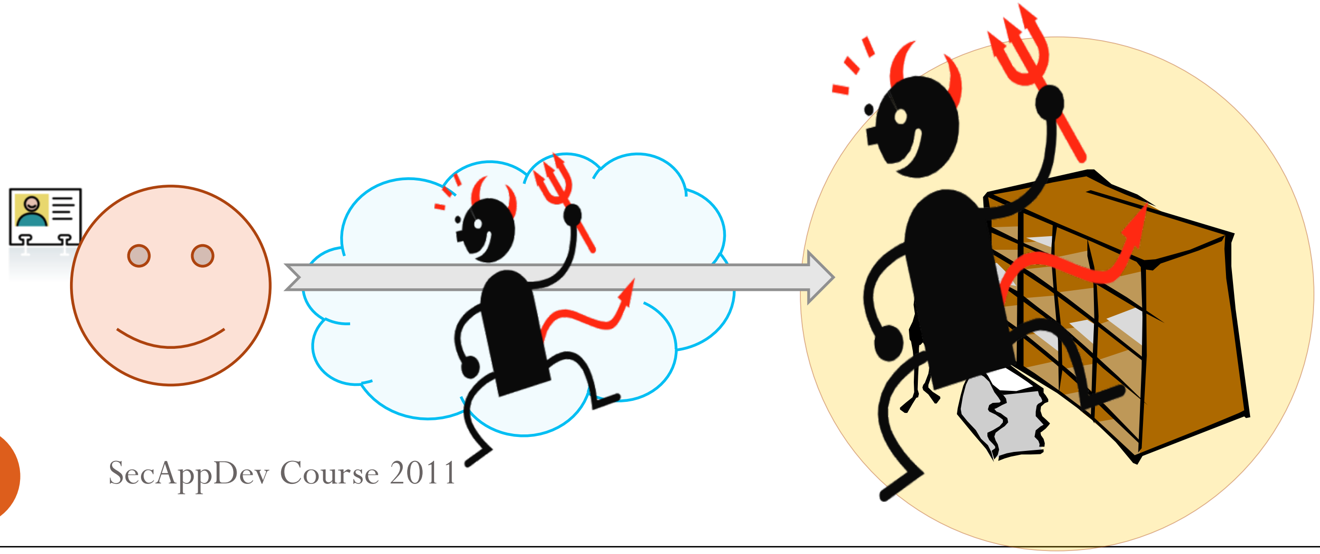
By Susannah F. Locke Posted 09.21.2009 at 12:27 pm 9 Comments

What are your friends saying about you? Online social networks like this Facebook one might reveal more about you than you think *jurvetson (CC licensed)*

A pair of MIT students claim that they have created an algorithm that outs gay members of Facebook by analyzing the sexual orientations of their networks of friends.

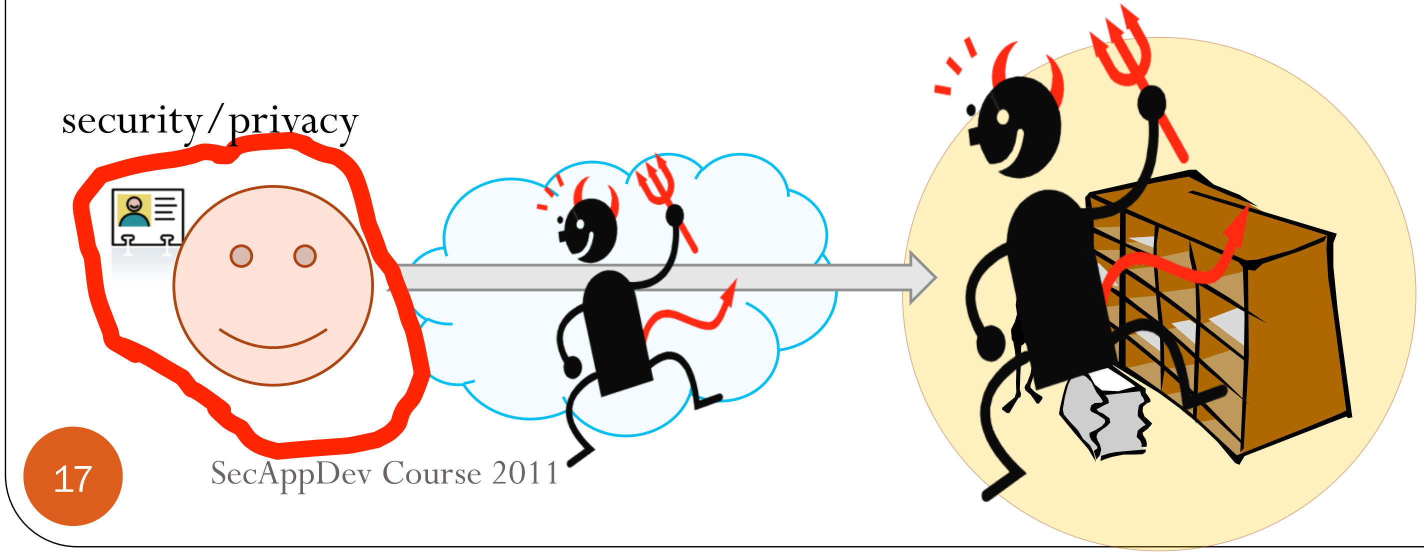
Hard Privacy (PETs)

- 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



Overview of PETs

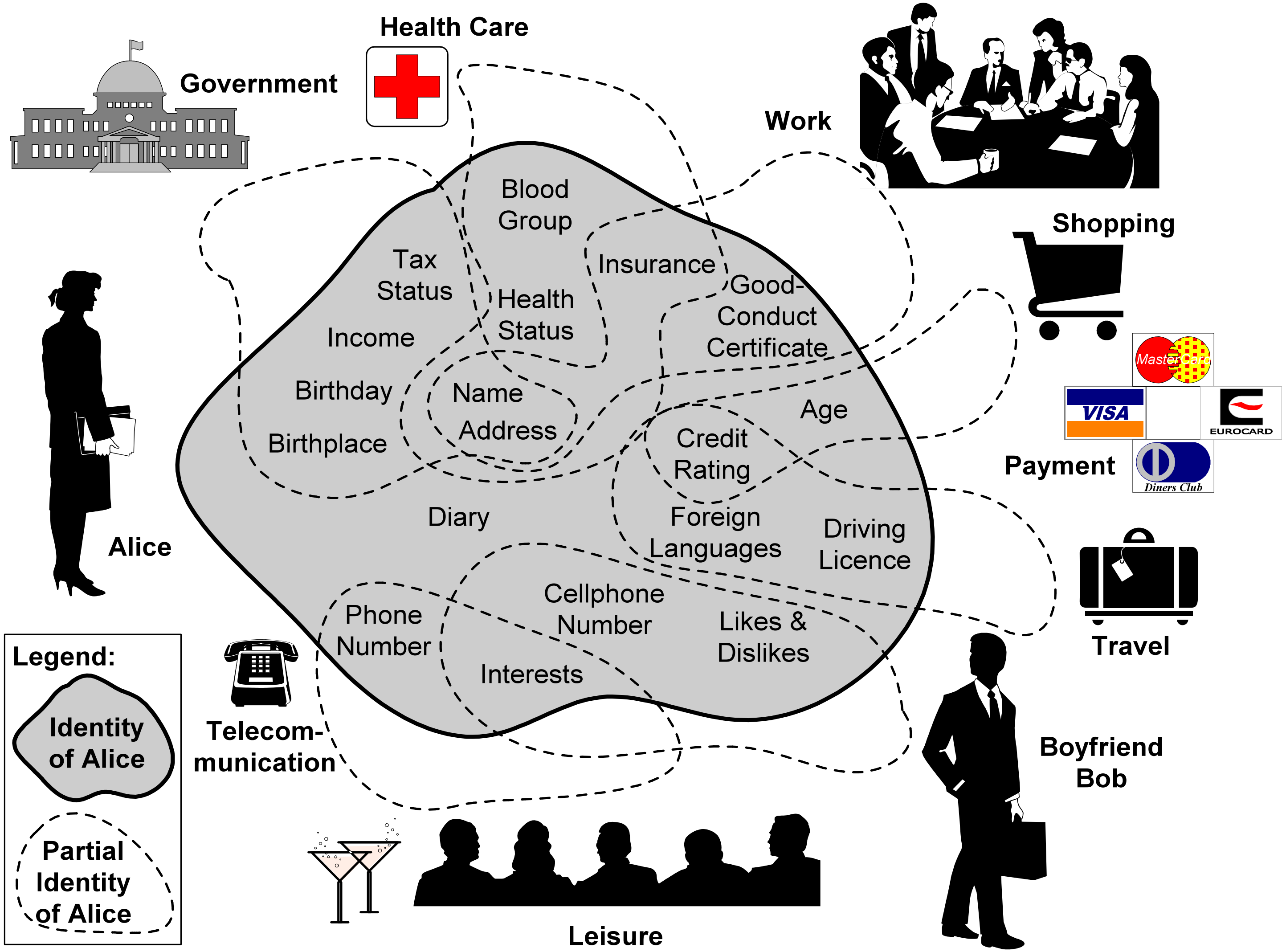
Authentication

- 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

Identity Management: partial identities



Ack slide: Marit Hansen

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	Anonymous credentials
Signed by a trusted issuer	Signed by a trusted issuer
Certification of attributes	Certification of attributes
Authentication (secret key)	Authentication (secret key)
Double-signing detection	Double-signing detection
No data minimization	Data minimization
Users are identifiable	Users are anonymous
Users can be tracked (Signature linkable to other contexts where PK is used)	Users are unlinkable in different contexts

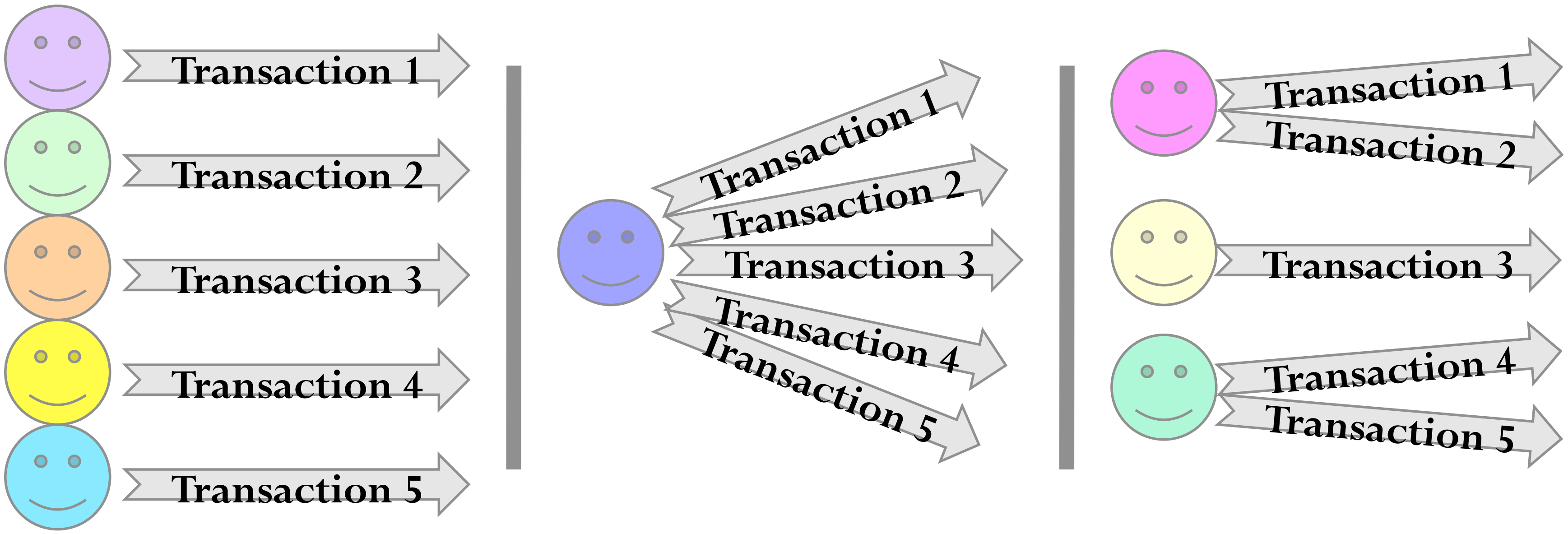
Types of anonymous credentials

- 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?

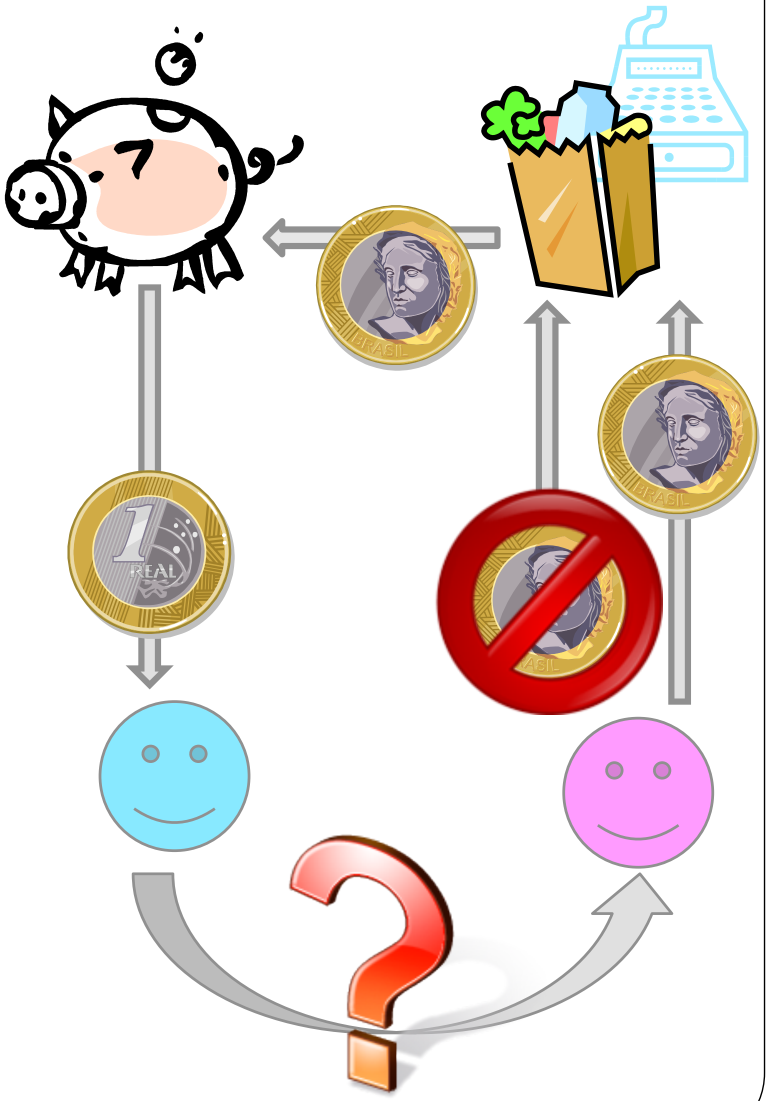
Pseudonymous identity management

- One-time pseudonyms: anonymity
- Persistent pseudonyms: they become an identity
- Solutions in between: context specific (partial) identities



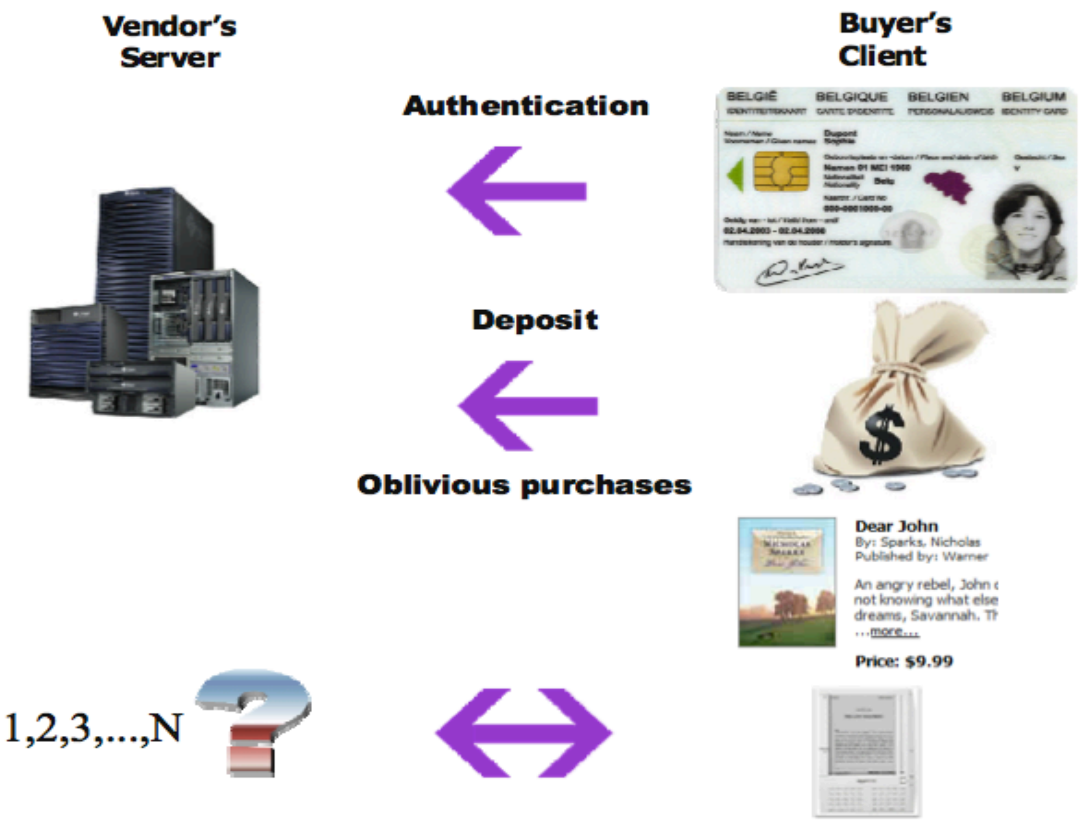
Anonymous e-cash

- 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!



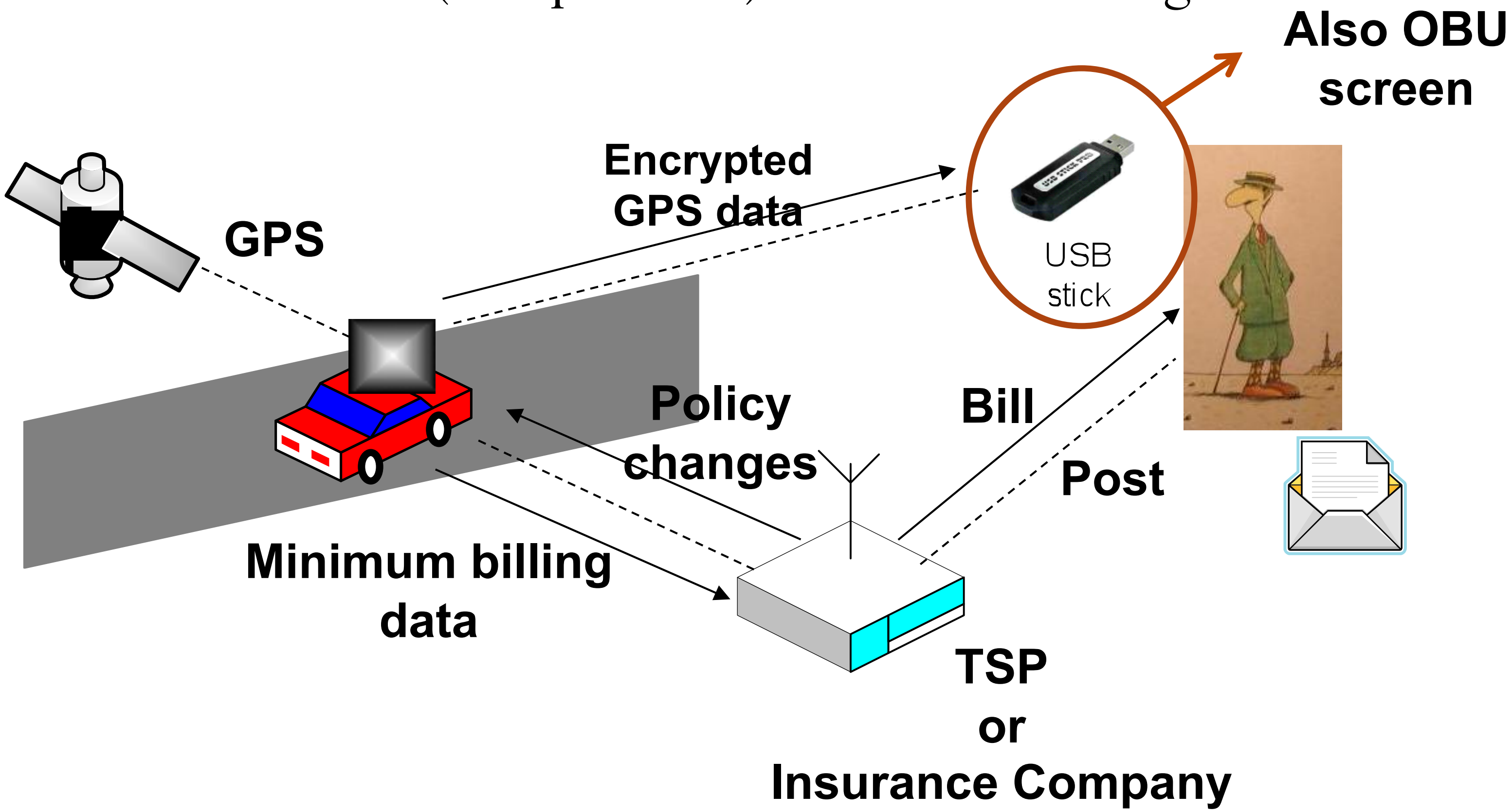
Private Information Retrieval (PIR) / Oblivious Transfer (OT)

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



PriPAYD: car insurance / e-Toll

- Keep data under the control of the user, and transmit minimal information
- GPS + Black box (computation) + transmit billing



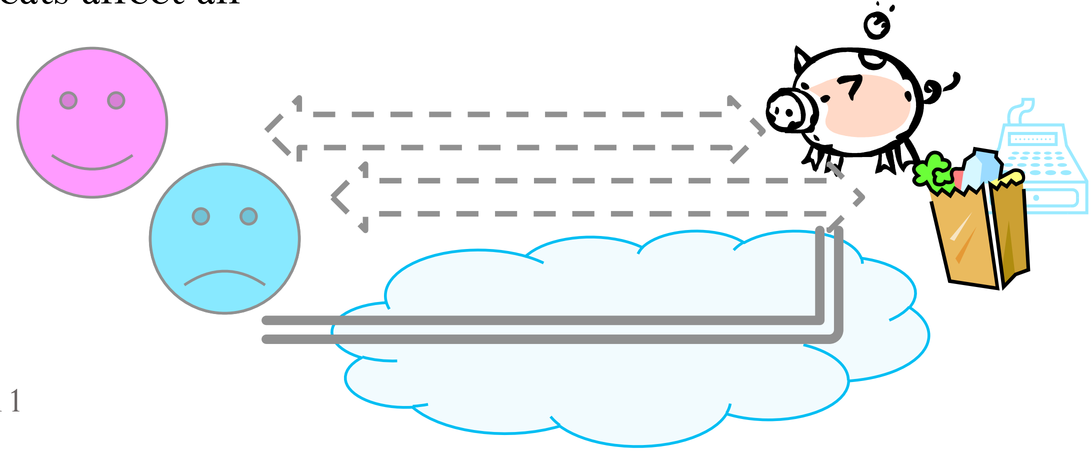
Off-The-Record (OTR) security

- 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).

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)

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)

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
  - 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

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)
  - Note: data protection does not apply to anonymized data

K-anonymity techniques

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

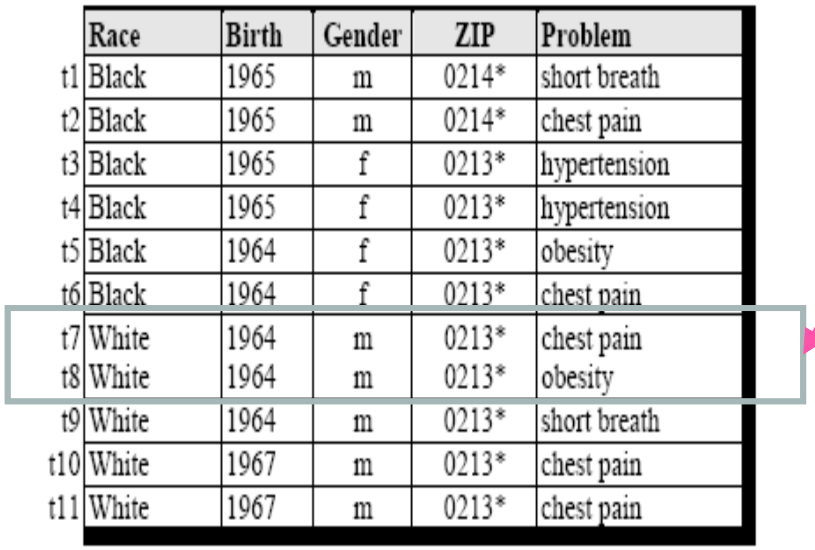
Vote Registration Data

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

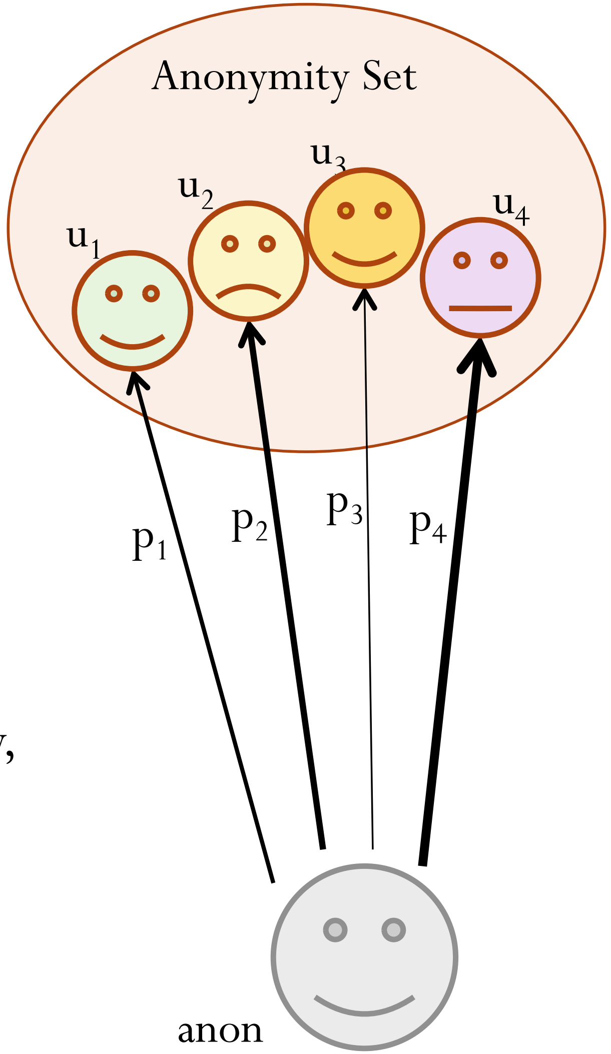
Figure 2 Example of *k*-anonymity, where *k*=2 and QI={*Race, Birth, Gender, ZIP*}

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

Defining anonymity

- Definitions [PH00]
  - "***Anonymity*** *is the state of being not identifiable within a set of subjects, the anonymity set.*"
  - "The ***anonymity set*** is the *set of all possible subjects who might cause an action or be addressed.*"
  - "Anonymity is the stronger, the larger the respective anonymity set is and the more evenly distributed the sending or receiving, respectively, of the subjects within that set is."
  - Probabilistic definition
    - Probabilistic definitions also possible for unlinkability, unobservability, deniability, …
- Probabilistic nature not captured by legal definitions

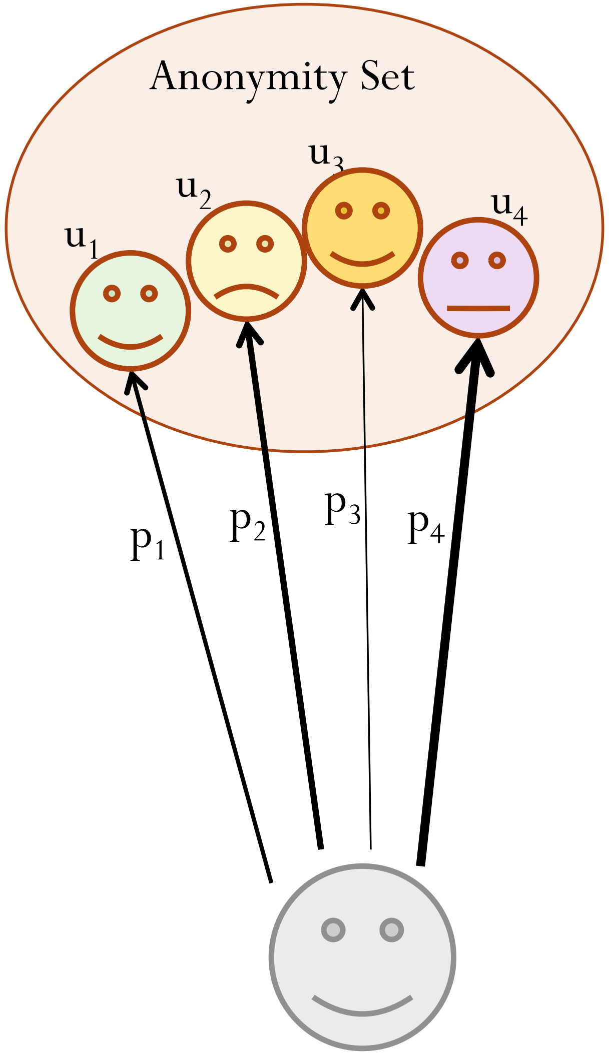


Quantifying anonymity

- Anonymity depends on *both*:
  - The number of subjects in the anonymity set
  - The probability distribution of each subject in the anonymity set being the target
- Entropy: measure of the amount of *information* required on average to describe the random variable

$$H = -\sum_{i=1}^{N} p_i \cdot \log_2(p_i)$$

  - Measure of the *uncertainty* of a random variable
  - Increases with number N of possible values and with the uniformity of the distribution



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

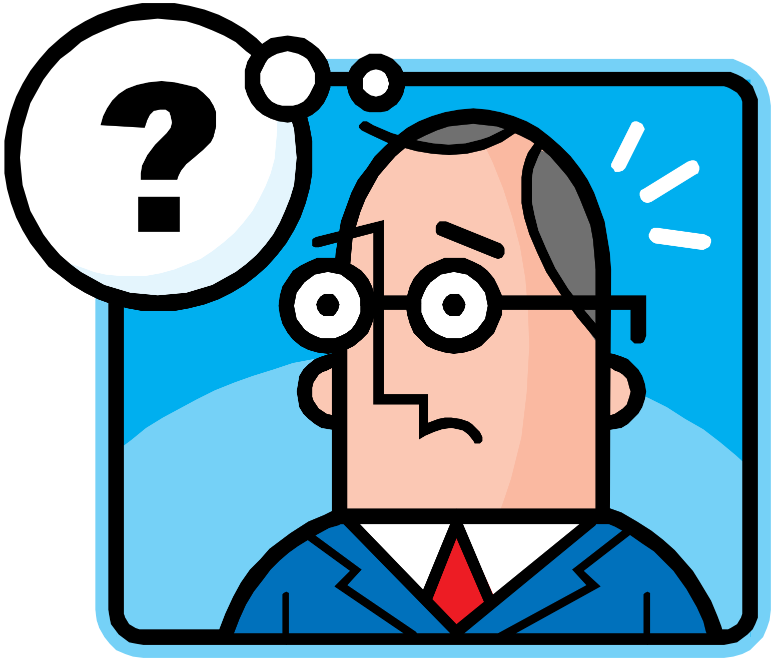
New challenging scenarios

- 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
- 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 (cost)

Thanks !



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