Introduction to Privacy

Claudia Diaz

K.U.Leuven ESAT/COSIC

June 11, 2009

About this course...

- Motivating why you should care about privacy
- Providing a basic understanding of privacy issues
- Broad overview of privacy research: problems and existing solutions
- Focus on interdisciplinarity: technical, legal, and social perspectives
- Aimed at a broad audience, but with a focus on Ph.D. students
- Supported by:
  - Leuven Arenberg Doctoral School
  - BCRYPT (Belgian Fundamental Research Network on Cryptology and Information Security)

Course Program

- Thursday, June 11 2009
  - 14:00 – 15:30 **Introduction to Privacy** (by Claudia Diaz)
  - 15:30 – 16:00 Coffee break
  - 16:00 – 17:30 **Privacy in the electronic communications sector: recent developments** (by Eleni Kosta)
- Tuesday, June 16 2009
  - 14:00 – 15:30 **Privacy and Web mining** (by Bettina Berendt)
  - 15:30 – 16:00 Coffee break
  - 16:00 – 17:30 **Privacy at the communication layer** (by Claudia Diaz)
- Monday, June 22 2009
  - 14:00 – 15:30 **The user experience of social media: implications for privacy** (by David Geerts)
  - 15:30 – 16:00 Coffee break
  - 16:00 – 17:30 **Requirements engineering and privacy by design** (by Seda Gürses)

Overview of this talk

- Motivation
- Privacy properties
- Privacy metrics
- Overview of privacy research challenges
- Questions and discussion

Motivation

Popular arguments against privacy

- "If you care so much about your privacy it's because you have *something to hide*"
- "Surveillance is good and privacy is bad for national security. We need a *tradeoff* between privacy and security"
- "People don't *care* about privacy"

"I have nothing to hide"

- "I don't care about surveillance because I have nothing to hide"
- "If you are so concerned about people/the police/the government knowing what you do, it's because you know you're doing something wrong"
- Solove:
  - "The problem with the 'nothing to hide' argument is its underlying assumption that **privacy is about hiding bad things**."

More from Solove

- "Part of what makes a society a good place in which to live **is the extent to which it allows people freedom from the intrusiveness of others. A society without privacy protection would be suffocation**."

“Learning to live with Big Brother” (The Economist,09/2007)

- “It used to be easy to tell whether you were in a free country or a dictatorship. In an old-time police state, the goons are everywhere, both in person and through a web of informers that penetrates every workplace, community and family.”
- “What they fail to pick up in the café or canteen, they learn by **reading your letters** or **tapping your phone**. The knowledge thus amassed is then **stored on millions of yellowing pieces of paper**, typed or handwritten; from an old-time dictator's viewpoint, **exclusive access to these files** is at least as **powerful** an **instrument of fear** as any torture chamber.”
- “…the **ubiquity** of electronic data-gathering and processing - and above all, its **acceptance by the public** - is astonishing”

East Germany's files

Choose your dystopia

Solove argues it is not so much Orwell's "Big Brother" as Kafka's "The Trial":

- "…a bureaucracy with inscrutable purposes that **uses people's information to make important decisions about them**, yet **denies the people the ability to participate in how their information is used**"
- "The problems captured by the Kafka metaphor are of a different sort than the problems caused by surveillance. They often do not result in inhibition or chilling. Instead, they are **problems of information processing—the storage, use, or analysis of data**—rather than information collection."
- "…not only frustrate the individual by creating a sense of helplessness and **powerlessness**, but they also affect social structure by altering the kind of relationships people have with the institutions that make important decisions about their lives."

Ordering pizza....

Surveillance = Security?

- Law enforcement keywords to justify more surveillance:
  - Terrorism
  - Child pornography
  - Money laundering
  - Crime
- Public opinion pressure on politicians fuelled by high-impact crimes
  - Making legislation as a response to concrete cases

The problems with using surveillance to achieve security

- Strategic adversaries (e.g., terrorists) will adapt to stay under the radar and evade surveillance, while law-abiding citizens will not
  - Surveillance systems can be evaded
    - Knowing the position of cameras to hide your face (CCTV not deterring crime)
    - Adapting behavioral patterns to remain undetected (financial transactions, mobile phone usage, etc.)
    - Vicious circle: all we need is *more* surveillance!
  - Indiscriminate instead of targeted (old times)
    - Shift of resources to electronic surveillance
    - False positives (e.g., no-fly lists, wrong people sent to detention centers)

The problems with using surveillance to achieve security

- Lack of transparency and safeguards may easily lead to abuses
  - Organizations are very keen on protecting their own secrets
    - How they create and use profiles
    - How they are conducting surveillance
  - Corruption of
    - Organizations themselves: Use against political opponents
    - Certain individuals within those organizations: Financial gain
      - Selling information about celebrities
      - Selling profiles

The problems with using surveillance to achieve security

- We run the risk that the surveillance facilities will be subverted or actually used for crime/terrorism
  - Diffie and Landau: "Communication is fundamental to our species; **private communication is fundamental to both our national security and our democracy**."
  - Example: Greek Vodafone scandal: "someone" used the **legal interception** functionalities (backdoors) to monitor: Greek PM, ministers, senior military, diplomats, journalists… (106 people)

The problems with using surveillance to achieve security

- Function creep: where do we stop?
  - Once the capability is in place, why to use it to do *more*?
- People change their behavior when they know they are being watched
  - Would you be spontaneous if you know anything you say may later be used against you?
  - Need for privacy to develop new ideas, new movements
- Current technologies enable surveillance capabilities for private entities
  - CCTV, Telecom operators, ISPs, Social Network providers
- Information asymmetries → power asymmetries
  - Made worse by the fact that you do not know what they know about you

People don't care about privacy?

- In the real world, people are keen on controlling information related to them
  - Who they tell what
    - You might be willing to tell your best friend that you had an argument with your espouse, but you don't want everybody to know about it
  - Concerns over information taken out of context
    - A picture taken at a crazy party being available to a potential employer
  - We value friends who are discreet and keep our secrets
    - We give more information to people we trust
- The cost of gathering and analyzing information without advanced technologies has guaranteed that we had a rather high level of privacy protection

People don't care about privacy?

- Impression management / self-presentation
  - The process through which people try to control the impressions other people form of them
  - Construct an image of ourselves to claim personal identity
- Personal safety
  - Valuable items in an empty house
  - Child alone at home
  - Vulnerability to manipulation:
    - Smart supermarket that makes you spend more
    - Identifying frustrated individuals and recruiting them for e.g., terrorism
  - Personal revenge
  - Identity theft

This information is not necessarily secret, but do you want to broadcast it?

- Identity attributes
  - Name, age, gender, race, IQ, marital status, place of birth, address, phone number, ID number…
- Location
  - Where you are at a certain point in time, movement patterns
- Interests / preferences
  - Books you read, music you listen, films you like, sports you practice
  - political affiliation, religious beliefs, sexual orientation
- Behavior
  - Personality type, what you eat, what you shop, how you behave and interact with others
- Health data
  - Medical issues, treatments you follow, DNA, health risk factors
- Social network
  - Who your friends are, who you meet when, your different social circles
- Financial data
  - How much you earn, how you spend your money, credit card number, bank account

Privacy = Security Property

- Individuals
  - Freedom from intrusion, profiling and manipulation, protection against crime / identity theft, flexibility to access and use content and services, control over one's information
- Companies
  - Protection of trade secrets, business strategy, internal operations, access to patents
- Governments / Military
  - Protection of national secrets, confidentiality of law enforcement investigations, diplomatic activities, political negotiations
- Shared infrastructure
  - Despite varying capabilities infrastructure is shared
  - Telecommunications, operating systems, search engines, on-line shops, software, . . .
  - Denying security to some, means denying it to all: *crypto wars redux?*

Privacy properties

What is privacy?

- Abstract and subjective concept, hard to define
- Dependent on cultural issues
- A couple of popular definitions:
  - "The right to be let alone"
    - Focus on freedom from intrusion
  - "Informational self-determination"
    - Focus on control
- How do we formalize privacy properties in computer systems?

“Soft” vs. “hard” privacy

- Hard privacy
  - Focus on data minimization
  - Adversarial data holder / service provider / environment
- Soft privacy
  - Keywords: trust, data security, liability
  - Policies, access control, right to correct information
  - Threats: 3rd parties, corrupt insider in honest service provider, errors
  - BUT user has already lost control of her data:
    - Millions of exposed records per year due to data breaches at businesses, government agencies and other institutions
- This talk focuses on “hard” privacy

Privacy properties from a technical point of view: Anonymity

- Hiding link between identity and action / piece of information. Examples:
  - Reader of a web page, person accessing a service
  - Sender of an email, writer of a text
  - Person to whom an entry in a database relates
  - Person present in a physical location
- Pfitzmann-Hansen terminology:
  - "*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"
  - "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

Privacy properties from a technical point of view: Unlinkability

- Hiding link between two or more actions / identities / pieces of information. Examples:
  - Two anonymous letters written by the same person
  - Two web page visits by the same user
  - Entries in two databases related to the same person
  - Two people related by a friendship link
  - Same person spotted in two locations at different points in time
- Pfitzmann-Hansen terminology:
  - "*Unlinkability* of two or more items means that within a system, these items are no more and no less related than they are related concerning the a-priori knowledge"
  - Focus on the information leakage of a system

Privacy properties from a technical point of view: Unobservability

- Hiding user activity. Examples:
  - Impossible to see whether someone is accessing a web page
  - Impossible to know whether an entry in a database corresponds to a real person
  - Impossible to distinguish whether someone or no one is in a given location
- Pfitzmann-Hansen terminology:
  - "*Unobservability* is the state of items of interest being indistinguishable from any item of interest at all"
  - "*Sender unobservability* then means that it is not noticeable whether any sender within the unobservability set sends."

Privacy properties from a technical point of view: Pseudonymity

- Pfitzmann-Hansen terminology:
  - "*Pseudonymity* is the use of pseudonyms as IDs."
  - "A *digital pseudonym* is a bit string which is unique as ID and which can be used to authenticate the holder"
- One-time pseudonyms / persistent pseudonyms / everything in between
  - One-time pseudonyms: anonymity
  - Persistent pseudonyms: become an identity
- Possible to build a reputation on a pseudonym
- Possible to have multiple pseudonyms for different purposes
- Examples:
  - Publishing a blog or comments under a pseudonym
  - Using a pseudonym to subscribe to a service

Privacy properties from a technical point of view: Plausible deniability

- Not possible to prove user knows, has done or has said something
- Examples:
  - Off-the-record conversations
  - Resistance to coercion:
    - Not possible to prove that a person has hidden information in a computer
    - Not possible to know that someone has the combination of a safe
  - Possibility to deny having been in a place at a certain point in time
  - Possibility to deny that a database record belongs to a person

Taxonomy of Privacy [Solove]

- Privacy threats we are trying to protect against (out of 16 identified by Solove)
  - Surveillance: monitoring of electronic transactions
    - Preventive properties: anonymity, unobservability
  - Interrogation: forcing people to disclose information
    - Preventive property: plausible deniability
  - Aggregation: combining several sources of information
    - Preventive property: unlinkability
  - Identification: connecting data to individuals.
    - Preventive properties: anonymity and unlinkability

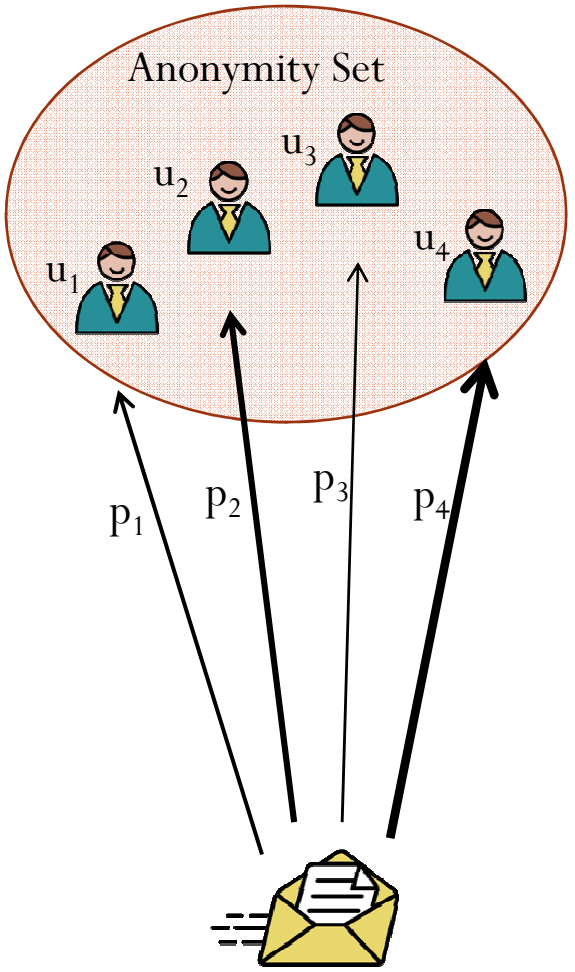
Privacy metrics

Can we “measure” privacy?

- Need to specify
  - Privacy properties we want to achieve
  - Details of the system (the Devil is in the details!)
  - Adversary model: goals and capabilities
- Typically, adversaries are able to obtain probabilistic information. Examples:
  - Probability of a person being the anonymous subject we want to identify (limited number of people in the world)
  - Probability of two information items being related to each other (e.g., two web page requests coming from the same user)
- Many proposals, open research field
  - Examples for anonymity metrics

Anonymity in communication systems

- First approaches
  - Number of subjects in the anonymity set
  - Probability assigned to a subject
- 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



Information-theoretic anonymity metrics [DSCP02, SD02]

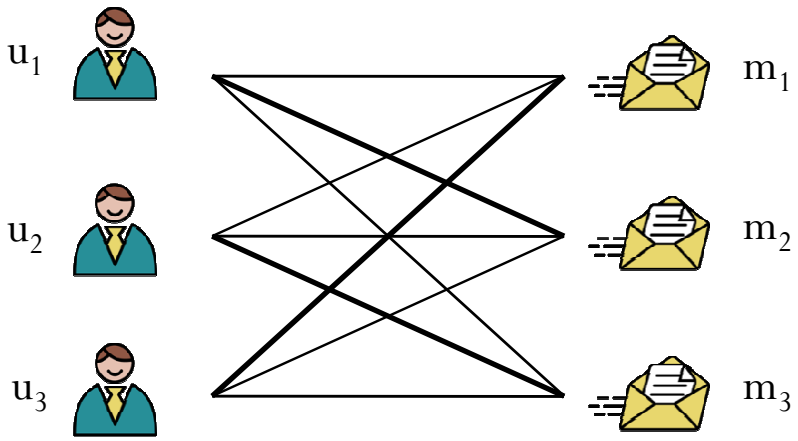
- Entropy: measure of the amount of *information* required on average to describe the random variable
- Measure of the *uncertainty* of a random variable
- Increases with number N of possible values and with the uniformity of the distribution

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

- Distribution with entropy H equivalent to uniform distribution with 2^H subjects
- Other information theoretic metrics: min-entropy, max-entropy, Rényi entropy, relative entropy, mutual information, ….
- A similar approach can be taken to measure unlinkability

Combinatorial approach [Edman]

- Consider deanonymization for a system as a whole (instead of individual users)
- Find perfect matching inputs/outputs
- Perfect anonymity for *t* messages: *t!* equiprobable combinations



Anonymization of data

- 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
  - Nodes in an anonymized social graph

Anonymized records?

- 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 [Sam]

- 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*}

Privacy metrics: challenges

- Modeling the background information available to the adversary
  - What kind of prior information / other sources of information does the attacker have access to?
- Modeling user behavior
  - Are users going to behave as we predict? What if they do not?
- Finding all the information leaks and the optimal way to exploit them
  - Is the privacy evaluation considering the best attack?
  - How does privacy degrade over time?
- Finding expressive metrics
  - How to interpret the result?
  - What is a "good" level of privacy?
- Metrics that generic enough for a variety of systems
  - Many proposals are ad-hoc

Overview of Privacy Research Challenges

Problems not quite solved yet...

- Defining the privacy requirements
- Including privacy principles in the design phase
- Hard to "add privacy" later on

Problems not quite solved yet...

- Finding robust and secure mechanisms
  - Proposed techniques keep on getting broken
  - Secure implementation is even harder

Problems not quite solved yet...

- Usability issues (e.g., design of privacy settings)
  - What can we expect users to understand and manage?
  - Automatization vs control: can we predefine all the possible situations that may arise?

Problems not quite solved yet...

- Economic incentives
  - Who pays for privacy?
  - Privacy techniques very costly: complexity, overhead, lower QoS, diminished functionality
  - Privacy invasive technologies are better funded than privacy enhancing technologies
  - Tragedy of the commons

Problems not quite solved yet...

- Awareness and transparency
  - Do we know what happens to our data?
  - Who collects it?
  - For which purposes is it used?
  - What profiles do they build on us?
  - What are the consequences?

Problems not quite solved yet...

- Legal compliance
  - Some hard privacy technologies may not be compliant
    - Legal systems are often national, while technologies are transnational
  - Implementation of soft privacy techniques (e.g. privacy policies and access control)

Active research areas

- Data anonymization of database records and other data structures (e.g., network graphs)
- Private communication (prevention of traffic analysis)
  - Anonymous and covert communication
- Crypto protocols
  - Privacy-enhanced authentication and identity management
  - Operations in the encrypted domain
  - Anonymous search and retrieval of information
  - Privacy-preserving biometric authentication
- Location privacy
- Ubiquitous environments
  - Principle of data maximization
  - Constrained devices
  - Securing the physical link
- Social networks

Conclusions

- Although we have an implicit understanding of what is privacy, privacy challenges and not yet fully understood
  - Define precisely what it means
  - Understand how our privacy is affected by new technological developments, and what it means for our social structures
  - Translate it into concrete properties for computer systems
  - Evaluate the degree of privacy protection
- Privacy enhancing technologies are far from mature
  - Security and robustness
  - Cost, incentives, usability, tensions with functionality
  - Active research area full of challenges
- Privacy is not "opposed" to security, but rather a security property
- You should care about privacy

Thanks !