universität wien

# MASTERARBEIT

Titel der Masterarbeit

Risks and State of the Art in Privacy and Trust

Verfasserin

Barbara Lang, Bakk.

angestrebter akademischer Grad

Magistra rer. soc. oec.

Wien, 2009

Studienkennzahl lt. Studienblatt: A066 926
Studienrichtung lt. Studienblatt: Wirtschaftsinformatik
Betreuer/Betreuerin: Prof. Dipl.-Ing. Dr. A Min Tjoa, Dipl.-Ing. Mag. Dr. Thomas Neubauer

## Eigenständigkeitserklärung:

Hiermit versichere ich, die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie die Zitate deutlich kenntlich gemacht zu haben.

Wien, den 5.März 2009 Barbara Lang

## Abstract

The WWW offers easy access to information. Few keystrokes and mouse clicks suffice to catch up on certain topics a user may be interested in. Modern technologies also offer (cheap) ways for people to communicate with users all over the world. Writing email messages, entering chat sessions or even browsing one's favorite website may not stay as private as most people think though. Businesses already discovered possibilities for advertising products by monitoring people's browsing behavior; moreover, hackers may be able to steal sensitive data, even without people's knowledge.

This thesis will cover different aspects of privacy issues. The first chapter will focus on privacy in general and how the attitude toward it changed in history and after September 11 2001, respectively. Measures and laws related to privacy in the US and Europe will be compared (e.g. juristic approaches to foreigners, measures allowed during prosecutions) and critiques triggered by changes after 9/11 are introduced.

The main matter of this thesis will concentrate on risks and threats (e.g. spyware, data mining, collection of data) to privacy and sensitive data. Users are often not aware of the extent of these threats when using the WWW. To counter the issues related to these risks, Privacy Enhancing Technologies (PETs) aim to improve the protection of private information. Different PETs, which provide different levels of anonymity, are introduced and discussed. These PETs include anonymizers that mask a user's IP-address as well as pseudonymizers, which help to manage online identities.

The final chapters deal with future challenges concerning PETs, which still need to be taken into account, and several current examples of privacy issues.

# Contents

# CHAPTER 1

## Introduction

Since the 90's of the past century, the number of people with access to the *world wide web* has been rising fast. Because of cheaper and more powerful technologies many households in industrial countries do not only own a PC or similar devices, but have also an option for internet access. With easy access to information and the possibility for cheap communication with people all over the world, 75 %[1] of households in Germany and 69 % in Austria currently have internet access. Staying in touch with family and friends even over great distances, shopping without leaving home, planning vacations or just collecting information about interesting topics, everything is now possible, even for people which may otherwise not be technically experienced. Inexperience, curiosity and unsophistication is often utilized by attackers (hackers) using malicious codes like Trojans or Spyware applications (cf. Chapter Risks to Privacy and [TFS04]) in order to get access to remote machines and alter, delete or steal (sensitive) data. Worms and Trojans are often attached to apparently legitimate email messages or the message content itself contains links to fake websites, which ask people to disclose sensitive information (a furtive method called *phishing*). Malicious software could also be used by (marketing) companies to gain data about people's interests in order to adapt advertisements or by employers seeking to monitor their employees. New possibilities offered by modern technologies are also utilized by law enforcement and governments respectively. The ability to monitor transactions and access huge amounts of data in a short period of time (including the controversial method of getting furtive access to local machines, which are not necessarily owned by suspects of an investigation) supports investigations and prosecutions, but could as well pose a threat to people's privacy because of the huge potential of abuse.

So people are exposed to serious threats and risks while surfing on the internet or receiving messages. Every file which is loaded onto the local machine (e.g. images, videos, executable files ...) jeopardizes the system's security. Malicious codes could be masked as legitimate files and spyware may even be bundled with commercial software (cf. [TFS04]). Security and privacy is not only put at risk by downloading and installing files, but by every request made to webservers/applications (e.g. loading a website). Data packets sent over the data highway are often transmitted in plain text (unencrypted and readable by

1 according to a study conducted by BITCOM, which was mentioned in `http://www.pressetext.at/pte.mc?pte=090108023`

everybody). Tools like communication monitors and *packet sniffers* offer an easy way to take a look at the content and to discern the sender and receiver of these packets. Sender and receiver are represented by *IP-addresses*, which are issued by *internet service providers* (ISPs). The address is sent with every request made and is used by ISPs to identify their customers (companies as well as private individuals). The address is visible to targeted webservers and may also be logged in order to be able to send a response to the right location. Because every request includes the originators IP address, actions taken online by one single user can be linked together. Businesses and marketing companies already discovered great advertisement possibilities by monitoring online behavior of single users to derive possible interests and therefore issue more appropriate ads. On the other hand, law enforcement and other services use these technologies for investigations and prosecutions.

To address the issue of potential privacy loss online, this master thesis will give an overview of how privacy is defined in general and how privacy was handled at the beginning of the computer age (60's-70's). Already in 1948, the *United Nations* included the *right to privacy* in the *Human Rights Declaration* (cf. [Pes05]). During the 60's it was already important to protect stored data from unauthorized access, partly because people were concerned that the information (which may in the future be collected in huge databases) related to them could violate their privacy (cf. [Par68]).

The improving technology and increasing popularity of the WWW, not only made it easier for "ordinary" people to access information of everyday's life affairs, but as well raised the risk of privacy violation. The advantage of faster affair handling offers disadvantages too. Many services provided online require disclosure of sensitive data, which is often stored on remote servers (to make reuse easier). To prevent abuse or unauthorized disclosure, which could be embarrassing, distracting as well as harmful in a financial way, data stored and transmitted information must be protected accordingly. When issues with private information became more pressing the US Federal Trade Commission introduced *Fair Information Practices* (cf. [Pes05]), which consist of guidelines for handling and security of private information. Owners of webservers and services, who ask for sensitive information, have to advise customers of the handling of disclosed data (privacy policy) and besides provide access to keep facts up to date or delete them if desired, while still prohibiting unauthorized access.

Shortly after the turn of the millennium, however, one incident seemed to change everything. During the months after September 11, 2001 (colloquially called *9/11*) many laws were changed and others were created to *prohibit future assaults*. These changes, which also affected people's privacy, were not only limited to the US, where the attacks on the WTC and the Pentagon occurred, but happened all around the world. Several sections of this thesis will address the changes made in US and European nation laws (especially Germany) and the effects of these changes (not only on people's privacy). After comparing juridical and technical changes, one section includes an overview concerning criticism related to privacy loss, triggered by the taken measures.

The main matter of this thesis, however, focuses on the current risks people are exposed to, when participating in online transactions. Like already mentioned above, easy accessing and storing of (sensitive) information poses a threat to people's privacy, because of the huge potential of abuse. Risks users are exposed to range from tools enabling the tracking of actions (*cookies*) and access to remote clients (*spyware*) over organizations, which store (and may share) private information (e.g. *ISPs*), to certain methods (e.g. *social*

*engineering*), which are used to gain information.

After describing the risks to a user's sensitive data, *Privacy Enhancing Technologies* (PETs) (cf. [Cra99]) are introduced, to show how people could protect themselves from the above described threats. The chapter addressing these technologies is separated in different sections, where each part represents one approach. Each approach provides a certain level of anonymity, depending on how much data must be disclosed in order to use a service.

*Anonymity* provides the highest level of privacy. The IP address related to a user is masked (replaced by another IP address), hence the targeted webserver is not able to determine the actual originator of the request (the response is sent to the IP address attatched to the request). Anonymity tools like *TOR* and the GUI-based *JAP* offer an easy way to access anonymity networks.

The second approach, *Pseudonymity*, allows users to create pseudonyms (partial identities), usually containing login information like (unique) usernames and passwords. The login data is then used to authenticate at a remote server. With this approach a webserver is still able to "recognize" a user even if he/she uses a different IP address or client.

The third section in this chapter shortly introduces *Encryption*, a method, which aims to protect the actual sensitive data, either during storage or transmission.

*Identity Management* takes a look at the authentication models currently in use and future developments [Cam05]. The basic silo model, which is often used by public boards and discussion groups, is included as well as more complicated architectures, which include *certificate authorities* (CAs) and *personal authentication devices* (PADs) like cellphones, PDAs and chipcards.

*Privacy and Trust Engines*, on the other hand, do not actually protect sensitive data, but support people and user agents (e.g. web browser) in dealing with privacy policies offered by webservices. The protocol *P3P* (Platform for Privacy Preferences) offers a standard, which enables user agents to automatically parse privacy policies and compare them to user preferences - hence helps to improve trust in businesses.

The last section of this chapter will deal with several responses to the mentioned threats. It includes user initiatives, which will show that even the attitude a individual user has toward online transactions could protect or threaten privacy of sensitive data.

A short comparison of different approaches is followed by *future challenges* that still need to be solved to improve the protection of privacy, but nonetheless offer a satisfying level of usability.

The thesis is then concluded by a chapter addressing current privacy issues. These issues include *computer surveillance*, which is already (or will probably soon be) conducted by law enforcement and governments, applications and search engines that may help creating profiles and dossiers about people, and the abuse of easy sharing of information and media (e.g. spreading of rumors, propaganda and copyright infringement).

# CHAPTER 2

## History of Privacy and Current Situation

Walters [Wal01] tried to summarize different views on privacy and the issues (trade-off between privacy, freedom of individual citizens and security of individuals, whole community) that have arisen throughout the centuries. In Greek and Roman eras, people were already seeking to limit surveillance by authorities.

### 2.1 Situation with privacy in general

In the survey by Walters (cf. [Wal01]) different views on privacy were collected and compared. In the literature concerning that subject, privacy is among other things defined as

- *human right* to determine which personal information is communicated to others
- *control* over access to personal information of oneself
- *limited access* to a person (thoughts, body, ...)
- *respect* that basic right of other people

Losing privacy causes deep concern in most of the people. In this situation they feel that their *right to be let alone* and their *realm of solitude* is broken, when (some of their) privacy is taken away. [Lin07]

In 1948, the United Nations ratified a *right to privacy*. Article 12 of the Universal Declaration of Human Rights states (cf. [Pes05]):

> No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks.

Years later, a new threat to one's privacy (and sensitive data) emerged. New technologies became open to the public, one of which was (and still is) the Internet. Many businesses avail themselves of the manifold possibilities of this new platform. In the last couple of

years the importance of *e-commerce* increased. Hence, protection of user data became crucial too.

In general, there are two different approaches to handle privacy of personal information over the Internet. Among EU countries the *opt-in* approach is the most common. This approach requires the collector of personal information to ask users, when disclosed sensitive information is used in a secondary way. The *opt-out* approach, on the other hand, informs users when their information is used in a different way. In this case the user has the right to 'opt-out'.

People have the right to control the use of their personal information and to know why their data was collected in the first place. Exceptions to that right are, for example, a compelling public interest like law enforcement, fraud and health concerns.

Practical experience showed, however, that these concepts of giving people options (to opt-in, opt-out) lack efficiency, particularly during online transaction. People are seldomly asked to make choices concerning the use of their personal data, because the 'opt-out' approach is not available for private use. A study conducted by Peslak (cf. [Pes05]) found that private companies are less inclined to reveal their privacy policies to the public, than (the 50 largest) companies led by the government.

Privacy, however, is not only an issue concerning online shops or businesses. By using online boards, chat rooms and other larger communicating platforms, new privacy issues arise. In order to get an account on such servers you often need to give away personal information, which can be (mis-)used in an illegal way. Several methods exist to protect identity or to establish an anonymous connection (see below '*Privacy Enhancing Technologies*'). Whenever one or more accounts are created that belong to a single user, there are impacts on privacy (listed below, cf. [Lin07]):

*price discrimination:* The shopping habits of users are recorded and used to offer buyers products according to their interests (e.g. Amazon). This information could equally be used to charge different prices from different buyers, depending on the price limit of the individual. Many users are not aware of that. To manage linking of actions (and hence creating a profile) only one identifier (e.g. IP address) is needed.

*manipulation and personal manipulation:* Just the user's profile is needed to link actions. By logging interests and preferences of certain users it is possible to manipulate the presentation of a website to better fit user's interests. Another possibility would be to show only certain articles to the user (e.g. online versions of news papers and magazines) and by that to manipulate the user without even knowing his real identity (only his online profile is known).

*industrial espionage:* Certain online boards are used by professionals and research workers to ask questions or discuss issues concerning their current work. Besides being helpful, such boards offer an opportunity for *industrial espionage*. The history of actions of a particular user can help to determine the current status of development of the company which the user is working for.

Apart from the impacts mentioned above, people often refrain from pronouncing their opinion (online), due to *fear of retaliation*. If a posting is tracked back (linked) to a

certain user (e.g. online boards), he/she is most often not willing to talk about medical or psychological problems. In this case, the maintainance of a higher level of privacy increases the disposition of the users to communicate issues.

Another issue considering the protection of privacy comprises the security of the stored information/data. This sensitive information is stored to make re-login easier (and save preferences). Security (as mentioned above) is therefore mandatory. Developers, however, show little regard for the usability of these security measures. In [Jø07], numerous principles for security usability were defined. Involvement of people in security can be distinguished in two types:

*Security Action:* users have to trigger a security action themselves (e.g. submitting a password)

*Security Conclusion:* evidence that leads users to the conclusion that the connection/data is secure (e.g. icons in browsers that represent secure connections)

To accompany these items, usability types were defined for each of them:

- Security Action Principles:
  1. users must understand which security actions are required
  2. users must have the ability/knowledge to take the appropriate security action
  3. mental and physical load of security actions (for repeated transactions) must be tolerable
- Security Conclusion Principles:
  1. user must understand what is required to support a secure transaction
  2. it must be possible for users to derive the security conclusion from given information
  3. mental load of deriving security information (repeatedly) must be tolerable

Authors of [Jø07] argue that usability is far more important than e.g. stronger cryptography. If users fail to make use of the offered security measures, they are useless.

## 2.2 Before 9/11

Already in the 1960's (cf. [Hof69]), people expected that databank grids would be created in the near future. These grids would have allowed inexpensive sharing and reproducing of data. The establishment of a national databank would have eliminated a lot of defects like the loss of records, the absence of standards and problems during the development of statistics. Organizations like law enforcement and credit bureaus could have benefited from the possibility of creating dossiers and enabling access to criminal records. Due to insufficient safeguards, this scheme could not be realized.

Besides the access, also input and retrieving of data should be controlled in order to avoid misuse and consequently the decline of public trust. Trust[1], however, is important for a fully functional data collection system. Users providing incorrect or false information renders the collection of data useless.

In 1966 a *Computer Bill of Rights* was suggested (cf. [Hof69]):

- no organization is allowed to store data that covers a large number of people
- rules which control access to data are publicized and programs that enforce such rules are open to everyone
- one individual has the right to read his own file and restrict access to certain entries
- access of files is recorded, as well as authorization
- if illegal access occurs, the targeted individual has the right to sue

The same year (November 11, 1966) the Association for Computing Machinery (ACM) composed guidelines named *Professional Conduct in Information Processing* (cf. [Par68]), which were partially adopted from ethical guidelines of other professions. This association tries to communicate technology and motivate students to work in that profession. These tasks, however, require guidelines of conduct. The guidelines published in 1968 include principles of loyalty (towards the employer, the client and public). A member of the ACM needs a certain qualification before he/she is allowed to publish articles (or to build applications). Moreover, passing information or advice to another profession is strictly forbidden. The member is requested to take responsibility and consequences for potential harm (and malfunction) of their work. When that article was published, it was not clear how the enforcement of the guidelines should be effected.

One year later, in 1967, a *Rights to Privacy Act* was proposed, which banned wiretapping and electronic eavesdropping. Contemporaneously, the state California had passed laws that recognize a person's right to privacy. The current *Privacy Act* in the US dates back to 1974 (cf. [PTA]).

Apart from protecting people's privacy by issuing laws, concepts were explored and developed to keep data, that is stored on electronic devices, safe. The technical methods listed in [Hof69] are intended for the realization and protection of time-sharing systems (some of them are mentioned below). The methods allow lock-out of unauthorized users and forbid certain users to interfere with the monitor by tampering with commands. Hence, only certain users have access to more 'privileged' commands. In addition to these methods that should protect computer memory, the article also suggests measures to control access to files.

A well known method to control access is *password protection*, although this measure can be compromised. During that time it was common practice to put data of different levels of sensitiveness into one pool. Thus the data could be used by everyone who had

1 ability to rely on a (business-)partner that agreements are followed and transaction will be concluded as agreed

access to that pool. Hence people could have access to superfluous data, whereas at the same time they can not gain access to essential data, since it is stored in a pool with a higher protection level.

A scheme that included so-called *authority items* was indeed a working solution, but raised new issues as well. The access of every single user to a certain entry within a certain file has to be recorded. Large numbers of entries would also raise the consumption of storage, memory and processing time (issues that vanished in modern times or were slightly altered, due to economic storage and extended memory capacity).

For finding adequate measures to control file access, Hoffman also identified the *most important threats to information privacy.*

| | |
|---|---|
| **Accidental** | User Error<br>System Error |
| **Deliberate, passive** | Electromagnetic Pick-Up<br>Browsing |
| **Deliberate, active** | Wiretapping<br>Masquerading as another user<br>'Between Lines' entry while user is inactive but on channel<br>'Piggy Back' entry by intercepting and transmitting an error message to the user<br>Core Dumping to get residual information |

**Table 2.1:** Threats to Information Privacy (1969) [Hof69]

Suggested measures already developed in the 60's to protect information privacy are listed below:

- *Access management* should prevent unauthorized users from gaining access to sensitive data. Passwords alone, however, are in this case insufficient. The important problem remains the authentication of the user's identification. *One-Time passwords* are a solution to that issue. As the name suggests, a certain password is just used once, afterwards it becomes invalid.
- *Privacy transformation* is used to encrypt data, hence conceal it from unauthorized users. Types of transformation are substitution, transposition and addition, which can be combined to increase the *work factor* to break the code. Variables that influence the work factor are:
  - length of the key
  - size of the key space: area of the key (e.g. letters with or without numerics)
  - complexity
  - error sensitivity: effect of transmission errors and malfunction of processor could avoid deciphering

- *Threat monitoring* includes, among others, recording of all rejected attempts to enter the system and certain files. Regular reports could reveal possible misuse and tampering.

Years and decades later (in 2000), while technologies kept advancing, a new threat to people's privacy was identified: data transfer and collection over the Internet. As an answer to that issue the *Federal Trade Commission* composed the *Fair Information Practices* (cf. [Pes05]):

1. *Notice* - data collectors must disclose their information practices before collecting personal information from consumers
2. *Choice* - consumers must be given options with respect to whether and how personal information collected from them may be used for purposes beyond those for which the information was provided
3. *Access* - consumers should be able to view and contest the accuracy and completeness of data collected about them
4. *Security* - data collectors must take reasonable steps to assure that information collected from consumers is accurate and secure from unauthorized use

Users visiting the website must be notified of privacy policies and the options about storage of sensible data. Access to personal information should be granted to that particular user (in case the data stored is not correct or has to be changed). Eventually, owners of such web pages are responsible for adequate implementation of security measures to protect important or sensible data.

## 2.3 After 9/11

After the attacks on the World Trade Center a new fear of terrorism emerged. In the current situation, people of certain religious affiliation or academic interests may be selected as potential targets. Logging and analyzing people's online actions would render the localization of suspects easier (cf. [Lin07]).

The terrorist attacks not only fueled renewed fear of terror in the United States, but also in many countries all over the world, because the alleged terrorists were previously inconspicuous and law-abiding. Only months later laws were changed and legislated to answer the terror threat (e.g. *Patriot Act* in the US).

On September 28 2001 the United Nations adopted Security Resolution 1317 (cf. [Ban08]), which asks member states to take measures (e.g. adopt laws) to fight terrorism.

### 2.3.1 Changes in Europe and the European Union (EU)

Few weeks after 9/11 the European Union introduced a framework, which included definitions of terrorism and criminal sanctions. Apart from definitions, laws were adopted

that *restrict access to information in the name of national security.* According to [Ban08], a paper that analyzed the situation of journalists (access to information, protection of sources, ...) after 9/11, these new adopted laws were used in many cases (e.g. France, Germany, ...) to investigate and prosecute journalists. Many countries amended their laws in adjustment to the EU recommendation, regarding information access. Ireland, for example, amended the Information Act so that *no harm must be shown* in order to block public access to information on behalf of national security (and national security-related secrets). Laws not only affected access to information, but also the ability of *taking photos of notable events, public protests and places.* Photographers and journalists are often forced to delete the photos or otherwise these photos are used in prosecutions. In Greece, 14 UK and Dutch "plain spotters"were arrested for several months after taking photos of military planes during a public air show. Freedom of expression was also limited in form of laws, which criminalize *speech that seem to encourage, either directly or indirectly, terrorism.* Laws were adopted, which even criminalize glorification, apology and public promotion of terrorism (e.g. UK Anti-Terrorism Act 2006). Additionally internet-based speech is also strictly controlled. Websites were blocked or removed, which contained controversial material. As of August 2008, Turkey blocked 853 websites, including famous international social networks like YouTube and sites like Geocities. Russia has even stricter regulations where over 1000 sites were banned in 2007. The Terrorism Act in the UK requires ISPs to remove materials that *promote or glorify terrorism.* Some member states (e.g. Czech Republic, Bulgaria) even adopted laws stricter than required by the international agreements. In Russia Anti-terror laws adopted in 2006 define extremism as follows: *"Public defamation of the person, deputy public office, the Russian Federation or public office, subject of the Russian Federation, in the performance of their duties or in connection with their performance."* [Ban08]

The Commission of European Union (CoE) in 2006 and the UN Human Rights Committee in 2008 found that many national laws adopted were problematic because of the broad scope and lack of limitations (many laws were not limited to terrorism-related investigations/actions). One year later the assembly suggested that the vague provisions should be replaced by more clear and specific provisions.

Laws that were legislated in Germany (cf. [MH05] and [ECK]) gave law enforcement and intelligence services new and enhanced competencies. Now informationexchange between these two services is easier - hence it is easier to find and investigate potential suspects.

The so-called "Terrorismusbekämpfungsgesetz" (anti-terrorism law) includes changes to identification and privacy laws. At the moment it is possible for certain services (like Bundeskriminalamt - similar to the FBI in the US) to get information about "suspicious" people without their knowledge, while in the past a court order or similar was necessary to start an investigation. In fact, the law enforcement or intelligence service is not allowed to let people know that they are being investigated.

Long before 9/11, a paragraph existed that made it illegal to create or promote terror organizations. It was part of the *Anti-Terror Gesetz* already issued in Germany in 1976 to address problems caused by the *Rote Armee Fraktion (RAF)*[2] (129a StGB "Lex RAF"

2 terror organization involved in several murders and killings throughout the 70's and 80's.

cf. [AUS06]). Since this law was established it was often misused to threaten or actually conduct arrests (cf. [FP04]). On April 26, 2002 that law was further enhanced. So already the promotion of an alleged foreign terrorist organization is a criminal act.

Changes made by the 'Anti-Terror Paket'[3]:

Changes made to laws by the "Anti-Terror Paket" are listed below (cf. [ECK], [FP04] and [MH05]):

*Enhanced Competencies for the Federal Office for the Protection of the Constitution:* The BfV [4] in Germany is authorized to collect and analyze information, when the linked investigation serves purposes of protecting international understanding and peaceful cohabitation of people. Under certain premises the BfV is allowed to collect information from telecommunication and bank companies. Strict limitations control the use of so-called IMSI-Catchers, which allow the localization of people through active mobile phones and the collection of device- and SIM-card numbers. Information gained by collecting telecommunication data can be helpful when investigating the environment of people (e.g. to find suspicious behavior).

*Security Checks:* If people want to attend a job, which is *vital* and *essential for defense*, security checks are conducted. Before 9/11, this security check was only mandatory for people working for civil services etc. With these new laws in place nearly everyone can be checked up on, to make sure that trustworthy people are employed at sensitive places. Which jobs actually fall into this category is not further detailed. The investigated person has to disclose personal data about himself, his family, relatives and sometimes even friends. This data includes among name, address also information about former professions, vacation destinations and of course rap sheets. After the data is collected, several agencies have access to that information and are consequently allowed to store and make use of it.

*Passports and Identity Cards:* Since 2005, biometric data is stored on passports. Strict specifications are used to validate the taken picture, whereas a copy of that picture is saved on a RFD-chip, which holds fingerprint-data since 2006. For foreigners the situation is even more alarming. Biometric data has to be stored within the resident permit, following that they are proverbially criminalized by law.

*Monitoring of Bank Accounts and Transactions:* The Federal Office for the Protection of the Constitution has the authority to monitor bank accounts, transactions, telecommunication and electronic transmissions (e-mails, visited web pages, ...). With the *IMSI-Catcher* it is possible to get data like IMEI (International Mobile Equipment Identity) from a certain suspect, who tries to log into a mobile net. The monitoring of calls is possible too.

*Enhanced Organizational Affairs Law:* Societies, clubs, unions, ... can be forbidden, if they are likely to become a medium for terrorism and intolerance - observing the

3 coloquial term for the 2nd draft of anti-terror laws
4 Bundesamt für Verfassungsschutz

activities of such foreign clubs becomes important regarding possible terroristic actions.

*Enhanced Competencies of Federal Office of Criminal Investigation:* The BKA [5] has the authority to investigate cyber-crimes, which could endanger national inner and outer security. To complete already collected evidence, the BKA has the right to gain more information, without ascertaining that other agencies and organizations are already in possession of that particular findings.

Federal agencies and law enforcement are now allowed to share collected information with each other, hence these two authorities are not clearly separated anymore. The authorities were separated in the first place due to experiences in the past. During World War II as well as in *der deutschen demokratischen Republik* (DDR) national security was maintained by one single authority (Reichssicherheitshauptamt and Staatssicherheitsdienst respectively) (cf. [MH05]). In both cases predominated a totalitarian regime where nearly every little aspect of people's lives was monitored and controlled. Nowadays, since the separation of powers was abolished, privacy and freedom of people is again in danger.

*Immigration:* Foreign people, who were part of terror activities are no longer getting visa or stay permits. In favor of those new rules, people are expelled from the country at once. Recorded speeches of immigrants (conducted with their knowledge) allow an analysis of the immigrant's land of origin. The storage of fingerprints allows a comparison with findings of BKA at crime scenes.

*Dragnet Investigation:* was established in the 80's as an activity against organized crime.

In Austria the *Bundesamt für Verfassungsschutz und Terrorismusbekämpfung* (BVT) fulfills the function of protecting the nation's constitution. Similar to Germany the BVT is allowed to collect sensitive information in order to conduct security checks. Scrutinies like these are strictly regulated (cf. [GES]). As long as requirements for permissions are fulfilled security checks are repeated regularly. The sharing of data (between security agencies, law enforcement) is allowed, if permitted by the investigated individual him-/herself. In some cases circumstances are disclosed for data sharing without permission (e.g. information is vital and time is too short to ask for permission). The BVT is allowed to process collected information, if a warrant is issued or an investigation is conducted against an individual.

### Critiques

Since the new laws have been executed, it got much easier for law enforcement and agencies to obtain (private) information about people (even without people's knowledge). Biometric properties are stored in ID-Cards and passports. It is not determined whether that information is just used for better and quicker identification, or for prosecution as well. Concerning foreigners, information stored on ID-Cards (like visa and stay permit) could be used for crime investigations.

5 Bundeskriminalamt

### 2.3.2 Changes in the United States (US)

The attacks on the WTC and the Pentagon left US citizens shocked. As a reaction to this new terror threat and the fears that emerged thereby, new laws were proposed. Shortly after 9/11 the *Anti-Terrorism Act of 2001* (also known as *Patriot Act*) was created.

The term PATRIOT is an abbreviation for "Providing Appropriate Tools Required to Intercept and Obstruct Terrorism", which should demonstrate the reason for these rather fast changes in the formulation of laws. On October 26, 2001 the amendments were legitimated by George W. Bush, former President of the United States.

That Act contains numerous amendments which *expand the authority of law enforcement and intelligence agencies to monitor private communication and access of personal information.* [OVE]

Furthermore it increases the possibilities to implement and place pen registers (devices which collect outgoing numbers from a specific telephone) as well as trap and trace devices. To adapt the amendment to modern technologies, it contains recordings of all computer routing, addressing and signaling information. The act created, changed and affected many different kinds of laws. Interception of communication was supported before by a framework consisting of several laws (cf. [OVE]):

*Title III:* requires a probable cause when placing wiretaps and similar technologies and imposes strict limitations upon the ability to get hold of information content. The interception must be connected to a certain crime and is only authorized for a limited period of time. Any information that is obtained while these statutes are violated can not be used at court.

*Electronic Communication Privacy Act (ECPA):* controls the access to stored emails and other electronic communication, including real time interception. Requires just a governments attorney's statement about the usefulness of information concerning an ongoing criminal investigation.

*Foreign Intelligence Surveillance Act (FISA):* allows electronic surveillance of (American) people or targets, which are suspected as a foreign power or an agent of an foreign power.

This framework was later amended by the Patriot Act.

#### Contents of the Patriot Act

The Act induced many changes in different kinds of laws and amendments. The most important changes (and which are most troubling in terms of privacy protection) are listed below (cf. [FP04] and [ANA]):

*Probable Cause:* Owing to the new laws, people can be investigated, who are suspected to be terrorists or connected to a terrorist organization, although it is not clearly defined how a person can achieve that particular status of suspicion. It is not obligatory for investigators to name a reason for their actions.

*Wiretapping:* Before 9/11 for every single telephone (line) or computer to be wired, a warrant was required. With the Patriot Act this requirement was removed, rendering multi-point (roving) wiretap possible. Although a court order is still required to use wiretapping devices, the display of probable cause is not mandatory. Voice-Communications, like email can now be obtained by search warrant, whereas the more strict wiretapping limitations do not apply. Intercepting the communication of a "computer trespasser" is allowed, as long as the owner of the particular machine authorizes it. Hence investigators are allowed to wire every telephone (computer), which will probably be used by the suspect. In such cases publicly used telephones could be wired too. Thus calls of people may be recorded, which have nothing to do with the ongoing investigation. Conversations of people, not yet incriminated, are recorded, which means a clear offense against one's privacy.

*Immigration:* Non-US citizens, who want to immigrate into the US can be arrested, if suspected to be terrorists or being part of a terror organization. The suspect can be held in custody for the maximum time period of seven days. Unless charges are pressed during that period of time, the suspect has to be set free. Shortly after the attacks many Muslim people were victims of that particular rule. On travels to the US people of Muslim origin may be arrested and questioned. During the questioning that followed, they had to disclose the reason of their visit to the US, moreover, they had to give personal information as well (connection to Military Forces, weapons-training, ...).

*Information Sharing:* Law enforcement and intelligence organizations are now allowed to share data and investigation results with each other. Every agency usually collects data concerning a limited number of topics: personal data, bank accounts, medical information, etc., thus forming a "big picture" of a suspected individual. The lack of separation between law enforcement and intelligence organizations raises a new threat to privacy.

*Money Laundering:* The corporation of foreign financial institutes with the US government is obligatory, otherwise these countries could face serious sanctions. This corporation includes sharing of customer information and their financial status. If this information is not provided, the corporation along with the bank in question could be canceled by US government. This amendment literally removes banker's discretion.

*Foreign Intelligence (FISA):* Proof that the investigated target is an agent of foreign power is no longer required. Hence the reason of gathering foreign intelligence could probably be used for investigations which in fact do not affect foreign powers. Although FISA specifications can not be applied during investigations against US citizens "solely based on activities, which are protected by the First Amendment [6]"

*Secret Searches:* This particular amendment raises the most important privacy issue concerning the Patriot Act. Thanks to that changes, searches without presence of

6 Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances.

the owner are allowed. Furthermore, these searches and collection of evidence can now be conducted without the owner's knowledge. Notice of searches has only to be given within a "reasonable period" (the term is not well defined) after the searches have been conducted. This searches include all acts which *constitute evidence of criminal offense in violation of the laws of the United States.*

*Sunset Clause:* Certain amendments of the Patriot Act contain a sunset clause (also known as *sunset provision*). These specific laws expire after a short period of time, if no further legislative action is taken. In case of the Patriot Act amendments concerning wiretapping, privacy is bound to this provision. Until today these laws have always been renewed after the period had ended.

#### Critiques

Although these extensions to authorities of law enforcement and intelligence organizations came into existence to fight terrorism, particular amendments are not limited solely to investigation of terrorist activity. Collection of information got much easier and privacy got even more limited. For most investigations a probable cause is not needed anymore and searches can be conducted even without prior notice of the target's owner. Nearly every piece of information that was collected can be shared with other agencies and organizations. Every foreign bank institute, which has a branch office in the US, is bound to corporate with US authorities (and consequentially has to disclose information about costumers). Investigation of foreign agents got much easier so that without proof, resources can, in fact, be used against US-citizens. Non US-citizens, on the other hand, have literally no rights anymore. People, who plan to enter and stay in the US are checked and investigated more closely, before they are allowed to settle down.

### 2.3.3 Comparison and Effects of Changes

After the attacks, the US as well as the EU were surprisingly fast in creating new laws and changing already existing ones, to be prepared for the terror threat. Most of these changes that were made in the US and the EU (Germany) are indeed quite similar (cf. [PEU05] and [PUS05]):

*Anti-Terror Package:* While the US adopted the *PATRIOT Act* in October 2001, the EU introduced the *Hague Programme for Freedom, Justice and Security*, which focused on illegal migration, organized crime and terrorism. The program increased international data sharing and monitoring of communication as well as financial transactions.

*Registration and Monitoring:* When the US (Department of Homeland Security) asked for access to data of (flight-)passengers, who want to enter the country, the EU denied the request at first. After extensive arguing, however, the data was not only sent over to the US, but stored in a central database, allegedly for cost and efficiency reasons. Access to that database is not limited to combating terrorism and capital crimes though.

To monitor and register movement, the *Schengen Information System (SIS)* was already introduced in 1995, when the internal borders between France, Germany, Luxembourg and the Netherlands were removed. To the IS people were added, who are wanted for arrest, mostly subjects of investigations, people who were not permitted to enter at external borders. Data can be accessed by border authorities, police and security/intelligence services. In the future biometric data could be added to the information stored in the database (SIS II[7])

*Registration of Foreigners:* The US *Visa Entry Reform Act* was introduced in 2002 and suggested to combine data of different databases for automated monitoring of entry and departure of foreigners, leading to the *National Security Entry-Exit Registration System* (NSEERS). Not only immigrants of "terrorist nations" are fingerprinted and processed, but also visitors, who may pose a threat to national security. Since 2003, the NEERS is integrated in the *U.S. Visitor Immigration Status Indication Technology System* (US-VISIT), which registers all visitors entering the US.

Justification for the implementation of US-VISIT was the fact, that Europe already kept track of visitors. After 9/11, however, the EU decided to implement the *Visa Information System* (VIS), which should also hold biometric data.

*Registration of Citizens:* In 2004, the European Commission decided to include biometric data like face and fingerprints in passports. The data is stored on a chip on the passport and the data can only be accessed by "authorized entities". Similar standards apply to the US, while only a small percentage of American citizens actually posses a passport. Hence the driving licenses are considered to be standardized in order to provide a nation-wide identification.

*Data Mining and Profiling:* Data mining could be a powerful tool for combating terrorism by finding hidden patterns, relationships and predicting future results. While in Europe enhanced access to SIS II could achieve that goal, the US introduced a program called *Total Information Awareness*. A system that is able to scan and process personal data of people in- and outside of the US. TIA raised concerns, because of the vast potential of abuse. The concerns resulted in a change of name (Terrorism Information Awareness) and in restriction of privacy invasion. Despite these improvements the program was ultimately canceled in 2003.

To improve airport and flight security after 9/11 a newer version of the *Computer Assisted Passenger Pre-Screening System* (CAPPS II) was introduced. When reserving a ticket, the passenger is asked for certain information. To authenticate the passenger's identity, the data is compared with information from numerous public and commercial databases. Because of the concerns TIA raised, CAPPS did not store large quantities of data, but was also canceled in 2004. The replacement *Secure Flight* only accessed local databases, but failed again because of public privacy concerns.

7 information taken from Europa info site, which provides fact sheets on EU legislation `http://europa.eu/scadplus/leg/en/lvb/l33183.htm`

After public solutions for data mining failed because of privacy concerns a private-sector solution was suggested, called *Multistate Anti-Terrorism Information Exchange* (MATRIX). The system analyzes existing data sources and integrates data of web-storages. Because of doubts according to accuracy of stored data, the program was shut down in March 2005, while in Europe *SIS II* is currently in testing phase.

*Surveillance of Transactions:* The US Section 215 of the Patriot Act permits the FBI to *gain access to any relevant tangible item held by business.* A gag order is related to the law, which orders people receiving the request to deny the existence of the request. Because of concerns Section 215 raised, so-called *National Security Letters* are used by the Department of Justice. These letters first became law in 1978, whereas powers related to the letters were extended since then. Nowadays NSL's have similar powers compared to Section 215, but even more limited oversight. Investigators are permitted to request records without juridical approval (by a judge). Reports claim that 30.000 NSL's are issued per year (cf. [PUS05]).

| **Category** | **US** | **EU** |
|---|---|---|
| **Anti Terror Package** | PATRIOT Act | Hague Program |
| **Registration a. Monitoring** | US-VISIT | VIS<br>SIS |
| **Data Mining/Profiling** | TIA (canceled)<br>CAPPS II (canceled)<br>Secure Flight (canceled) | SIS II (testing phase) |

**Table 2.2:** Comparison of Changes US/EU

Instead of getting better equipment for law enforcement and more thorough security checks (except for tighter checkups at airports), politicians took the easy and fast alternative of modifying and creating several laws, which usefulness remains questionable (cf. [MH05]).

Shortly after the new laws had been legislated (in the US) the effects could already be seen (cf. [FP04]). Numerous people of foreign origin were taken into custody because they were suspected to be potential terrorists. A great deal of these people stayed imprisoned and were interrogated for a considerably long time, despite none these accusations could be prooved. Dragnet investigations were conducted, justifying arrests due to minor crimes (mostly people of foreign origin were affected). Telecommunications were wiretapped and numerous people were questioned to gain as much information about them as possible. Bank accounts owned by suspects of terrorist activities were closed.

A major part of the amendments of the Patriot Act are limiting the privacy of people by disposing the need of a probable cause. In many cases law enforcement and other agencies do not need a warrant anymore to start an investigation or for the usage of certain tools (e.g. wiretapping of telecommunications or networks). However, the application of these new laws in terms of restrictions is not well defined yet. Furthermore, additional information about such investigations has not been made public.

Access to stored data about people can nowadays be shared among different organizations and traded to marketing agencies too. Address information that is needed by online shops could show up in databases of marketing companies or call centers (see *spyware* for more

information). People have to specifically opt-out[8] from the privacy clause.

Data about marital status, family, friends, prior occupations, criminal records and even vacation destinations of the past are now collected and stored. During so-called *security checks* this information is used to detect potential terrorists, who want to advance to positions in society or the government. The information could be used by law enforcement during dragnet investigations (most of the modifications are not limited to terrorist activities).

These changes enhanced the competencies of governments and law enforcement, but also affected people greatly. Their suspicion towards other people is more easily aroused (maybe fueled by media and governments). The assassins of 9/11 appeared to be low-key citizens and did not attract attention until the day the attacks took place. It seems like the next attack could happen anywhere and anytime. The target of the assaults seven years ago were not just the US (and their progressive politics) but every western industrial country. Every citizen of these countries is likely to be a potential target.

After the attacks people felt more insecure. The measures taken by governments were apparently aimed to re-establish the sense of security. Although the measures were officially taken to prevent future assaults, the usefulness is not obvious. Experts argue that the applied modifications would not have been able to prevent the attacks of 9/11 (cf. [MH05]).

### 2.3.4 Critiques in Literature

The incidents of 9/11 triggered many changes in politics as well as in society. Reactions to these changes were both positive and negative. Especially privacy advocates were concerned that these new measures to fight terrorism would considerably diminish freedom and privacy of people.

Laws were altered rather quickly, members of parliament often had not enough time to deal with drafts that were assigned to them. The Patriot Act was the fastest legitimated law in US history (cf. [FP04]), the negotiations took only five weeks. These hasty actions raised questions, whether these taken measures actually be beneficial (cf. [MH05]). The fast presentation of alleged culprits could also be considered as an excuse to give people some sense of security back. Politicians may also have used the chaotic situation to their advantage. With quick issuing changes they could easily pretend that measures had been taken, despite the impossibility of prevention of terrorist attacks like 9/11.

After the catastrophe the whole world had changed. People were shocked and distracted. People wanted something to be done and authorities obeyed almost with pleasure. During that time it was easier to change laws so that the government (and linked agencies) gained more power and at the same time took away more and more freedom from the citizens. The most troublesome changes concerning privacy were already discussed above.

In Germany, laws concerning ID-cards and passports were altered, but privacy advocates argued that these modifications are probably useless in terms of fighting terrorism. In fact it is more likely that biometric data and similar information is used for prosecution by law enforcement (e.g. for dragnet investigations) or monitoring purposes by governments. There is always the possibility of abuse, despite the statement that fingerprint data

8 passive privacy approach. people are informed that data will be used in different ways and have to explicitly say so, if they do not accept further actions

stored together with (Austrian) passports is only used for identification of people and ID-validation purposes. Reportedly no law enforcement organization has access to the databases, where the information is stored.

Even if biometric data stored with IDs is not used for prosecution the simple act of storing such informations violates the right of self-determination (including sensible data), which is an important part of German law [9] (cf. [MH05]).

In addition, collected data as well as the act of collecting the data could be harmful for people's lives. After 9/11 the competencies of many agencies and law enforcement were enhanced. During an investigation (or in order to prevent a crime) collection of information from bank institutes, telecommunication companies, ... is allowed. Nowadays companies or bank institutes seem more suspicious when federal agents (or the BKA) ask for information about an employee or costumer than before 9/11. It is alarmingly easy (especially in the US) to get denounced as an alleged terrorist because of religion, destination of last vacation or even hearsay. People should not be surprised if suddenly a certain airline does not issue a ticket of fear to transport a potential terrorist or the bank institute you had already been costumer for years doesn't grant a requested loan or even cancel your account.

Human rights, one of the most fundamental rights, are often violated. Above all in the US a simple voicing of suspicion is nowadays sufficient to justify custody or even arrest. Checking or validation of arrests is not required, but should be applied, in order to keep human rights intact (cf. [FP04]). Wiretapping and the fact that data from numerous people can be collected at once could easily be seen as a violation of the right of self-portrayal. Furthermore the collected information is shared with other agencies and organizations. Hence the creation of profiles about "suspects" became much easier (*transparent society*).

Information about investigation, collection of data and even arrests is held back by the US-government and forbids people involved to make connected incidents public. Authors of [FP04] claim that restrictions like this would not be conform with the policy of constitutional legality of German law. Even the term terrorist is not well defined. According to the new laws every action that could probably harm governments politics (e.g. unannounced demonstrations) is seen as terroristic and people connected to that act can be treated as terrorists. This includes collecting and storing of DNA samples and the arrest of these suspects.

Companies and even individual people are granted immunity for disclosing information about customers and other people (e.g. neighbors) - the US-government sees this as a preventive measure. In [FP04] that preventive measure is even compared with *Minority Report* - a movie in which future crimes were prevented by arresting the delinquents before they could even commit the criminal act.

The situation in Germany is troublesome too. New German anti-terrorism laws even tend to criminalize whole ethnic groups. Data of people who request asylum is sent to intelligence agencies, which are allowed to use that information almost without limitations. Hence it seems that according to these laws all non-citizens are treated like potential

9 "Grundrecht auf informationelle Selbstbestimmung": people have the right to know what is known about them and when and where this data is stored

terrorists, despite the fact that terrorist assaults could also be committed by (German) citizens. The modifications and violations are frighteningly similar to principles and practices used in so-called *police states*. Till Müller-Heidelberg in [MH05] is going to such lengths as calling the practices used nowadays as *sedition* [10].

In the US people who have Muslim beliefs and want to enter the country can be taken into custody. Questions asked range from the reason for their stay in the country to their knowledge about weapons or connection to Armed Forces (military). People are suspected because of their beliefs - a clear violation of religious liberty.

In terms of the measures which were taken after the attacks, the terrorists seem to have already reached their objective (cf. [MH05]). People living in western countries are living in constant fear of assaults committed by terrorists. Freedom and human rights of people have become limited; therefore, measures have been welcomed in order to get a (delusive) sense of security back.

## 2.4 Summary

This chapter gave an overview of the history of privacy and the different definitions of the term and how it changed over time (because of technology and certain incidents). Research on that topic showed that (protection of) privacy is indeed a permanent problem. It had been an issue long before the term of privacy was firstly used in the 19th century. Several times it was attempted to define the term of privacy, considering mainly the mode of protection.

The definitions as well as the principles of privacy are difficult to determine. Privacy can be viewed under different aspects: On the one hand, serious violation of privacy can be done by using confidential data, but on the other hand sensitive information can as well represent desperately needed information (or for that matter, information that is given away willingly by someone in order to use a service).

Already in ancient Rome and Greek governments and political rulers restricted people's privacy in order to have increased control. The difficulty of hiding actions or personal views makes it easier for authorities to prevent incidents that may harm society (or the ruling classes).

The US constitution adopted in late 18th century already included (rather unclear) amendments[11] regarding people's privacy. Privacy was paractically non existent in parts of Europe during WW2 (NS-Regime) and during the Cold War in eastern Germany (Totalitarian Regime in the *DDR*).

During the onset of the computer era (in the 60's) the technical privacy concern was just focused on access limitation to data stored, but the public seemed already aware of the potential (for privacy violation) upcoming technology could hold (cf. [Hof69]). Because of technological advancements made over the years, it was easier to access and

10 "Volksverhetzung", as defined in §130 StGB: assaulting of human dignity by insulting or decrying certain parts of society

11 Bill of Rights, 4th Amendment: "The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon probable cause ..."

store information (about people). During the 90's many countries found it necessary to adjust laws in order to protect people's privacy better (data protection).

The incidents of 9/11 seemed to turn everything around. New laws and amendments adopted shortly after the attacks tend rather to reverse the "advancements" made in protecting people's privacy in favor of security and control. Technologies were introduced that made it possible to store and access databases, which hold detailed information about people (data mining). These measures were usually justified with maintenance of *national security*, by identifying terrorists before they are able to inflict harm on society. Especially in the US several attempts of implementing database networks (TIA, SecureFlight, ...) were officially canceled because of public uproar, while the European pendant SIS II is currently in testing phase.

These invoked measures, which cut on people's privacy and freedom triggered criticism all around the world. The author of [MH05] even believes that the measures taken by governments, which have been justified through national security for the prevention of future assaults, play into the hands of the terrorists.

# CHAPTER 3

## Risks to Privacy

After outlining the history of "privacy" and introducing different approaches to define the term during past decades (and centuries), this chapter gives an overview of risks to privacy, while mainly focusing on the risks people encounter while taking part in online transactions.

When surfing the internet user establish numerous connections to different (web-)servers. Each of this servers has the ability to record certain data about users (and connections). In addition to logging the content, which is accessed on the server, the IP-address can be recorded too.

This IP-address is not just a combination of numbers, but an opportunity to determine

- domainname
- workplace
- location
- operating system and browser

which this particular user is using.

While the IP-address is likely to change between different connection sessions, privacy is still threatened, because it is possible to link different sessions to one user. With so called *cookies* (see below), which are saved on a users harddrive it is not only possible to resume sessions, save preferences, but also to identify a user, who already had been connected to that server in the past.

*Session-linking* can also be used for monitoring purposes. Businesses as well as consumers can benefit from that fact. Web servers are not only able to personalize layout of the site, but also the content, which can be accessed by a certain user. Businesses on the other hand could use monitoring tools like cookies to log activities of their employees (cf. [Rei99]).

In this following chapter not only risks and tools like IP-addresses and cookies are discussed, but also threats like spyware and an organization most of us depend on to get a connection to the internet in the first place: the ISPs (Internet Service Provider).

## 3.1 Tools

In the chapter below the most common threats to users, when users surfing the net or simply working on their computer, are discussed.

### 3.1.1 Cookies

*Cookies* are small files that are placed on the user's harddrive. Data that is included in these files is usually used by websites and web applications to 'remember' users. Every time some user is visiting the webserver it searches the users harddrive for the cookie. If the file is found, the information included is parsed - if the search is not successful, a cookie is created and placed on the user's harddrive.

When the webbrowser *Netscape* used cookies for the first time back in 1994 (cf [Fox00]) the main purpose for these small bits of data was to make surfing the internet easier for people. Websites were able to save login information, hence user didn't waste time with reentering that information. Other uses for cookies back then were

- implementing a shopping cart
- saving user preferences
- saving preferred layout
- saving information for use in future

Website owners already used cookies back then to determine which kind of websites/content attract people the most and which parts of the website/application need to be changed or altered to fit more users interests. In the past however Netscape didn't inform their users about the existence of cookies (cf. [Fox00]) and the capabilities of that particular feature. Not until January 1996, when the media reported about cookies, Netscape promised to add a tool to their browser, that should allow people to disable the cookie feature. Cookies however were still enabled by default (standard 'opt-out' practice) and finding that option within rather complicated menus, wasn't that easy for most users.

Later cookies were also used to customize/personalize results of search engines (e.g. Yahoo, Google, ...), content of pages to better fit peoples' interests (cf. [COO08]). Information hidden inside can be accessed by the website that had placed it on the harddrive in the first place as well as other websites. Cookies weren't intended to be harmful to peoples privacy, but it took businesses and companies (e.g. advertising companies like *'double click'*) not that long to discover opportunities cookies offer.

#### Structure of Cookies

Cookies consist of several parts/variables, which are filled with information:

*Name:* Cookie name, given by the server

*Version:* Cookie Management Specification

*Expires:* Date, when cookie is automatically deleted

*Max-Age:* Period of time the cookie is saved (cookie is deleted afterward)

*Domain:* Name of domain the cookie is valid for

*Path:* Cookie is only valid for this path

*Port:* Cookie is limited to one or more ports

*Comment:* describes cookie more detailed

*CommentURL:* URL to comment

*Secure:* cookie is sent trough secure connection

*Discard:* Cookie will be deleted when browser is closed

#### Threats to Privacy

Many websites save cookies on your harddrive - hence a 'track' is left of the sites you visit. Companies already discovered the usefulness of these small files. *Doubleclick*[1] for example displays advertising banners on numerous websites. When a user is visiting on of those websites, the server searches the harddrive for an 'doubleclick-cookie'. When search is unsuccessful a cookie is created and sent back to the user. In case of succesfull search the webserver sends the in the cookie included 'user-ID' to doubleclick and requests information about that particular user. The request itself adds new information to the 'profile' doubleclick owns on that user. With a huge network of websites that are connected with doubleclick, it is possible for that company to easy derive profiles of interests on user, which help to customize advertisement to fit needs better. All these actions are taken without the user's knowledge (cf. [COO08]).

At first glance the creation of these profiles seem to be rather useful. Only advertisements and banners are displayed that are of peoples interests. But some people could find is invasive to their privacy, when data concerning their interests and preferences is stored somewhere without their knowledge. Other users are not even aware of cookies and what these files enable webservers to do.

With newer versions of different webbrowsers it is not only possible to disable all cookies, but disable cookies sent by certain websites or cookies from other websites (*third party cookies*, e.g. doubleclick). Hence cookies don't necessarily pose a threat to users privacy - if security settings of browsers are set accordingly, of course.

### 3.1.2 Spyware

Spyware represents software that installs itself on a users computer - usually without the user's knowledge and permission. Tools like that conduct an indirect infiltration and are considered as *the largest threat to internet users since spam* and poses a *serious harm to consumers* (cf. [Sip05]). Such software could be the cause for slower internet connections, poor computer processing times and privacy concerns (cf. [Fre05]). Spyware is often used

1 www.doubleclick.com - Internet Advertising Solutions

by businesses and companies, who want to know something about consumers or their computing habits and interests (marketing segmentation and audience targeting). On the other hand spyware can be utilized by hackers, who want to gain access and control of the victims system (to steal information or create networks of compromised computers).

Once installed the code collects information (computer uses and online behavior) sends it back to the creator of the spyware. Actions taken by such software not only include the gathering of data that is considered private, but access, altering or even deletion of sensible files as well. Usually processes which are utilized by spyware are running in the background and sometimes can't even be seen in the taskmanager of the operating system (or they pretend to be a legitimate process). With spyware users interests and online behavior could be tracked and collected data not only be sent back to the creator of the spyware, but also be sold to marketing or ad companies (and used for direct targeted advertising [2]). Different kinds of *update agents*[3] can be considered as spyware. Because of poorly written code such applications can have appreciable effects on the systems performance - which is one of the common symptoms of spyware.

Other symptoms are

- slower internet connection
- advertisements are popping up, that have no obvious connection with the visited website
- different starting page in web browser
- bookmarks are added
- system is connecting to the internet (dial-up connection only)
- firewall alerts are popping up regularly, because some kind of application tries to make a connection
- files are deleted without user authorization
- new wallpaper in the background, which advertises download of some anti-spyware application

The Center of Democracy and Technology defines spyware as follows:

> software which hijacks web traffic, tracks internet users without their knowledge and consent, sends information back to a third party and is not easily removable

In numerous articles spyware is described as:

> software that is surreptitiously installed on a users computer and monitors a users activity and reports back to a third party on that behavior

2 advertising, web banners are displayed according to users interests and preferences
3 application that notifies users, when updates of installed software is available

The *Federal Trade Commission*[4] defines spyware as:

> software that aids in gathering information about a person or organization without their knowledge, and that may send that information to another entity without user consent

As of 2003 85% of personal computers had spyware installed (usually without users knowledge) and an average of 28 spyware items per machine. Some of these computers came directly forom the manufacturer with spyware already installed on the system (cf. [TFS04]). On the other hand spyware can also be used by parents, who want to keep track of thair childrens use of internt capabilities or by managers, who want to keep an eye on their emploees. Most of the time however spyware is used *for purposes that accrue to the commercial, financial, or personal interest of some third party* and represent a *lack of transparency and absence of respect for users' ability to control their own computer and internet connections.* Having said that the Business Software Aliance (BSA) sees spyware as *a tool for businesses to provide more effective service to customers.*(cf. [TFS04])

### Types of Spyware

The most common types of spyware are listed below (cf. [Sip05] and [TFS04]):

*Adware Cookies:* Files are stored on the user's hard drive when visiting a website. These files contain different kinds of information. Originally *cookies* (see section above) were created to make it easier for people during login or saving preferences. Cookies can also be used to customize information provided to the user's needs or interests. On the other hand adware cookies are used to create profiles of user's online behavior (and sell it to marketing companies).

*Adware:* tracks online behavior (visited sites, interests, ...) of users, usually without their permission or knowledge. With gathered information detailed marketing profiles are created. These profiles could be used to provide users with customized offerings, adverts, web banners (according to the users interests) or even triggering of webpage redirects. But adware could also change the behavior of web browsers (with the installation of browser helper objects) or default settings like starting page, bookmarks, default search engines or displayed toolbars. In some cases adware could make changes to the systems registry to prevent users from simply change back to their previous (preferred) starting page. Changes in registry is one of many reasons why getting rid of spyware can be very time consuming for users.

*Keystroke Loggers:* Software (also known as System Monitors), which is usually used to capture user interaction with the system and gather information like usernames, passwords (e.g. to log onto someone's online bank account) or creditcard numbers (to commit *identity theft*[5]). Sensitive information like that can also be used (by hackers) to infiltrate target networks and sometimes these loggers are installed as part of an Trojan Horse attack (see below).

4 US government agency, which promotes consumer-protection and prevents tactics that could probably be anti-competitive

5 pretending to be someone else in order to gain money (e.g. by using the victims creditcard)

*Trojan Horse:* Remote Administration Trojans (RATs) are a malicious form of spyware. They are often part of free software (e.g. games, peer2peer software, ...) or attached to an email. These Trojans usually take control of an user's computer after installation (and trigger installation of tools that monitor and control the victims system) and access, alter, damage or even delete files, while maintaining a connection back to the Trojans creator. By making use of address books (e.g. MS Outlook) on the hijacked computer, the Trojan sends itself to other users (per email). As a side effect of this actions the collected information could be used for distributing spam. Examples for Trojan Horses are dialers, code that makes use of dial-up modems to connect to expensive telephone numbers, drive-by downloads (code is downloaded by simply visiting a particular website) or threatening applications that exploit vulnerabilities in operating systems like Microsoft Windows to cerate a backdoor[6]. Despite network administrators often make use of these tools there is no other legitimate use for such software. Most of the time it is utilized by hackers to get access and control over a victims system.

*Web Bugs:* Graphics that are not bigger than 1 pixel. They are often included in email in HTML format (to determine whether the message was read or not). Usually a cookie is placed on the user's hard drive, which can be retrieved by spyware later. It's alarmingly easy to become a victim of this kind of spyware. Actions are triggered simply by visiting a website or reading a HTML email. Hence even the most careful users could be victimized by web bugs.

*Scumware:* Changes particular content of websites to link back to third party sites (e.g. every mention of a particular company is enhanced with a direct link to the company's homepage).

#### Issues with Spyware

With spyware (alongside with it's distribution and capabilities) there are numerous ethical and legal issues connected. Some of them are listed below (cf. [Sip05]):

*Privacy Invasion:* Spyware is capable of intruding into a users computer and collecting data without the users knowledge. Information that could probably be considered private could be dismissed and sold or used for direct marketing (see below). Usually the user doesn't grant permission for installation of spyware. Hence such software can be easily seen as a violation of privacy.

*Surreptitious Data Collection:* Spyware is able to collect private information and send it back to the software creator, marketing companies or used for other fraudulent purposes (e.g. Backdoor Santa [7]). Data stolen could include private data as well as valuable or secret information that could probably be useful for business competitors.

6 bypassing of authentication (by exploiting of vulnerabilities) to get remote access to a computer and take control (at a later time)

7 slang term for spyware which has no othher obvious function than gathering information about users online behavior

*Direct Marketing:* Software companies are paid by creators to include spyware with their legal products to gather information. While gathered information could be used for improving offerings for users, it could also lead to promoting of pornography gambling and other fraudulent schemes. Although the presence of spyware is often disclosed within the Licence Agreement, such agreements are seldom read by users installing software.

*Hijacking:* Trojan Horses usually not only gather information, but make as well use of the hijacked computers resources. Installed spyware could alter preferences like starting page of the browser, bookmarks and toolbars. With Torjan Horses are often tools installed that are capable of monitoring the victims computer. Hence the attacker is able to collect private information (by utilizing keystroke loggers, screen capture tools, ...), which can lead to threats like identity theft.

*Trespass:* Spyware is also often a part of a software that was obtained through file sharing networks or in unwillingly triggered downloads. Sometimes the malicious code disguises itself by appearing to be a legitimate browser plug-in or as a important part of an "auto-update".

### Survey on Spyware

In [Fre05] a survey was conducted. People had to decide whether they agree or disagree on 11 statements. Asked questions and results of the survey are shown below.

1. The privacy concerns of spyware outweigh the potential benefits.
2. The focused marketing efforts of spyware outweigh the potential privacy concerns.
3. Spyware should be controlled, monitored, and regulated by industry.
4. Spyware should be controlled, monitored, and regulated by government.
5. The number of CPU cycles (computer processing speed) lost to spyware applications is acceptable.
6. There should be a maximum amount of CPU cycles that can be acceptably used by spyware.
7. When I install software, I read the license agreement.
8. When I install new software that I believe contains spyware, I read the clickwrap agreement.
9. In general, I feel that spyware is more beneficial than harmful.
10. In general, I feel that spyware is an issue of privacy more than one of CPU cycles.
11. Before completing this survey, I was aware of and knew about the issues associated with spyware.

**Figure 3.1:** Questions asked [Fre05]

This survey was conducted with 70 students (average age of 23.5 years) residing all over the world and results showed that the greater part of participants found spyware to be harmful (not just to privacy). The marketing benefits by spyware also don't out wight privacy concerns that are raised. Issues with privacy are even more important than the performance loss caused by spyware. Nevertheless performance loss (or CPU cycles) remains an important issue. 63% of questioned people felt that the (sometimes enormous) number of lost CPU cycles is unacceptable and that there should be a limit on the amount of CPU time the spyware is able to use. Participiants aslo felt that limitations should be applied by the industry or the government to regulate the use of spyware.

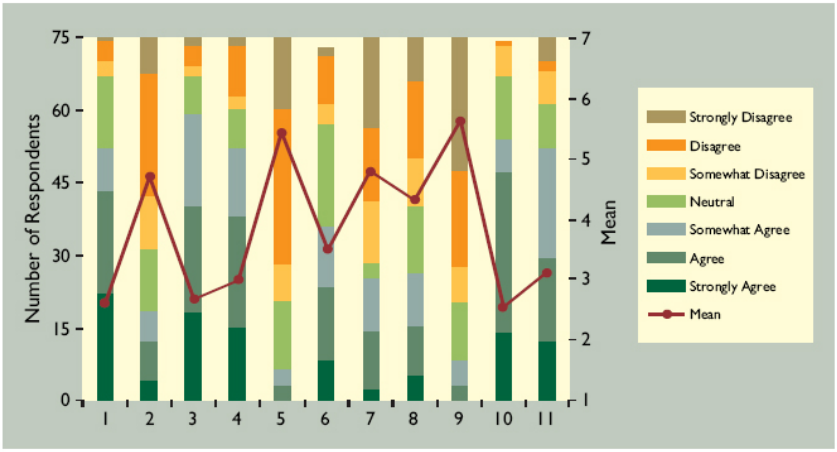


**Figure 3.2:** Results of survey [Fre05]

### Examples of Spyware

Some Examples of spyware are listed below (cf. [TFS04]):

*Microsoft:* The *Windows Update* feature is designed to update Microsoft software with new features (and bugfixes) as soon as they are available. The update function establishes a connection with a Microsoft server and looks for updates for already installed software (or the operating system itself). Depending on the users preferences these searches can be conducted manually or automatically (at a given time). Furthermore users are able to decide, whether these updates are downloaded/installed automatically or require the user's permission, before any of these actions are triggered. Despite a connection with a third party server is established, the Windows update feature isn't acting like a typical malicious spyware program. The user is clearly notified and seeks permission (if so chosen in preferences), when a update is ready for download or installation - hence users are aware that an connection to Microsoft's servers is established and therefore maintains some degree of transparency. Many other spyware applications don't use notifications at all, when communicating with third party servers and sending (often valuable) data. *Windows Media Player* on the other hand is a whole different story. The application keeps track of media (music, videos, ...) that is played and sends data back to Microsoft. By default that feature is enabled and serves also for validation purposes of files in wm-formats[8].

*Google:* is using the terms entered into the search engine to calculate trends - what are people most interested in. This data is disclosed to companies and businesses, who can make use of it (e.g. for more customized advertising to gain even more profit). With *Google ToolBar* and other applications like *Google Desktop* (including Google's newest invention - it's very own web browser named *Chrome*) searching got more effective for people - at least at first glance. Google keeps track of terms a particular

8 Windows Media Player format, video (wmv) as well as audio (wma)

user is searching for and customizes the results that are shown by the search engine. Hence Google is capable of influencing which sites are found and visited by people. Google Desktop on the other hand is not just an enhancement to the operating system by making it possible to search the internet without the need to start a web browser first, but the application is also capable of searching the harddrive of your local computer. That particular feature could however easily be misused, because people could never be sure, whether information about the contents of one's harddrive is sent back to Google.

*RealNetworks and other companies:* Companies like RealNetworks offer limited versions of their products for free download. Before/during the installation process (or even before people are able to download the product) they have to give away information (like name, email address, place of residence). When filling the form people are offered an option of opting out, e.g. they have to tick a box if they don't want the information entered to be used by other companies.

*KaZaa (p2p applications):* Years ago KaZaa was one of the most famous p2p[9] applications around. KaZaa was (and other applications like Limewire and Bearshare still are) widespread and therefore a perfect opportunity to include spyware with the software bundle. In case of KaZaa the spyware-application is called *Gator* and users installing the software are only notified by its existence in the package by a brief mention in the EULA[10].

*Gator:* Gator is also known as GAIN (Gator Advertising and Information Network) (cf. [PCF]) and can be installed with software bundles as well as by drive-by downloads[11]. The tool includes a feature that remembers login information on different websites. At first glance it seems useful, because people have not to reenter the information every time they visit a website. On the other hand storing that information could be quite dangerous, especially in case of creditcard numbers. Gator not just keeps track of visited websites but also displays advertising banners - ads of a particular company when visiting the competitors website.

*Kodak:* Legitimate Software that is bundled with e.g. hardware could also include applications, which behaviors and actions are somewhat similar to spyware. In this particular case the Kodak software bundle included an application named *BackWeb Lite*. That software is obviously used to notify user when updates for the installed software are available. Because of poor written code however a huge part of system resources is used and taken away from other running processes - hence overall system performance is suffering. Although that particular part of the software was never intended to be harmful to the system, rather helpful to people and customers, the effects it has on the system rather resemble the impacts (e.g. performance loss) which spyware has on computer systems.

9 peer to peer connection: connection between clients, without the need of a server
10 End User License Agreement
11 download is triggered by visiting a website, clicking an web banner, or opening an email attachment

### 3.1.3 Proxies

When connected to the internet, data packets are sent to and from the target server. These packets usually include the IP-address of the sender (as well as the receivers address). That address is a risk to privacy, because different actions can be linked to one particular user. When the worst comes to worst the user could even be identified, if the user's ISP is asked to disclose ID information.

To protect privacy data could be sent through a so-called *proxy* - hence targeted web servers are only able to see the proxy's IP-address. Proxies however can not only be a tool to protect privacy but a threat as well. The situation is similar with network administration tools, which could be used by network administrators for legitimate purposes as well as by hackers for Trojan Horse attacks (see section above). All data packets that are sent to the proxy can easily be stored there. That data includes of course login information, creditcard numbers and other private and sensitive data.

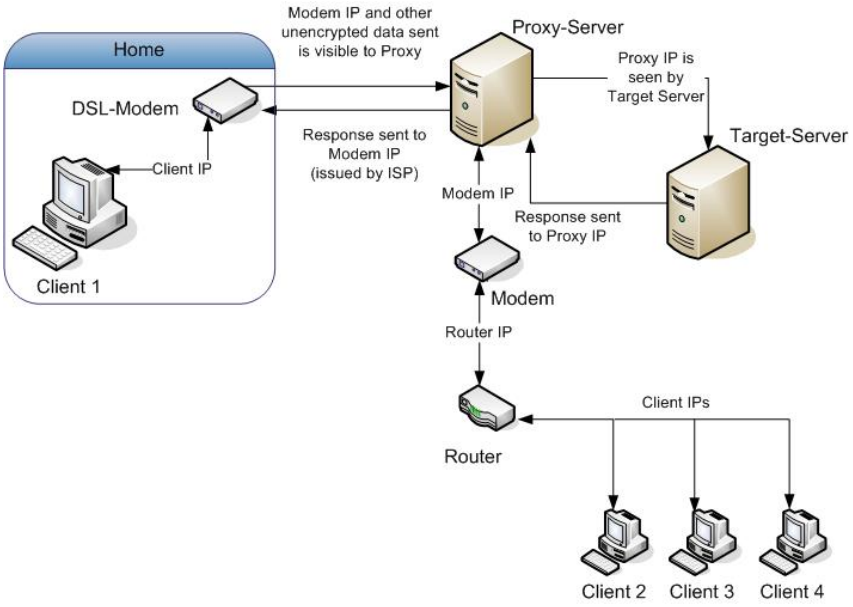


**Figure 3.3:** Proxy Architecture

## 3.2 Organizations

Below are some organizations (business companies, ...) listed, which could probably pose a threat to peoples privacy, when surfing online.

### 3.2.1 Internet Service Provider (ISP)

An ISP company (e.g. AmericaOnline - AOL) provides services and contents that are needed to use the internet or to make content available online (by providing webspace). Every single data package is transmitted through some kind of gateway[12] which enables customers of such companies to access the WWW. All data that is sent and received by subscribers to that ISP is put trough servers owned by that company. When a contract is made with an ISP (private) data of customers is stored (in order to enable easy communication, payment, ...).

To get access to the internet an IP-address is issued to every customers. This address is usually included in every request a subscriber sends and is visual to every server that receives a data packages. Like already mentioned in sections above it is not only possible to link several actions to one particular user, but also to gain (more or less detailed) information about people, like the domain name from which the request was sent. When the ISP is determined, which issued that particular IP to the user, who sent the request. To get to know private information (e.g. name, address, ...) about that particular user on barrier is still in place: the ISP, where the IP-address can be connected to one subscriber.

As suggested above ISPs are far easier to identify than their subscribers. In addition to internet access ISPs often provide a certain amount of webspace (storage) on their servers for customers to use.

Concerns were raised by ISPs whether they are liable for (copyright protected) data (cf. [Mid05]) that is

- stored on servers owned by their company
- channeled through their servers (gateways)

With the DMCA[13] ISPs have to comply to *notice and take down* procedures. Such notice has to include

- name, address, telephone number, email address ...
- information to identify media, which infringes copyright
- statement of 'good faith' that distribution wasn't authorized by copyright owner
- signature of person giving that notice

If an ISP is notified of media (stored on it's servers or distributed by on of it's subscribers) the company has to

12 single point of hardware (e.g. router, server, ...) where data is channeled through
13 U.S. Digital Millennium Copyright Act

- take down or block access to media that infringes copyright
- block or delete account of copyright infringer (subscriber) or similar effective measure

#### Disclosure of private Information

Usually ISPs are not allowed to disclose private information about their customers for other purposes than the data was collected in the first place. According to the FTA[14] ISPs should not be required to give sensible data away unless a according court order is issued (cf. [Mid05]). Copyright owners who want to obtain information about some alleged copyright infringers, can do so by *requesting a clerk of a United States district court to issue a subpoena to an ISP*. In order to get such an subpoena copyright owners have to provide

- a copy of the take down notice under US Copyright Act
- sworn declaration stating that the purpose for seeking the subpoena is to obtain the identity of an alleged copyright infringer
- obtained information will only be used for the purposes of protecting that persons copyright

If these requirements are met, an ISP has to disclose information, which could eventually identify the alleged copyright infringer. After however a take-down notice has been issued by a copyright owner, obtaining a subpoena is just *an formality* (cf. [Mid05]). ISPs and privacy agents got more concerned about

- privacy of subscribers
- the burden of compliance on service providers through being inundated with subpoenas[15]

According to privacy obligations in US and Australia ISPs are not allowed to disclose personal information[16] unless it is required and authorized by law. To get these authorizations got far easier since 9/11 and the launch of the *Patriot Act*.

### 3.2.2 Google

After the Google company was launched on September 7, 1998 by Larry Page and Sergey Brin their search engine went online just few months later. The service should pose a threat to *Yahoo*[17], the market leader in that branch back then. Shortly after the launch

14 Free Trade Agreement between USA and Australia

15 as of 2003 the RIAA (Recording Industry Association of America) had served almost 1000 subpoenas [Mid05]

16 any information about an individual whose identity is apparent, or can be reasonably ascertained, from that information (cf. Australian Privacy Act Section 6)

17 `www.yahoo.com`

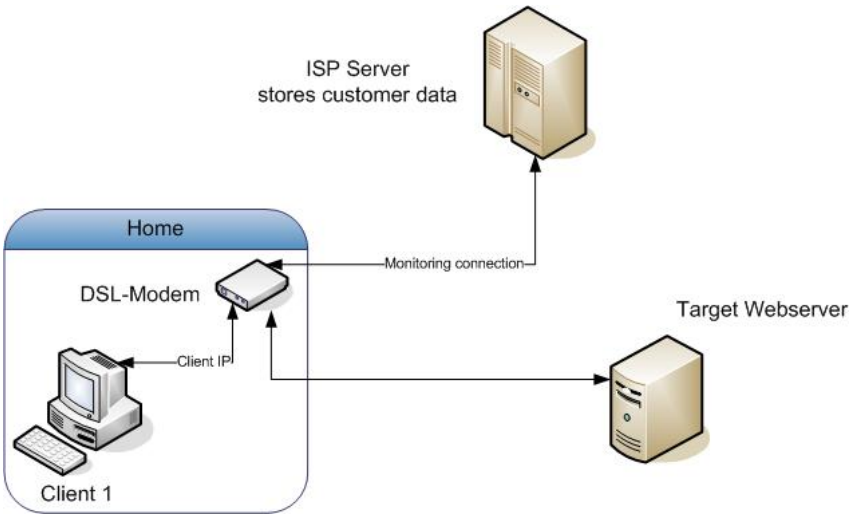


**Figure 3.4:** ISP Architecture

the service became a success and many other services (usually financed by advertisement) followed.

Rapidly Google became more than just the owner of a successful search engine. Apart from a mailing service (see GMail), Google News provides some kind of daily news portal with a feature to send news directly to users per mail. People can built their own personal starting page (`www.google.com/ig`) and keep track of their appointments (see Google Calendar below). Even an specially designed version of the search engine exists for PDAs and mobile phones (`www.google.com/mobile`). (cf. [WEL08])

The Google company, more so it's search engine, became famous and w4ell konwn by so many people that even the verb "to google someone/something", meaning looking uo information of something/someone on the web (not just with Google search engine), was added to (online-) dictionaries (e.g. LEO Dictionary[18], German DUDEN).

All this useful services provided by one single company could as well pose a threat to peoples privacy. Not just search terms are sent to Google servers, but people put content and data on servers owned by Google willingly (see Google Docs, Google Code, ...), so works in progress, developments are accessible almost from anywhere and can easily be shared with co-workers and friends. Data put on these servers however is not that safe as it may seem. The privacy policies and terms of use connected to these services should be read carefully, before Google is given access to sensitive or valuable data.

Some of the most important services provided by Google and privacy issues that may arise according to them are discussed below (cf. [DRW]).

### Search Engine

As of today Google seems to be the most important and the most famous online search engine. The engine is translated into numerous languages (including German, Italian, ...).

18 `dict.leo.org` - comprehensive online dictionary/translator for numerous languages

In addition to the text field, where people type the search terms into they are looking for, option buttons give people the opportunity to filter search results easy and fast (e.g. only sites are shown that were written in an certain language). The extended search offers even more options and commands to choose from, in order to improve search results.

The search term put into the form is of course sent to Google, where the request is processed, stored and the results are sent back to the client. Results that are sent back to the user are filtered and often even censored. Google is removing sites from results that break laws of certain countries (e.g. sites that break copyright laws according to DMCA[19]) or their own in-house rules. Filtering on the other hand is used to improve search results, by storing terms that a certain user was looking for in the past. Hence Google is also able to derive profiles of peoples interests - an ability that is especially useful for advertisement that is tailored for every single user. Privacy agents got even more troubled when Google took over *Doubleclick*[20] in 2007.

### Chrome (Webbrowser)

Google's newest development is called *Chrome* and should battle Microsoft's market leading position with Internet Explorer. Since September 2008 Google's very own web browser can be downloaded for free (`www.google.com/chrome`). Chrome is still in development and additions are made nearly everyday, but this new browser has already earned praise and reprimand from users and critics (cf. [SUR08]).

Chrome should be way faster than it's competitor developed by Microsoft and easier to use by customers, but tests proofed otherwise. Google's browser is not faster than other web browsers, especially running with Microsoft Vista. But even istalled on the faster XP Chrome is still far slower than FireFox (`www.mozilla.com/firefox/`) or Opera (`www.opera.com`) (cf. [SUR08]). The user interface is more minimalistic in order to increase usability, but is in fact missing many functions, which are standard in other browsers (e.g. administrative functions for bookmarks).

Most important features of Chrome are listed below:

- With *anonymous mode* no traces of online behavior is left on one's computer (similar to Apple's Safari Browser `www.apple.com/safari`)
- when terms are entered into the URL-textfield already visited sites are searched for this terms (like features already implemented in FireFox and Opera).
- the nine most visited pages are included into the starting page

**Security and Privacy** Every new Version of Chrome gets an distinct version number - hence the used version of Chrome is easy to identify.

While surfing Chrome is connecting regularly to Google servers to transmit data, while the actual nature of the sent information can't be determined. Likely the connections are established to enable features like the build-in phishing- and maleware-filter (similar

19 US Digital Millennium Copyright Act

20 Online Advertisement Company - displays advertising web banners on third party websites according to users interests

to Microsoft's Internet Explorer). Hence content of data isn't transparent (private and sensible) data could be transmitted, which could be useful to Google in one way or the other.

### GMail (E-Mail Service)

Google's e-mail service (`gmail.google.com`) was first announced on April 1 2004 (cf. [GOO04]). Because of the announced storage space of 1 Gigabyte for mailboxes (other free-mail providers like Yahoo, GMX and Hotmail provided just a few Megabytes per customer back then), the announcement was first wrongly regarded as an April Fool's joke.

The e-mail service was advertised with the following advantages (cf. [GOO04]):

*Storage:* Google provides every single account with 1 Gigabyte of storage, so customer are able to *hold their emails forever* (at that time more than every other free-mail provider)

*Search:* with built in search technology, people are able to quickly search their emails (sent or received) for specific keywords

*Speed:* emails aren't filed into folders, but parsed into "conversation - trees", hence it is easier (and not that time consuming) to follow sent emails and following replies

**Security and Privacy** Hence to the huge size of storage available for single customers, people are seduced to never delete their messages from their mailbox. However even if some user deletes a message from his/her account, it is not completely deleted (cf. [GMA]). In fact the mail is still stored on Google owned servers and the content can be still accessed internally by the company (for customized advertising purposes). Google argued that *no human-being* is parsing mails stored in mailboxes and that the information collected is used for advertisement purposes only (cf. [BEI04]).

Furthermore Google stated in its privacy policy that

> they will be pooling all the information they collect on you from all of their various services. Moreover, they may keep this information indefinitely, and give this information to whomever they wish. All that's required is for Google to "have a good faith belief that access, preservation or disclosure of such information is reasonably necessary to protect the rights, property or safety of Google, its users or the public." [GMA]

According to the Electronic Communication Privacy Act (ECPA) emails lose their status as protected communication after 180 days have passed. After that deadline has been reached an email messages is just an ordinary entry in a database and it is no warrant (but an subpoena) needed to make Google to provide a copy. Due to the fact that Google's databases are spread all over the world and therefore subject to the laws of respective countries, it is unclear how the privacy of email's content is kept intact (after some time has passed).

With the ability to parse email's content (for certain keywords) and store information about email authors as well, there is now a massive potential for abuse (cf. [GMA]). These practices sound alarmingly similar to measures CIA, NSA used for years already with their communication monitoring technology (see section below) for national security's sake.

### Google Desktop

The search feature this application (`desktop.google.com`) adds, doesn't just cover content available online (similar to the standard search engine provided by Google), but it is also able to search local computers for specific content. All files stored on a computer are indexed by Google Desktop for faster execution of searches. It is possible to search for numerous file types, including image, music and video files. With available plug-ins the list of supported file types can be even enhanced. Results of local searches include comprehensive information about the files found and a short abstract of the files content with search terms clearly marked. Because of Goggle Desktops indexing it is even possible to search for files that were deleted (by accident).

Apart from the search engine, Google Desktop includes the feature to add certain gadgets to the sidebar offered. Examples for useful little gadgets are analog/digital clocks, news-tickers, weather information ... new gadgets are available almost every day.

**Security and Privacy** Since Version 3 of Google Desktop it is possible to expand a search to include more than one computer. Privacy agents however advise against using this feature„ because index data is sent to Google servers when using that option.

### Google Docs

With Google Docs (`docs.google.com`) it is possible to store documents on remote servers and edit these documents in real time. Hence people are able to create and share documents with friends and co-workers easily. Features are similar to capabilities of Microsoft's Office.

After complains (not just made by privacy agents) Google altered it's terms of use and terms of privacy. Before the edit the copyright of content put onto the servers migrated to Google. Now the copyright remains at the user who created the document. Google however still reserves the right to alter, translate, share and distribute the content in order to enable full functionality of the provided service.

### Miscellaneous

Other services provided by Google Inc. are listed below:

*AdWords and AdSense:* AdWords and AdSense are services that display advertisement on third party websites.

*AdWords* displays ad-banners according to search terms entered into the search engine. These resulting ads are clearly marked as advertisements for better distinction from search results (with background coloring and the "Ads by Google" caption).

*AdSense* is a service that allows webmaster to display Google Ads on their own websites. When these advertisements are clicked by users, the owner of the site is rewarded with a small salary, which is derived by the actual amount of clicks.

*Google Earth and Google Maps:* With these two services (`earth.google.com` and `maps.google.com`) provided by Google it is possible for users to virtually travel to every location on the planet. Numerous pictures taken by satellites in space are combined to create the most thorough map that is accessible by the general public. Unlike

Google Earth, which is an self-contained application, Google Maps can be integrated into websites, to provide Google Earth's features.

Simply by entering GPS data, address information or just by clicking at a spot on the map, people can have a look at requested locations. But the puzzle built of numerous pictures isn't complete, because of numerous reasons.

- pictures of a particular location are not added yet
- pictures taken are not clear enough
- Google isn't allowed to show pictures of certain locations, because of security and similar reasons (cf. [WAS07])

*Knol:* Only a few months ago (23.7.2008 - Beta Version) Google launched a new service named *Knol* (`knol.google.com`). It is meant to represent a collection of articles about numerous topics (Online Dicitonary), similar to Wikipedia[21]. It can be assumed that Google is trying to compete with Wikipedia - as of today the market leader in Online Dictionaries.

*Google Calendar:* This service (`www.google.com/calendar/`) provided by Google works similar to other electronic calendars. Unlike for example the calendar offered by Microsoft Outlook online calendars can be reached from every workstation or notebook, because the information isn't stored locally, but on remote servers. User just need to log into the account to get access to their data. Microsoft and Yahoo provide similar products, but Google's service has some useful additional features. New appointments can be added simply by typing in title, time, etc. into the text field (e.g. meeting 7 pm). For people who using *Gmail* as their mail-account Google calendar can add appointments automatically, if certain terms are included in the mail's content (cf. [WEL08]).

## 3.3 Miscellaneous Methods

After describing tools and organizations, which could threaten people's privacy, this section will include miscellaneous methods.

### 3.3.1 Data Mining

More and more data is collected and stored in huge databases. Stored information (about people) is mostly used by companies and governments to provide services. However the information could as well be used to predict trends or even behavior, habits of people (costumers). These predictions are created by a technique called *data mining*, which is often used by businesses and companies to gain certain benefits and advantages while competing against other businesses.

Unfortunately the huge amount of data saved about individual people poses a threat to privacy, because information collected is more often shared and sold to other companies.

21 `www.wikipedia.org` - comprehensive Online Dictionary existing since 2001 and that is expanding fast

Furthermore it is becoming more difficult for people to keep track of who knows what about them. A ability, which is an important part of the German *Grundrecht auf informationelle Selbstbestimmung*.

Basically data mining is used to discover interesting and useful patterns within (huge) databases to predict future trends or behaviors. Collection of data as well as data analysis and creation of profiles could be supported by automatic methods. During the process data collected is processed and new assumptions (due to results) are generated. Rules, which generate these assumptions (new data) are often generated by experts, who are familiar with the topic.

In [Sei03] data mining is described as software-supported, automated prediction due to known patterns of behavior and the calculation of yet unknown relations, patterns and trends within huge databases. Data mining should discover and of course use new information and knowledge. Other objectives of DM among others are segmentation, classification, forecast, description of concept, detection of anomalies and analysis of dependencies.

While some phases (from collection of data to actual use of new derived data) of DM can indeed be automated, others still have to be conducted manually. Human beings are still an important part of data mining (cf. DAT03), especially when it comes to protecting peoples privacy (e.g. reviewing of DM rules).

#### Data Mining Models

There are two different DM models (cf. [Sei03]):

*Verification Model:* Questions and hypothesis created by experts are verified or discarded due to analyzed by numerous tools.

*Discovery Model:* automated creation of hypothesis. These are validated due to data collected.

#### Data Mining Phases

The process of data mining consists of 4 different phases (cf. [Sei03]):

*Planing Phase:* Definition of objectives and choice of experts.

*Preparation Phase:* Collection and processing of data, includes editing, deleting of obviously false or misleading data. Creation of a "Mining Base" for following procedures.

*Mining Phase:* actual search for interesting patterns. Sometimes it's necessary to repeat steps from earlier phases to add new data to Mining Base.

*Analysis:* Results of search is analyzed and processed in order to make them understandable for non-experts. This objective is achieved by interpretation and visualization of gained results.

### Data derived by Data Mining

Data and patterns which were derived by data mining should posses certain attributes (cf. [Sei03]):

*understandable:* patterns have to be easy to understand and graphical visualized

*valid:* patterns should be valid too for data collected in the future

*useful:* patterns have to be useful and relevant for the particular objective

*non-trivial:* only interesting patterns are presented

*interesting:* discovered knowledge has to be interesting too

Most importantly the data should be described in a way that is easy understood by people. Common patterns of mined data is listed below:

*Rules:* Data in databases is divided into classes. Attribute classes describe attributes of objects and regression rules predict numeric values.

*Cluster:* Data is divided into groups, due to statistical procedures. To each group a description is added, to make them better understandable.

*Dependency Patterns:* These patterns determine dependencies between variables of relevant data. Which groups, attributes turn up together?

*Connection Patterns:* Determines regularities among different objects within one database or among different databases.

*Sequence Pattern:* Search for regular occurrence of event sequences

### Data Mining Techniques

Common Data Mining techniques are listed below (cf. [Sei03]):

*Cart Analysis:* tries to find products, which are often bought together. When associated with customer-data it is possible to derive what will be bought by certain customers in the future. This analysis is also used for direct marketing, by remembering customer preferences and giving notices of special offerings according to these. The cart analysis is part of the cluster analysis group.

*Case-based Reasoning:* By remembering past experiences these method tries to derive future decisions. Attributes and parameter are stored in a database and when a decision has to be made, this database is searched for attributes and parameters, which resemble the decision the most. The more resemblance is found with stored attributes, the more accurate the prediction will be.

*Decision Tree:* More complex decisions contain of numerous smaller decisions. A decision tree is build of nodes and at each of these nodes, attributes and parameters are queried and a decision is made until a node is reached, where it isn't possible to make further decisions. The technique looks for the most fitting attribute and parameter at every single node that is reached during decision making.

*Neuronal Network:* A neuron processes an input to produce some kind of output and can be associated with other neurons to share information. Input as well as output is traveling between neurons - hence a network is created that is capable of processing information. Particular input produces particular output, but like in a *black box*[22] the processing process, that creates the solution, isn't visible.

*Genetic Algorithms:* tries to derive an optimal solution from randomly picked base solutions. To do this, genetic algorithms are used. Most fitting proposals for solution are picked and recombined. The result builds the new base for solutions. These solutions are sometimes changed randomly to resemble something like mutation[23].

*Automatic Cluster Analysis:* This technique is mostly the first step during data mining of huge databases in order to build groups of information. These groups are further processes with other DM techniques.

*Relation Analysis:* is trying to associate individual elements of information in a database. This analysis is only able to process structured information that is specifically conditioned.

### Data Mining and Privacy

Main objective of data mining is to create new knowledge from already existing knowledge. This ability could be used for great benefit for society, but could also be abused. After 9/11 data mining got even more important for the search of potential terrorists.

Often non-anonymised data is needed and it is impossible to know the nature of results that are delivered beforehand. Hence the fact that the already existing knowledge does not violate privacy, does not necessarily mean that newly gained information continues privacy protection (e.g. by taking public data to derive sensitive data in private or ethical way). Numerous studies and surveys (cf. [Vai02], [Ful04] and [Kan04]) covered the issue of data mining and (protection of) privacy.

To determine whether some data mining method violates privacy it is necessary to know (cf: [Kan04]):

- Which particular information is considered sensitive?
- To whom it is sensitive?/Whose privacy is at risk?
- What else is known?

22 system, where input and output is visible, but the (mostly complex) procedure in between isn't
23 spontaneous changes in a system - in this case in order to investigate new possibilities

- What is an acceptable trade-off between privacy and benefit of the result? How is this trade-off measured?

In [Kan04] it was also tried to build a so-called classifier that keeps privacy violation at a minimum during data mining. This classifier conducts the actual data mining process and is designed as a black box in order to hide the procedure of deriving new data. When the procedure is visible (like in neuronal networks or decision trees) sensitive data used could easily get exposed. Someone who wants data mining to be done, has only access to the classifier itself and gains no further information about the classifier (e.g. how it actually works) by usage of cryptography.

The question, however, what is considered private information and what not, is not that easy to answer. Many privacy laws include a so-called trade-off between cost and benefit for use of private information. Some of these laws include a provision *'in the public interest'* when it comes to the usage of sensitive information and loss of privacy caused.

The study conducted in [Kan04] identified the amount of given data examples to the classifier as the most important threat to privacy. The more examples are offered as input the more difficult it gets to trace information back to its origin. Every time sensitive data and unknown data is strongly associated with one another, privacy is threatened.

The survey in [Ful04] identified rules that are used to derive information during data mining pose a threat and should be monitored (also according to their *interestingness*[24]). These rules are usually created by domain experts - people who have profound knowledge of that particular branch data mining will be used in. Other additional measures which could protect privacy are

- limit access to data
- restrict scope of queries
- hide/delete data
- alert users, when rules are used that could probably violate privacy

*Privacy preservation* is needed in many different situations including:

- secure sharing of data between companies
- guarantee confidentiality of data available to the public, so people can't be identified by aggregated data
- anonymisation of private data by mutating or randomising data
- Access control not just to databases

Techniques to maintain privacy preservation are listed below (cf. [Ful04]):

*Authority Control and Cryptography:* data is hidden from unauthorized access, but inappropriate use of authorized users isn't prevented

24 measures whether a rule useful for accurate results or indeed misleading

*Anonymisation:* any attributes, which enable identification, are removed from source data

*Query Restriction:* tries to detect threats to privacy caused by combination of queries

*Dynamic Sampling:* chooses a different data set for each query and reduce size of data set

*Noise Addition:* change results to protect privacy, while still keeping accuracy

*Multiparty Computation:* data mining is conducted on numerous sites and results are combined

To protect privacy during the process of data mining, the main objective is to find a DM model, which doesn't reveal the data used. This could be achieved by altering the data before use (and accept a loss of accuracy in results) or by secure multi party computation.

### 3.3.2 Social Engineering

In order to protect sensible data, not only hardware and software measures (e.g. virus scanner, firewall, ..) are important, the human aspect involved should not be neglected. In fact hard- and software mechanisms are useless, when it comes to social engineering attacks.

Security of (intellectual) property usually relies on a three step process (short: *IAA*) (cf. [Tho04]):

*Identification:* Identify the person

*Authentication:* Is the identification information given valid?

*Authorization:* Is the person allowed to retrieve asked information?

Security maintained by following these steps is in danger, because people are easy to manipulate and are often the most dangerous vulnerability. Here comes social engineering into play, which is described as the obtaining of restricted information. Contact with secret carrier is established by (pretended) social contacts (cf. [SOC]). Hence social engineering could be defined as *a hackers clever manipulation of the natural human tendency to trust* (cf. [SOC01]). In [Tho04] SE is simply defined as a *social/psychological process by which an individual can gain information from an individual about a targeted organization.*

Social engineers, people who use these tactics to gain information, are not necessarily hackers, but hacker-enablers (cf. [Tho04]). They collect information during their attacks, which can be used by hackers to break into the company's network with technical measures.

Reasons for Social Engineering are

- industrial espionage
- identity theft
- boredom
- longing for power

- financial reasons
- social reasons

Social aspect is often as important, because the social connections needed are sometimes already established and ready to be used. For example ex co-workers are picked as victims because they are close, well-known or probably chosen for revenge reasons. Targets for such SE attacks are usually telephone companies, famous companies (e.g. ebay), financial institutions, military and government organizations (cf. [SOC01]).

### Types of Social Engineering

[SOC] tried to describe different forms of social engineering (SE). These are listed below.

*Computer based Social Engineering:* Like in other forms of SE, the attacker tries to pretend to be someone else in order to get access to valuable data. But this kind also uses technical measures like e-mails or websites to get the job done. A famous example for this kind of SE is phishing.

*Phishing:* is some kind of SE, where the attacker uses technical measures like websites to get people to disclose valuable information. Usually emails are sent to numerous users and some of the receivers are actually customers of the company (e.g. bank institute, ebay, ...) that seems to be the originator of the message. The content of these mails could include a url-link to a site that was created by the attacker. It usually asks users to enter valuable, restricted information like username and password (in order to proceed business with these company). Data entered by victims is then saved by the attacker and used of course.

*Human based Social Engineering:* This kind of SE establishes the actual human contact between attacker and victim, which isn't channeled through email-messages or websites. Usually the attacker poses to be a person of authority (which people are likely to obey to) or some technician, who not actually works at the company, and has to solve a problem for the firm. Creating a fake stressful situation helps to increase the victims stress level and make them obey to rather unusual demands.

Sometimes hackers even cause a problem in the targeted network, only to provide help solving it directly at the company's building. While being at the building, the hacker has the chance to get access to even more information and because the problem is indeed solved afterward, people don't tend to get suspicious.

*Dumpster Diving:* To get access to private information, attackers first have to get their hands on information, that seems rather trivial at first glance. Trash at companies is searched for interesting information like company phone books, organizational charts, memos, company policy manuals, calendars of meetings, events or vacations, system manuals, printouts of sensitive data or login names and passwords, printouts of source code, disks or tapes, company letterhead and memo forms, and outdated hardware (cf. [SOC01]). Data like these files can become useful, when searching for people the attacker can impersonate while conducting the attack. Organization charts are useful for determining

which employees have authority at the targeted company or calendars show which employees are not working these days.

So dumpster diving (and *shoulder surfing*[25]) can be seen rather as a research tool, than a actual attack method (cf. [Tho04]).

This kind of attack/tool isn't necessarily illegal. Trash in a dumpster on the street isn't property of the former owner anymore and can't be considered private. As long as the attacker hasn't to access restricted grounds it is legal.

*Reverse Social Engineering:* When using this method, attackers often simulate some kind of stressful situation, a (technical) problem, for which the attacker later pretends to provide help. Hence the attacker seems so be a ally rather than a dangerous attacker or impostor and people more likely obey to unusual, strange demands.

This kind of attack can provide even more valuable information than the types discussed above. However procedures like this need more preparation (and pre-hacking).

### Procedure of Social Engineering Attack

During an attack the SE attacker tries to by-pass any security precautions that are probably in place. This often includes hard- and software implementations. Every little bit that is gained during an attack can be used to by-pass security systems or saved to be part of a future attack on the same target.

According to [Tho04] the profile of an SE attack is similar to a Software Development Life Cycle (SDLC), called *Social Engineering Cycle*. Is described to consist of 4 distinctive steps:

1. research
2. developing rapport and trust
3. exploiting trust
4. utilizing information

Below the steps, which are usually taken when conducting a social engineering attack, are described in detail (cf. [SOC]):

*Collection of Information:* First collection of Information in order to prepare the actual attack, usually involves measures like *dumpster diving*. This step is important in order to conduct a successful attack on the target. Information gained helps during decision making, when deciding on identities the attacker could adopt.

*Establish Contact / Fake Identity:* During this phase the contact to the victim and the fake identity are established. Contact could probably be established by

25 peek someone "over the shoulder" when he/she is handling/entering valuable information

- phone poll
- query from end-user
- call from ISP
- conversation with supervisor
- etc.

To develop trust, attackers usually place calls while posing as another member of the organization, service provider, costumer. Through using verifiable information of employees and use of organizational jargon, attackers often succeed in making people believe, that they actually trustworthy (cf. SOC04). For attacks conducted by email the situation is quite similar. Attackers try to compose emails that are difficult to distinguish from actual officially sent messages.

Call centers, or rather the employees working there, are extra vulnerable to attacks by phone - people there are especially trained to answer questions. By asking the right questions (and faking the caller ID by technical measures) information could be gained to get access to higher authorities.

Attacks conducted by email (also known as *phishing-mails*) can also contain links in the mails content, which lead to a website created by the attacker. Another version of pishing is to pretend to be for example a system administrator of the targeted company, who asks employees for their passwords because of some problem.

SE attacks, however, aren't necessarily conducted by phone or email. Strolling into the company's building, while pretending to be some maintenance guy or janitor, could do the trick to. When not acting suspicious, it could be possible for the attacker to search workplaces for notes and similar that contain information about passwords or access codes. These obtained codes could be used later to gain remote access to the companies network, e.g. login as a normal user and gain administrator rights through privilege escalation. One of the easiest ways remains to simply watching users while they are enter their passwords(cf. [SOC01]).

*Gain Information at the Target:* This step should provide the actual access to the desired information. With clever questions the attacker tries to make the victim disclose the wanted information, without asking for it directly. The collection of information has to be conducted very carefully in order to not raise suspicion at once.

*Steps after successful Attack:* After the attack has succeeded and the wanted information has been collected, it is important to keep certain things in mind:

- remain contact, for later usage and future attacks
- (malicious) collection of information shouldn't be revealed or obvious. Otherwise the information could become useless (again).

*Composition of gained Information:* Usually the information gained arrives only in bits and pieces and yet has to be composed to be useful, e.g. derive passwords for future attacks from trivial data like date of birth, names of children etc.

### 3.3.3 Identity Theft

To be able to use services provided online, people need to disclose (sensitive) information. Webshops ask for data like resident address and creditcard number in order to complete transactions and deliver ordered products. Another serious issue related to data disclosure online regards resumés and CVs. People in pursuit of employments provide detailed information additionally to their job history e.g. their Social Security Number (SSN) (cf. [Swe06]). Once the information is publicly available it stays that way (even after the data was deleted), thanks to search engine caches and Internet archives (e.g. `http://www.archive.org/web/web.php`). Identity Theft is defined as using *another person's identifying information without permission to commit fraud or other crimes.* To apply for a new creditcard in the US, people need to disclose their SSN, date of birth and the mother's maiden name. Hence people should be extra careful when choosing where to disclose this particular identifying data, which is very valuable for criminals.

US courts, however, decided that a victim is not liable for charges made by forged creditcards as long as victims did not know about these cards. Once a forged card appears on a credit report (and the victim is notified of the fraud), the victim has to immediately begin correction, otherwise, he/she is responsible for the dept.

More than a quarter of identity theft cases (in the US) involve creditcards (cf. [Swe06]) and the consequences of identity theft are versatile:

- lost job opportunities
- refusals for education, housing or cars
- arrest for crimes not committed

The risk of identity theft, however, is very high for US citizens, because of the numerous cases the SSN is needed for (e.g. applying for a new creditcard).

### 3.3.4 Communication Monitors

As the name suggests communication monitors are tools, that are able to monitor (electronic) communication. That includes packet sniffers for (Internet-)traffic analysis and wiretapping tools, that are capable of monitoring phone calls as well as parsing of e-mail or instant messages. Communication monitors existed for some time, until they became finally matter of public concern after 9/11.

#### Carnivore

Packet sniffer used by the FBI to monitor electronic communication. FBI claims that Carnivore was abandoned back in 2001 in favor of commercial software.

#### Echelon

While Carnivore is a packet sniffer, Echelon represents a global network, that is capable of parsing electronic communication (e.g. email, instant messages, ...) for particular keywords. The processors in the network are called dictionaries, which are connected to each other. Hence ECHELON works like a huge integrated system. [ECH]

In a report published back in 2001 (cf. [EUR01]) the European Parliament found that communication monitors do not conflict with EU laws as long as they are used for maintaining national security and not for clearly illegal purposes like industrial espionage. Nonetheless it is advised in the same report that European citizens should use cryptography and other methods to maintain privacy of communications.

Known members of ECHELON program are agencies from the USA, Canada, Great Britain, Australia and New Zealand. Stations that work as sender and receiver for the network's communication satellites are allegedly located all over the world (e.g. Misawa in Japan and Bad Aibling in Germany).

## 3.4 Summary

Apart from the risks and threats described in this chapter, the biggest problem poses the fact that many people are not even aware of the threats they are exposed to when searching for news about their favorite topic or checking their emails. The previous section gave an (incomplete) overview of different kinds of threats user face when participating in the online world.

The descriptions given included tools like cookies and spyware. While spyware can pose a serious threat to people's privacy by collecting data stored on local machines, cookies used by most websites, on the other hand, act like a tracking device, which could be used to derive profiles of people's interests.

Barring more or less dangerous tools there also exist organizations, which could threaten people's privacy. An organization people have to rely on for establishing a connection to the Internet is the Internet Service Provider (ISP). To provide the service the ISP needs to know master data related to a particular customer including address and payment information. This information can be disclosed, if the ISP is forced to do so (by law enforcement). When providing access to the Internet, every customer is provided with a (in most cases fixed) IP-address that is sent with every request which is relayed through the ISPs servers. Hence, every request sent by a particular user can be linked to this issued IP-address.

Huge companies, which attend to many different branches, could be dangerous to people's privacy in terms of collection and combination of information. One particular company is the well known *Microsoft* company. The market leader in operating systems and related software, however, is not the only one posing a threat to privacy. In past years another company, now well known as *Google*, provided a search engine, which soon became the most widely used. Shortly after this huge success, Google started to provide other services which currently include an email-service (GMail), opportunities to store and code documents online. Few months ago the company even released its own web browser called *Chrome*. All these (often free) services owned by one single company need different kinds of data in order to provide functionality. This information is stored and used for advertisement purposes by Google.

Other methods, which were described in this chapter, included *data mining*, an approach trying to predict customer behavior based on collected data and *social engineering*, which mainly manipulates human behavior in order to gain information.

After giving an overview of existing threats and risks, the next chapter will focus on

| Threat | Action Track. | Privacy Viol. | Data Theft |
|---|---|---|---|
| **Cookies** | x | x | |
| **Spyware** | x | x | x |
| **Proxy** | x | x | |
| **ISP** | x | x | |
| **Google** | x | x | x |
| **Data Min.** | x | x | |
| **Social Engine.** | | | x |

**Table 3.1:** Summary of Threats to Privacy

methods to protect one's privacy while surfing on the Internet.

# CHAPTER 4

## Privacy Enhancing Technologies (PETs)

Privacy Enhancing Technologies (PETs) are an approach that seek to protect ones privacy, while transacting data. Burkert identified in [Wal01] four types of PET concepts:

*Subject-Oriented:* tries to eliminate or reduce the ability of identifying the transaction subjects (e.g. proxies).

*Object-Oriented:* a transacted object leaves traces and this concept seeks to eliminate them (e.g. paying cash, not with credit card).

*Transaction-Oriented:* this concept covers the transaction itself, without touching the exchanged object (e.g. video surveillance).

*Systems-Oriented:* tries to combine all concepts by creating zones in which the identity of transaction-members is hidden, the exchanged object and the transaction itself leave no traces.

The figure below shows three principles to protect privacy while surfing online. These principles will be described more detailed in this chapter.

*Anonymity:* the users identity is fully concealed while being online and making requests to servers.

*Pseudonymity:* many services provided on the web need at least some level of identification and data about users (e.g. to customize layout, content, design according to user interests or needs). Users are represented by some kind of unique identifier (a pseudonym), which however can't be linked to the users true identity. Pseudonyms are often used by online services to "recognize" (identify) a user, when he is visiting a site again. User are only represented by their pseudonyms - hence other users get only to see these pseudonyms and never the identity hidden behind it. Examples for such services are online boards, news reader, discussion groups.

*Negotiation Agents/ Trust Engines:* enable some kind of standard for privacy policies. Hence they are easier to analyze and compare - it is easier to determine whether a particular site's policy is violating or satisfying a user's need for privacy.

Additionally this chapter will take a closer look on *encryptions*, which aims to protect data while it is transmitted.

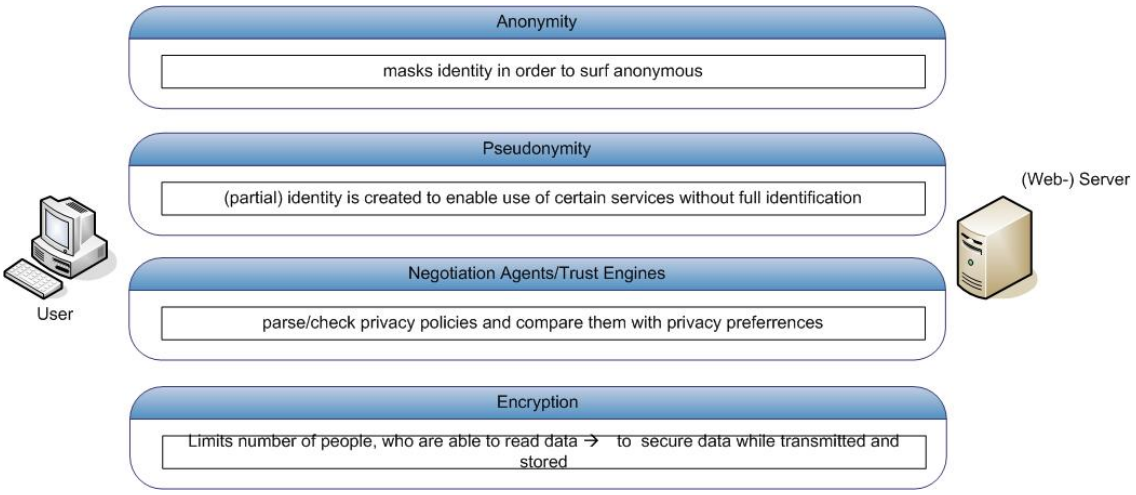


**Figure 4.1:** (Online-) Privacy Enchancing Technologies

## 4.1 Anonymity

With an *anonymous internet connection* it should be impossible to link a number of actions to a particular user. To maintain such a connection two tasks have to be accomplished:

*user anonymity:* Sent packets are modified to contain no identity information.

*unlinkability:* It should impossible to link actions and transmissions to a particular user.

Furthermore there exist different types of anonymity that should be taken in account when implementing some kind of anonymity (cf. [Kob03]):

*Environmental anonymity:* this kind of anonymity is influenced by external factors (e.g. number and diversity of users) and can't be changed by the implemented system - hence they have to be monitored over time.

*Content-based anonymity:* is in tact when no identification is possible by data exchanged. Can be broken by analyzing the content itself (e.g. by looking for addresses, names, ...), the data structure or by looking for certain sequences.

*Procedural anonymity:* depends on the communication protocol, therefore can be provided by the application. Hence this type of anonymity should be considered already in design phase.

*Sender anonymity:* Sender can't be identified by receiver within the set of potential senders.

*Receiver anonymity:* Receiver is not known by sender.

To build a system that provides anonymity, all these types described above must be implemented (cf. [Kob03]). There are not only different types of anonymity identified, but levels too, which can be taken in account by application/tools considering anonymity:

*Super-Identification:* user is clearly identified within the system/application by some other certification authority

*Identification:* user identifies himself to the system

*Latent Identification:* user identifies himself and a pseudonym (see section below) is assigned to him, by which he is recognizable by others in the system

*Pseudonymous Identification:* User himself chooses a unique pseudonym, which stays valid in following sessions

*Anonymity:* user uses the system/application without any identification (necessary)

Below are some techniques listed, a application, system or tools could provide when dealing with anonymity (cf. [Lin07]):

*Anonymous Routing:* A anonymity tool should provide a solution for both tasks (unlinkabilty and user anonymity). Unlinkability could be maintained through *anonymous routing*. Different approaches to achieve such routing are listed below:

*mixed net:* a computer, a so-called *mix*, randomly permutes received packets. Transformation of these packets is achieved by encryption. This method can be improved by using a number of mix-machines (*mixed net*). This approach works properly if at least one of the mix-machines is honest and all users (clients) and servers receive almost the same amount of traffic (otherwise is would be easy to determine which user connected to which server).

*onion routing:* A number of mixed nets is picked randomly. (see section below)

*dining cryptographers:* Parties broadcast messages to make it impossible to determine which party has sent which message. Once key and bits for encryption are set, no further interaction is needed. (see Encryption)

*crowds:* User form crowds before sending messages. A user sends a message to a random member of the crowd. This member forwards it to another member or to the server (task is randomly picked). So the server can not determine which user had originally sent the message. (see below)

*Anonymous Authentication:* While using anonymous routing however some server still require authentication (e.g. digital libraries). One solution could be *anonymous authentication*. The server knows that he is dealing with an user belonging to a group of authenticated users, but the server can not determine which particular user tries to connect. This method also works for applications with different permission levels. Requirements for this approach are:

*Secure Authentication:* should ensure that no unauthorized user gets access to the server

*Anonymity:* the server should not know with which user of the authenticated group he is interacting with

A so-called *ring signature* should maintain something like anonymous authentication. This signature guarantees that a signer of a message is a member of a set of users, without the possibility to determine which particular user actually signed the message. If a user uses *SSL* (Secure Sockets Layer) to connect to the server, the digital signature is replaced with the ring signature.

When hiding among a crowd, the size of the crowd should not fall below a certain limit. Otherwise it could be easy to investigate which user is currently communicating with which server. On the other hand it could be difficult to deal with huge crowds. Therefore user-sets are often divided into sub-sets.

*Password-based anonymous Authentication:* Lesser level of security than other methods. Standard Authentication Protocol is used to establish a connection to server. In the next step a temporary public key is generated. After disconnect the user is reconnected with anonymous authentication and the generated public key.

*Revocable Anonymity:* should allow to determine the identity of an particular user among a set of authorized users when needed. A user authenticates with encrypted info about his/her identity. This info should be decryptable to obtain the identity. The encryption of identity information should be bound to the transmission in order to prevent users from sending garbage instead of the actual identity. The server can not decrypt the information, thus it can not check if the transmitted information is valid. Therefore smartcard should be used to compute the ringsignature and the authority certificate (identity information) should only be readable but not modifiable by a user.

### 4.1.1 Proxies

Proxies were already mentioned in the risks chapter, because information transmitted over a proxy can be monitored by the proxy's owner. Hence the owner is able to determine the users identity. With this risk in mind proxies can be used to protect user privacy. Like stated in the chapter above user can transmit all requests over a proxy and the targeted webservers are only able to see the proxy's IP and not the address of the user.

So-called *proxy-lists* (e.g. `http://www.samair.ru/proxy/`) provide URLs to numerous proxies to be added in browser properties. Apart from support of different protocols, proxies can provide different levels of anonymity. Proxies which provide higher levels of anonymity are usually not free of charge. Hence proxies can be used for different actions (e.g. p2p sharing applications without the risk to be identified).

Proxies can also be used to "fake" the requests location of origin. Some websites restrict content or whole websites to IP addresses located at a specific place. While using a proxy located at that particular place (e.g. country), the restriction could be easily by-passed. While this possibility may come in handy for some users, which otherwise may not be able to view content, others may see it as a threat, that people may be able to access content, which wasn't intended for them in the first place. A popular example for content restriction are full episodes of TV shows provided as streaming videos on the CBS webiste (`www.cbs.com`). Access to that videos is restricted to IP addresses located in the US.

### 4.1.2 Onion Routing

An application that makes use of above described proxies to remain anonymous online is *Onion Routing* (cf. [Gol99]). When OR is activated the paths to retrieve information/data are constantly changed. Currently supported protocols are HTTP, SMTP, FTP and more. The proxy used by OR consists of three logical layers:

1. *application-specific privacy filter*: sanitizes the data stream
2. *application-specific proxy*: translates data into a application-independent stream that is accepted by OR
3. *onion proxy*: builds and manages anonymous connections

Data Movement in the network consists of different stages:

*Setup:* a onion [1] is created that defines the path through the network.

*Data Movement:* Each *onion router* on the path uses a public key to decrypt the entire onion. The information gained consists of the identity of the next onion router, cryptographic information and of course the onion (with the actual data) itself. After padding (to maintain a fixed size) the onion is sent to the next onion router. Once connection is established an required information is distributed, data can be sent in both directions. When data is moving along the predefined path every router removes a layer of encryption from the onion (as set in cryptographic information) and the data arrives in plain text at the recipient. To remove data backwards (as a response) the encryption layering occurs in the exact reverse order. The data is sent in fixed-sized cells, with numerous onion layers wrapping it. Every cell appears different to each onion router along the path, because of adding/removing of cryptographic layers. Hence the design resists traffic analysis more effectively.

*Tear Down:* connection tear down can be initiated anywhere (beginning, end, in the middle)

Different kinds of cryptography are used during (dis-)connection phases. Computationally intensive public-key cryptography is only used during setup, therefore connection overhead is rather small. During data movement only the much faster secret-key cryptography is used. Data throughput on the other hand depends on the number of routers along the path.

Onion Routing can be used with anonymizers and LWPA [2] (see below). Examples for applications that use Onion Routing are

- TOR (The Onion Router) `www.torproject.org`

  TOR is written in C and should work with Windows, Linux, Mac, .... Traffic analysis is still possible, when an attacker is able to watch both ends of communication [TOR].

1 has numerous layers, which define properties like cryptographic information at each point of connection
2 Lucent Personalized Web Assistant

- JAP (Java Anon Proxy) `www.jondos.de`

  This client application is based on java and therefore runs on every system with java runtime installed. It can be used as a client for TOR.

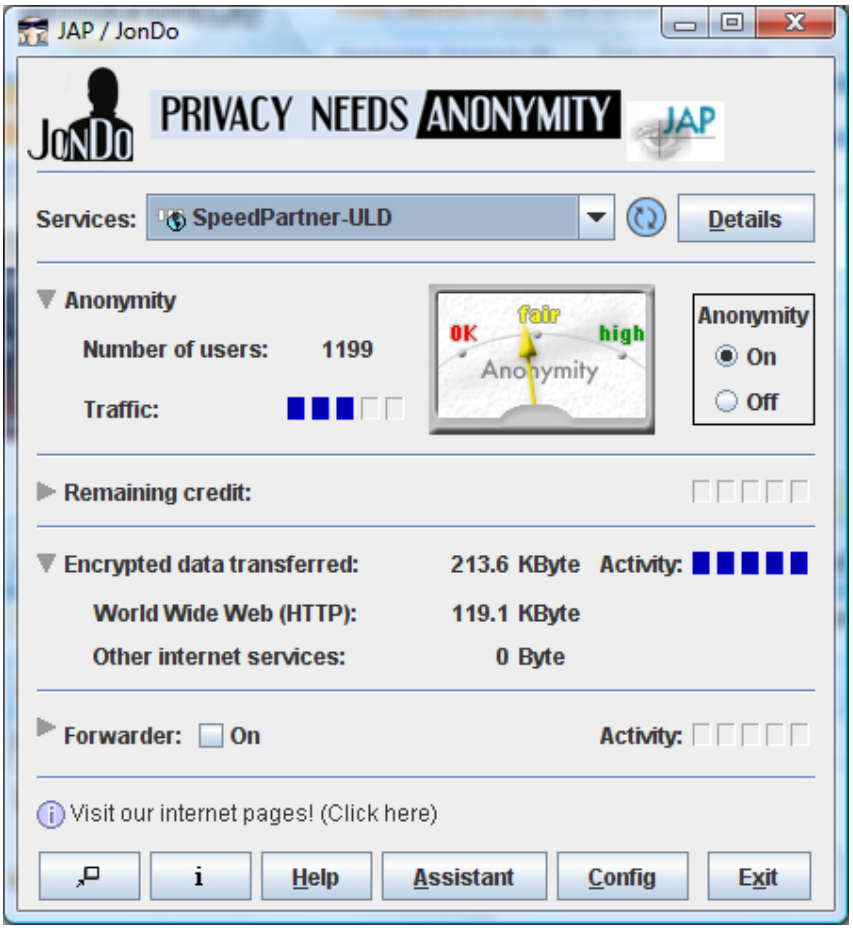


**Figure 4.2:** JAP/JonDo Main Interface

Once JAP is activated and the connection to one of the available proxies is established, all requests are redirected over this proxy. Targeted webservers then are only able to see the IP-address of the chosen proxy. Level of privacy as well as speed with which requests are forwarded, both depend on the number of users currently active in a particular network. The more people use the network at a certain time, the higher level of privacy is. Proxies are only able to forward limited amount of data at once, hence connection is slower when many people are using the proxy at the same time.

JAP advises however to disable Javascript and ActiveX functionality in webbrowser preferences to fully avoid identification.

### 4.1.3 Crowds

As mentioned in previous chapters surfing on the internet could be a serious threat to privacy. Every time a connection to a server is established, data is sent. Certain aspects

of that data could possibly be used to identify who is trying to make a connection. The developers of *CROWDS* ([Rei99]) try to give users the ability back to decide what kind of data is revealed and when this is happening (similar to *P3P*).

*CROWDS* hides (or corrupts) certain information (like IP-address, domainname, referer, OS, ... ) so a made request cannot be linked to a certain user. Hence these tool makes it possible to retrieve information online, without revealing too much information about oneself.

To protect users who utilize CRWODS they are added into a group called crowd. Every member of the crowd runs a process named "jondo" ("John Doe" - alias for not yet identified patient or person) on his/her local machine. When a user wants to join the crowd, the jondo engages into the CROWDS-protocol and the user is informed of the other users in the crowd and vice versa. After the new user was admitted to the crowd, he/she is able to use the crowd to send requests.

Sending an request anonymously is made possible by the jondo-process, which is working as a proxy that forwards every request made by the users browser. The request is then sent to another (randomly chosen) member of the crowd (jondo) or the target-server. Whether the request is forwarded to a member or the target server is chosen randomly. This choice is based on an system wide parameter ($pf$) that has an impact on anonymity properties.

While a request is traveling to the target-server is uses a certain path. This path consists of a number of hops across numerous members of the crowd. If a request is sent to an jondo, this process records the jondo which sent the request and the jondo (or server) the request is forwarded to. When the server eventually sends a reply, this reply takes the exact same path as the request took. By recording jondos that communicate with a certain jondo, it is possible to use a path more than once. A once established path (identified by path-ID) remains static as long as possible in favor of performance. A path through the crowd is only altered when one member of the crowd doesn't answer anymore or a new member joined. If a new member joined, all established paths are erased and replaced by new ones. Hence new members have the same privacy status as all the other crowd-users.

The communication between the jondos is of course encrypted by an cryptographic key. One particular key is only shared between certain jondos to enable different classes of privacy properties.

CROWDS provides certain *degrees of anonimity*:

*beyond suspicion:* attacker can see evidence of sent messages, but can not tell which member of the crowd originated the request

*probable innocence:* one crowd member is not more likely the originator of the request, than any other

Of course there are also *risks and limitations* to the potential of CROWDS:

- The IP-address of the jondo which ultimately sent the request to the target server can be logged, although it is impossible to determine who originated the request in the first place. The authors of [Rei99] state that the owner of that particular jondo which IP was logged, can simply explain, that he/she was using CROWDS at that moment and didn't originate the request

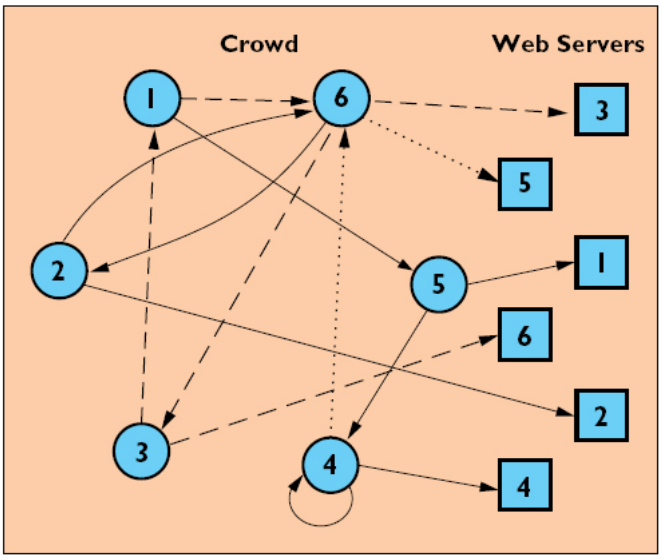


**Figure 4.3:** Paths in a crowd [Rei99], the number of request originator and target server is the same. Loops in paths are possible too

- CROWDS protects the anonymity of users, but not the confidentiality of the requests content. The content is accessible to other members of the crowd (jondos). Hence CROWDS should not be used, when transmitting confidential data, like accountnames and passwords (this data is processed by the target server anyway)
- With mobile code like Java and ActiveX CROWDS can be circumvented. Such code opens a direct network connection to the Web Server, which served the code. This connection is not directed through the jondo-proxy and the user is exposed to the server. Hence the authors recommend to disable Java, ActiveX, ... while using CROWDS

The publication [Rei99] also mentioned page retrieval times and network load as notable limitations. In 1999, when the article was published, technologies like modems and dial-up connections may had been a handicap for the usage of CROWDS, but as of today broadband connections allow much faster data retrieval than only a few years back, so the increase in traffic caused by CROWDS may be not as noticeable as back then.

## 4.2 Pseudonymity

The chapter above covered the topic of *anonymization*. Tools provide services that keep users, who are surfing online anonymous as good as possible, by prohibiting linkage of individual actions, respectively of numerous action to one particular user.

Many services however need at least some level of identification for proper usage (e.g. news reader, online boards, ...). The challenge here is to protect private information, while enable proper usage of services that need user identification, to “recognize” user, when they visit the site/application again, or even provide personalized interaction and try to

adapt to user needs. User-Adaptive Systems[3] (UAS) on the other hand need as much data about people using the application as possible. They provide customized services like displayed news according to peoples interests, storing preferences and preferred layout. User-adaptive system aim to *supply each user with all relevant information in form that is suitable for him and does not overtax him* [Kob03]. In order to build trust with people and therefore make them to disclose data in order to enable services to adapt to users needs, this (mostly private and sensitive) data has to be protected properly.

To gain the ability to adapt to users needs information is stored about characteristics each user posses. These characteristics are stored in user models, respectively stereotypes (for user groups). Collected data includes (cf. [Kob03])

*User Data:* Demographic data, user knowledge, skills, capabilities, interests, preferences like layout, preferred topics etc.

*Usage Data:* Which parts of site/application are used, which actions are taken and how often

*Environment Data:* software, hardware used by user or the location

This data is associated with each user in order to enable proper functions of services. User models, in which these information about users are stored persist over sessions and aren't deleted when a single session is terminated - hence users are "recognized", when visiting the site/application again.

Examples for such are user-adaptive systems are *GRUNDY* in late 70's and early 80's, which used stereotypes and default assumptions as characteristics and *DPS* since early 00's which works with content-based filtering and production rules (cf. [Kob03]).

User models (UM) are used for adapting and are stored on so-called *user model servers*, accessed there by the system, whenever needed. When one particular user interacts with an user-adaptive system (e.g. a news reader) the system contacts the user modeling server and accesses the fitting user model. With the retrieved information the system is able to adapt display, layout, content etc. to the users needs, preferences. In fact more than one UAS can access an single UM server and retrieve needed user models, while these different UAS not necessarily need to be aware of each other and more often they aren't aware of each other at all. It is also possible that two or more user models exist on a single server, which are associated with one single user (e.g. for use in two different UAS). To guarantee full functionality, different relationships possible for user models should be taken in account when implementing user-adaptive systems. Content stored in user models could be accessed by two different UAS (e.g. one content used by one application includes content that is used by another application) or could be entirely separated for these two applications and don't interfere with each other.

In order to maintain protection of privacy when working with user models user demands have to considered that can be influenced by numerous factors (cf. [Kob03])

- general preferences in privacy concerning internet technology

3 systems that change the behavior according to user needs, respectively to information available on user

- should different user models connected to one single user kept apart from each other?
- personal roles an user wants to adapt when dealing with the system (e.g. employee at work, private person at home, ... )
- are the benefits gained by adaption worth disclosure of information?

To deal with aspects of privacy the connection between users and their associated user models is concealed to maintain some level of anonymity and encryption protects from unauthorized inspection, while data is exchanged between application and server.

Privacy is best protected with fully implemented anonymity, because there is no identification needed. User-adaptive systems however don't work properly with full anonymity in place. Applications like these need at least some kind of user-identification, in order to provide their services. These compromise is best achieved by *pseudonymity*, meaning that pseudonyms (automatically or manually) are assigned to users.

Below is a list of properties pseudonymous users of user-adaptive systems should maintain (cf. [Kob03]):

*Unidentifiable:* pseudonymous user should not be identified by third parties or the user adaptive system itself

*Linkable for the User-adaptive System:* UAS is able to link (inter-)actions made by single users (even across sessions)

*Unlinkable for Third Parties:* Third parties cannot link individual interactions to one single user

*Unobservable for Third Parties:* Third parties should not be able to recognize, whether a user is using the user adaptive system

Furthermore user should be able and allowed to adopt not only one single pseudonym, but several pseudonyms (e.g. for different situations/environments):

*Role Pseudonyms:* to interact with one application/system in different roles

*Relationship Pseudonyms:* for interaction with different systems in different roles

*Role-Relationship Pseudonyms:* combination of both types above

To keep peoples' private information or rather the relationship between on single user and his associated user model secret, certain tools and components are needed. These component needs to be trustworthy, because it is able to compromise privacy in a very harmful way. Hence there should always be more than one component for privacy protection. The more components are in place, the more secure a system becomes, because in order to defeat privacy of data all these components have to work together. The so-called *privacy complexity* is measured by the number of components, which are implemented.

Anonymity of private information is more than needed, because people tend to communicate more freely, when they vain their identity and information are kept secret. These information is necessary for user-adaptive systems to work properly. The more data they

gain on one single user, the better they can adapt to the user's interests and preferences. Ideally other users get only to see the user's pseudonym and only learn information, that particular user is disclosing willingly (e.g. in a posts content).

Conclusively pseudonymous access to user-adaptive systems should be provided, because

- demanded by users
- relationship between users and user models (private information) is hidden
- people are communicating more freely, when they have trust in the application's privacy protection - hence more data is available to the user-adaptive system

To maintain secrecy of data in user-adaptive systems, several techniques should be taken in account:

*Encryption:* user information is encrypted while transfered.

*Anonymization:* implemented by anonymous re-mailers, crowds, onion routing or mix networks. A mix network (as described in section above) is not only able to protect the user himself, but also the user model server. It is kept secret with which UM server a particular user is communicating, hence it is more difficult for an attacker to retrieve stored data for an chosen user.

*Selective Access:* when applications should only be able to access certain parts of user models, encryption and anonymization can't be applied anymore, because clients have to be identified in order to be authorized to access requested data. Security models providing selective access include:

*Chinese Wall security policy:* client chooses for itself which (conflict-) classes should be accessed - therefore it is not possible with this policy to prevent access to classes containing sensitive data

*Non-interference Model:* assigns clients to groups as well as keeping track of commands issued in the past. History saved in user models could possibly result in inconsistency

*Information Flow Control* on the other hand deals with the information exchanged or processed within the system. Describes flow of information and which flows are prohibited.

*Multi-level Security Model:* defines flow of information, hence prohibits unauthorized access. Downside to this is that the system can't alter data that was written by applications of lower sensitivity level. Untrusted applications are not able to access data that was written by application with a higher trust-level.

*Information Control Model:* enables definition of allowed information flows, hence this model is more flexible

Security models in general are not independent from the (user-adaptive) system they are implemented in.

*Hierarchical Role-Based Access Controll Model:* focuses on interface between user modeling system and user-adaptive system. Enables selective access to user models by defining roles, while being independent from structure of the system, it is applied to. Access models can be assigned to each role and these roles can be ordered hierarchically (including inheritance of access models). To each role a degree of trust can be assigned (untrusted, trusted and verified).

In [Kob03] a reference architecture for pseudonymous and secure interaction with user-adaptive system was introduced and it consists of following components:

*UM Client:* one or more user-adaptive systems or the user himself.

*Certificate Directory:* used to verify identities of UM and UM client.

*Role Server:* for definition of roles and the hierarchy and manages the assignment of clients to roles.

*Permission Server:* for assignment of permissions and authorization of requests.

*Mix:* for connection between user model and user model reference monitor.

*User Model Reference Monitor:* for control of information exchange between UM client and UM and conducts among others authentication, authorization and anonymization.

*User Model (UM):* for processing requests, which have already been authenticated and authorized by UM reference model.

In [BP05] Privacy-Enhancing Management tools (PIM) and secure representation of pseudonyms are discussed. While some examples for PIMs are introduces in detail later in this chapter, the following paragraph will focus on (secure) representation of pseudonyms.

The article takes a closer look on partitioning of sensitive data as a protection technique, while still enough data is provided to use services that need some level of identification (data about the user, e.g. addresses, bank account data, ...). PIM enable use of this technique. Identity of user is partitioned into several parts, e.g. every single website visited has access to another part of the identity, without knowing the entirety of the identity. Every site has access to exactly that information it needs and therefore enables the user to be recognized by the site. But these different parts of a users identity must not be linkable, otherwise user's privacy is violated again.

In general a pseudonym is used as an identifier, which can be created automatically or can be chosen by the user himself. Depending on the pseudonym chosen different degrees of anonymity are provided. For example when some created, chosen or often used pseudonyms are similar, the possibility is higher that these pseudonyms can be linked and the collected information can point back to the real identity of the user.

Degree of anonymity of pseudonyms is among other factors determined by the frequency a pseudonym is used. *Transaction pseudonyms* for example are only used for one single transaction, offer therefore the highest degree of anonymity. *Role-, relationship pseudonyms* offer a lesser degree of pseudonymity. They are used more than once, often while participating on online board discussion groups. There the user communicates with other

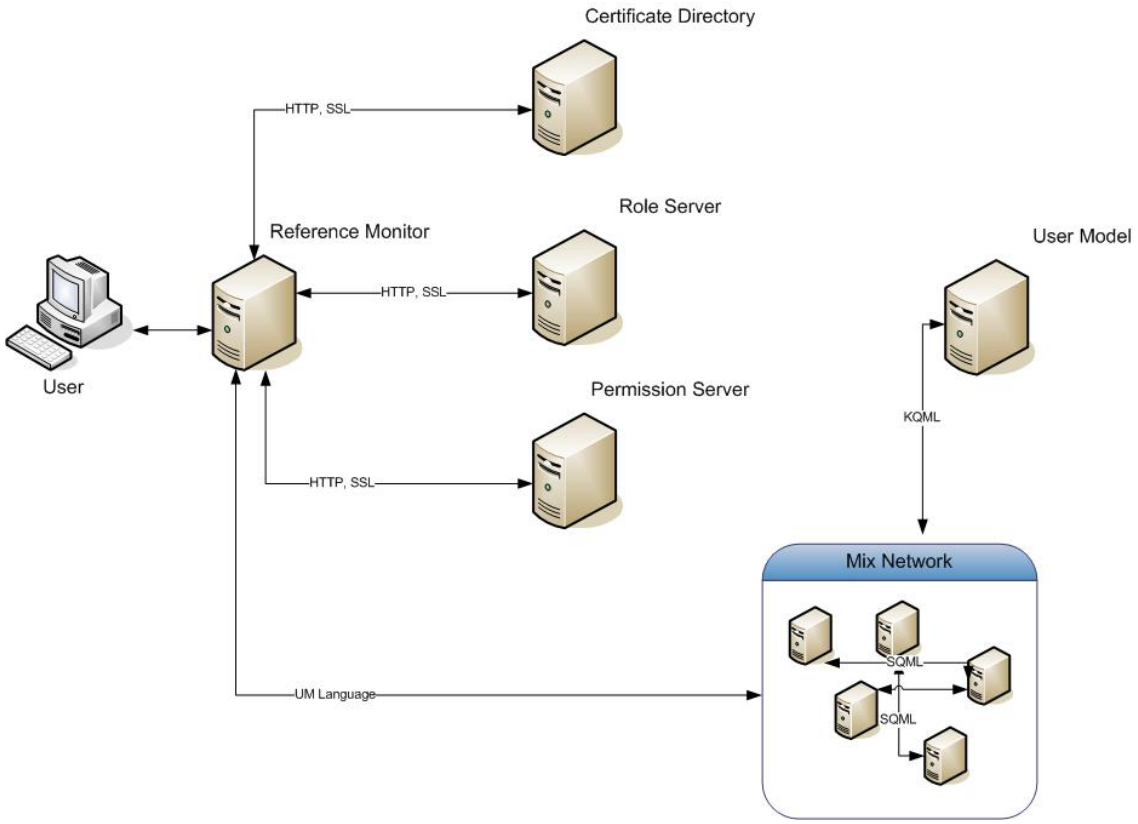


**Figure 4.4:** Reference Architecture for Pseudonymity in User-Adaptive Systems

users *while acting in a specific role* and his true identity is hidden from other users. The pseudonym class which provides the least degree of anonymity is called *person pseudonyms*. This class just replaces the name of a user by a pseudonym, hence unlinkability maybe not provided anymore.

Another factor that determines the degree of anonymity is called *internal representation* of pseudonyms. Automatic generated pseudonyms surely offer the best degree of anonymity, while usability is bad. User rather tend to use pseudonyms, which are easy to recognize. This preference is not that far fetched, when taking into account that recognition of communication partners is vital for a good discussion.

To solve this problem so-called *mnemonics* are used as an alias for automatic generated pseudonyms. These mnemonics are usually easy to recognize by users. Mnemonics are not bound to text-only representations, it is also possible to use images and sound. As long as mnemonics are used only locally (as representation on a single website) unlinkability isn't in danger.

Authors in [BP05] also identified *requirements of pseudonyms* and how they influence representation:

*Unlinkability:* pseudonyms must not be linked back to a single user.

*Recongnizeable:* on the other hand users must be recognizable by other users, to enable

communication, Some sites/transactions need authentication, authorization, so user must be identifiable somehow.

*Usability:* pseudonyms must be easy to use by users, otherwise pseudonyms may not be used at all. Usage of pseudonyms must be effective, efficient and should help people to achieve a desired goal. Pseudonyms are usable if they support users in managing partial identities (own or identities of other users). When usability isn't (fully) supported, security could be undermined, because user may try to simplify pseudonyms by themselves and unintentionally disclose information.

In general pseudonyms can be divided into two types:

*Randomly generated Pseudonyms:* are randomly generated by computer (e.g. arbitrary random bit strings, public keys), totally independent of environment. Two different pseudonyms can't be linked, because they are entirely independent from each other. Even observation of many pseudonyms doesn't help attackers to link individual pseudonyms.

*Pseudonyms chosen by users:* these pseudonyms are likely chosen according context it is used in. This helps user in recognize these pseudonyms (e.g. add some description to pseudonym). Because this kind of pseudonym includes semantics for better recognition, it is likely that it is possible for attackers to discover similarities in pseudonyms used by one single user.

Obviously there exists a trade-off between security and usability - pseudonyms that are secure, aren't easy to use. Hence generation and representation of pseudonyms have to be considered separately. Attributes of pseudonyms visible to other users must not disclose sensitive information, or help observators to link pseudonyms. Therefore PIMs should generate pseudonyms entirely at random, so that they are unlinkable and protect privacy. Users of the system then assign aliases, called *mnemonics*.

Security of mnemonics depends on the scope they are used in:

*Global Mnemonics:* are visible to all users of the site/application. But to ensure that mnemonics aren't linkable to single users, all mnemonics used in this site/application should be picked from one set. Otherwise if users are allowed by PIM to choose a mnemonic of their own, the PIM should at least check, if it is suitable and not undermining security. In order to keep security the dictionary of available mnemonics has to be large enough, so a individual (distinct and user-friendly) mnemonic can be chosen for every single member of the system. (Global) Mnemonics chosen by the system might not meet user preferences in terms of usability. People have to be familiar with mnemonics they use.

*Local Mnemonics:* Mnemonics are assigned locally by the PIM client. When communicating over the network only the automatic generated pseudonyms are used. Locally every user can assign his own mnemonic, hence usability here is higher than the method of assigning the mnemonics globally. Because the relationship between a pseudonym and a mnemonic is only valid locally, the same mnemonic can be assigned to a whole different pseudonym on another PIM client.

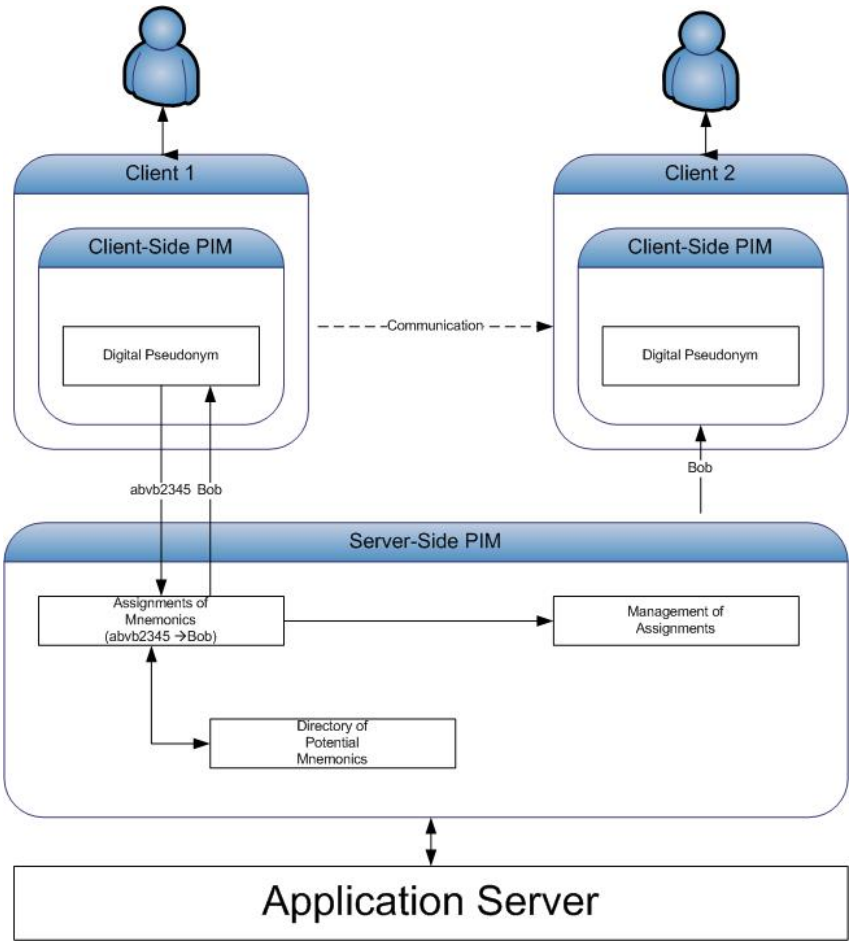


**Figure 4.5:** Assignment of Global Mnemonics

Global assigned mnemonics provide secure communication, while local mnemonics provide best usability. Authors of [BP05] advise to let the PIM generate random pseudonyms to which each user is allowed to assign his very own pseudonym. These pseudonyms however should be checked by the PIM for suitability and whether the pseudonym unintentionally discloses any sensitive information and messages should be checked if mnemonic was used and in case be replaced by pseudonym to avoid information disclosure. Furthermore PIM should inform users about the mnemonic/pseudonym currently used and the information that is linked to this partial identity.

When an implemented solution is working and people have trust in a system to protect their private data stored, they are more likely to disclose information when using the system. The more information a UAS is able to gain about a individual user, the better it it can adapt to the user's needs and preferences. Hence anonymity (as well as protection of sensitive data) is equally important in systems that need (more) detailed data about people in order to provide full range of services.

In the following subsection of this chapter numerous solutions and tools are described

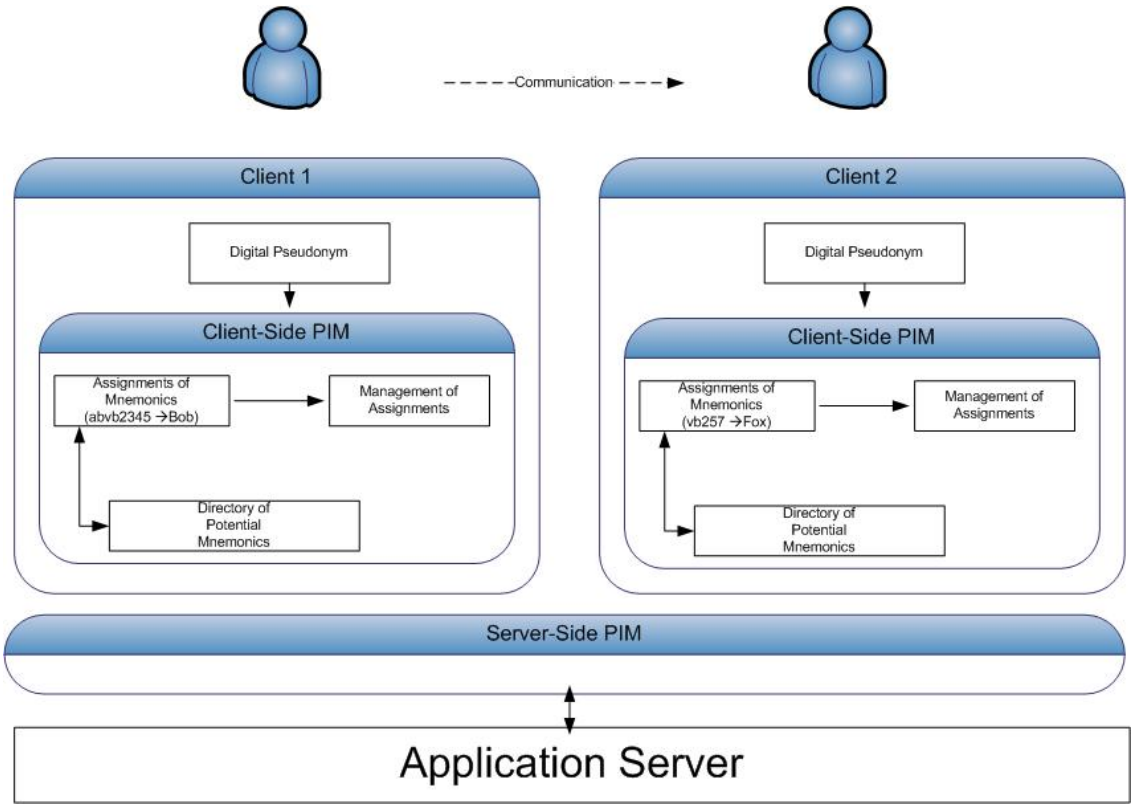


**Figure 4.6:** Assignment of Local Mnemonics

that use the principle of pseudonymity:

### 4.2.1 Lucent Personalized Web Assistant (LPWA)

Like mentioned above anonymizing tools are good at keeping peoples identities secret while surfing the internet. However often services provided online need at least some information about user visiting the site, in order working properly. Online discussion boards or simply sites, where people are able to create accounts, need at least data like username, password and often a valid email address (for account activation and validation of given email address). Some sites just ask for an email address, like websites providing newsletters. When email addresses are disclosed however, there is always a potential for abuse (e.g. spam, junk mails).

Unfortunately people tend to use the same information for identification among different sites, because it is difficult to remember different usernames, passwords for every single site they use. Reuse of login information does not only enable linking of actions (simply by comparing usernames), but could also be very dangerous. If an attacker is able to gain access to one account (e.g. by breaking the password), he could easily log into another account created by the same user.

To solve this problem LPWA generates aliases for users, consisting of username, password (and email address if needed). When a user is asked to enter information, the user enters

predefined prefixes into the form (e.g. /u for username, /p for password), hence LPWA knows where to enter the automatically generated information. When this same user is visiting the site again, LWPA recognizes the site and login information can be entered the same way as during registration.

Features of LPWA include (cf. [Gab99]):

*Alias Generator:* LPWA generates automatically secure, consistent pseudonyms, which build different personae (partial identities) for different websites. Usually these identities consist of username and password. Sometimes email addresses are included too, when this information is requested by a website.

*Email Service:* the tool not only generates usernames and passwords, but email addresses too. When a website sends an email to the automatic generated address, the message is automatically forwarded to the users true email address.

*Anti-spam Support:* When an email is forwarded to the users true inbox, not only the address generated by LPWA appears as sender, but the address, from where the message originated, is stored in the mails CC-field (carbon copy[4]). Thus it is possible to block mails sent by this site in the future, in case spam was distributed. Despite a measure called *address spoofing*[5] people using LPWA are not only able to block spam, but can identify the site, which was responsible for the spam-mail in the first place (e.g. which site disclosed information to spam distributors). This is possible because LPWA-user are only known to different websites by the data generated by LPWA. Although spam distributors are still able to change (or spoof) originating addresses to avoid spam-filters, these emails can still be blocked, because the email can not only be identified by the originating email address, but also by addresses generated by LPWA. If a LPWA user wants to block mails sent to a certain email account generated by LPWA and forwarded to him, he has just to block mails sent to this account. This information can't be changed by spam-distributors, hence address spoofing is useless with LPWA.

*Filtering of HTTP-Headers:* sensitive information is removed from headers.

*Indirection:* every request made by the user is rerouted through a proxy - hence server are only able to see the proxies address, not the true address, where the request originated from.

*Statelessness:* LPWA doesn't store translation tables (associating data for relationships between pseudonyms/personae and websites) remotely. Hence LPWA can't be forced to disclose sensitive data stored on a central server - a potential target for hacker attacks.

After describing LPWAs features, the tools main components (cf. [Gab99]) will be introduced below:

4 to email addresses inserted into this field, a copy of sent message is issued. Unlike BCC (blind carbon copy) receiver-addresses are visible for all receivers

5 real address is concealed, so address may not be recognized by spam filters

*Persona Generator:* generates a persona by using the *janus function*[6]. LPWA needs a user ID (valid email address) and a secret serving as universal password. Additionally the generator uses the address of the target website to generate data of the partial identity. Given information is used to generate username and password. The email address for a particular persona is a combination of the website domain and a secret key. The generator itself can be integrated directly into the users browser (or on the Browsing Proxy - see below).

*Browsing Proxy:* The proxy used by LPWA is not only redirecting requests, but filters as well HTTP-headers. The proxy could be implemented remotely on a firewall, ISP access point etc.

*Email Forwarder:* forwards messages sent to generated addresses to the user's true address. The forwarder has to be placed somewhere remotely, away from the user, so numerous generated email addresses can't be linked to the user.

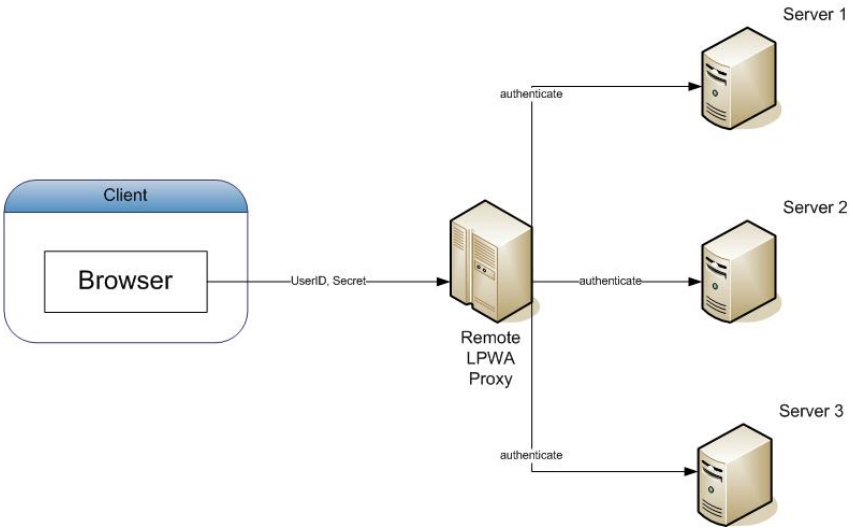


**Figure 4.7:** LPWA Proxy Configuration

In [Gab99] were numerous trade-offs identified, concerning LPWA:

*Trust:* LPWAs persona generator is provided with information, which is probably sensitive to the user. Therefore user have to trust the mechanisms used with LPWA and that the information is secure and isn't abused.

*Anonymity:* no components (Browsing Proxy, Email Forwarder) should make it possible for attackers to gain information about people using LPWA.

*Performance:* if proxies or other nodes, connections can't handle the amount of requests issued, a performance degradation can be expected. When [Gab99] was published

6 combination of cryptographic functions

in 1999, dial-up, not broadband connections, were widely used. ISPs may not have been issued with cables, servers to handle all requests properly made by customers.

When components of LPWA can be trusted and data is secure, users anonymity isn't violated. Servers are only able to see the proxy as originator of the request, not the user himself. Furthermore personae generated (usernames and passwords) can't be "decyphered" back to the users true identity. Because one persona isn't used twice actions, visits, accounts made/created by users remain unlinkable. Hence it is impossible to build profiles/dossiers of users according to their online habits. Although it could be possible to determine users identity by analyzing traffic from and to proxy servers used.

LWPA was introduced in June 1997 and 40 000 users were already attracted by this service, when in 1999 this article was published. The service was canceled after only a short amount of time. Lucent used the LWPA technology however to create ProxyMate, but the technology was sold to CMGI in 2000[7].

## 4.3 Encryption

Methods described above tried to prohibit unwanted information disclosure. *Anonymizers* enable sending of requests to webservers without giving away identification information like IP-address, used browser and operating system. *Pseudonymizers* enable at least some level of identification, without disclosure of a user's whole identity.

Before introducing PGP (Pretty Good Privacy, method for signing messages), a short overview of encryption (algorithms) will be given.

*Encryption* tries to maintain the privacy and integrity of data stored and the messages transmitted during transactions between transaction members. Encryption of messages during transmission prohibits eavesdropping and encryption of stored data makes it unreadable for hackers, unless they are able to decypher it.

The most important part of encryption algorithms is the *key*. It is used to encrypt the data/message and to decypher it. The length and the ingredients which were used for deriving the key, are mostly the measurements for security of a particular algorithm.

There are to different kinds of encryption methods, which are mainly differentiated by the use of keys:

*Symmetric Key:* Sender as well as receiver use the same key to encrypt and decypher the data. Symmetric keys are used for blockcyphers and streamcyphers.

When using a blockcypher algorithm, the plaintext (unencrypted text) is divided into blocks of the same size. Each block is then encrypted by using a key. Examples for blockcyphers are the Data Encryption Standard (DES) and the Advanced Data Encryption Standard (AES).

*Streamcyphers* on the other hand combine the plaintext bit by bit with the encryption key. An example for streamcypher is RC4.

*Hashfunctions* (also used by PGP, see below) produce a hash of fixed length, derived with the plaintext. With better hashfunctions it is quite difficult to find different

7 see `http://www.cs.tau.ac.il/~matias/lpwa.html`

inputs (plaintexts), which produce the same hash. Therefore hashfunctions belong to one-way encryptions. It is impossible to derive the plaintext from the hash.

One problem of symmetric key encryption is to find a secure way for distributing the keys that are used for both, encrypt and decrypt, the data.

*Public Key:* Unlike the symmetric method, where only one key is used, two keys are needed for public key encryption. The *private key* (mostly used for decryption) for a transaction is either known by sender or receiver, while the *public key* (mostly used for encryption) is known by both transactionmembers. Although these two keys are related to each other, it is impossible to derive one key after getting hold of the other. Public key algorithms like RSA can be used for digital signatures (authentication of messages).

### 4.3.1 PGP (Pretty Good Privacy)

The security system PGP was introduced in 1991 by Phil Zimmerman (cf. [Hen00]) and uses the public key method with public and private key (like described above). While the public key is used to encrypt messages and verify signatures, the private key is used to decrypt messages and creation of digital signatures.

The system of PGP seeks to ensure privacy of messages, meaning that only the owner of the data or the intended receiver of the message can decypher and read it. When a hacker manages to get hold of data or a transmitted message privacy is still intact, because the message is unreadable, until decyphered.

To not only ensure privacy but authentication of data and messages, the input is first passed trough a hash function. Then the derived hash is signed by PGP (rather than the whole message to be sent). Therefore it can be ensured that the message was indeed sent by the claimed sender and the message wasn't tempered with during transit. The signature is created with the private key of the sender, and the receiver is able to check the signature with the sender's public key.

Keys needed for encryption/decryption of messages are saved in a key ring file. This file should be kept at a safe place. Otherwise hackers could be able to get hold of private keys and decrypt messages or forge signatures. Alongside private keys the file not only holds the user's own public key, but public keys of other users.

[Hen00] includes a detailed description how PGP is set up on a (linux) system and how encryption, decryption and signing works with PGP. When installing PGP it is first necessary to crate the both keys, public and private. For the private key it is also necessary to choose a pass phrase (similar to a password), which must be entered whenever the private key is used. Then the public key needs to be published, in order to grant other users access to it. The key can be for example put on a personal website or the Public PGP Key Server. In order to ensure that a public key can be really trusted and that the message (or rather the message's sender) is indeed authentic, a key must be certified. This happens among other possibilities by using a model called *web of trust*.

### 4.3.2 Web of Trust

While the section above introduced PGP and how the model works, this section will give an short overview of the certification model for public keys used with PGP. So web of trust

tries to answer the question how user can be sure that a public key they use is indeed created by the user stated in the information associated with the key.

One way to verify the key (cf. [EXP04]) is of course to get in personal contact with the user by getting his number and calling him (if you don't know the user in person already). In this case however the voice of the to be identified user has to be known already to validate it.

To meet the user in person would solve at least some problems. The user could show his ID card to verify his identity along with his key properties.

The third concept, another person confirms that the key indeed belongs to the user stated, is used in web of trust.

Calling or meeting the user in question is often not easy to manage or even impossible, so the WOT is a welcomed alternative. With WOT people can benefit from the fact, that a key was already verified/signed by other users. When a user confirms/signs a public key, he sends that confirmation to the PGP key server, so other users have access to it.

Hence a key is valid, when it has been confirmed by enough users/key holders, which are part of the users trust ring (consisting of users, who the user trusts). The number of confirmations needed to validate a key is determined by the so-called *trust-flag*. Values for this flag range from (in PGP Version 5.x) untrusted (certificates done by this key are ignored) to ultimate (user owns the secret key associated with the public key –> key signed by a ultimate-flagged key becomes valid).

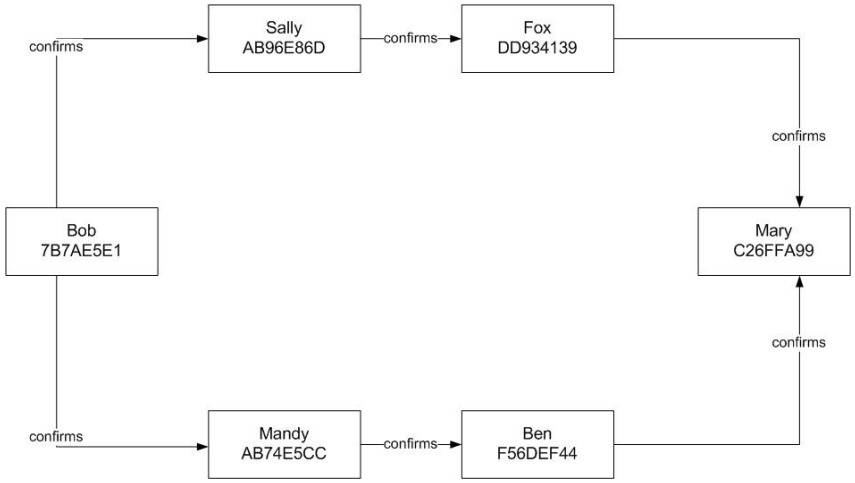


**Figure 4.8:** Web of Trust Principle

## 4.4 Negotiation and Trust Engines

When using the Internet people disclose data with every single request sent to servers. Information like IP-address, ISP, used browser and operating system is included in the data received by webservers. One method to prohibit this kind of information disclosure involves the use of anonymizers (e.g. mixed nets, CROWDS: see sections above). Most webservers and web services however need at least some level of identification. This is usually achieved by use of cookies, pseudonyms to "recognize" people, who revisit websites.

To be able to use some services provided, other, probably more private, information is required. Data asked for by webserver could range from email addresses to enable receive of newsletters, zip codes for statistical purposes or customization of services, to disclosure of complete mailing addresses, bank account data or credit card numbers.

According to privacy bills passed by nations all around the world, web servers need to clearly state which data is collected and how disclosed user data is handled, stored or if it is divulged to other companies. Privacy practices of a website are stated in the so-called *privacy policy* and links to these terms to handle private data are usually located at the bottom of a website's mainpage. Unfortunately these terms are often written in a way that is not easy to understand by most people, even if the text is written in a language that can be considered as the particular user's mother-tongue. These texts are often written by people who are educated in law and these texts often turn out complicated and not easy to understand by normal people. The use of legal terms, with which normal users are not familiar with and the often huge amount of paragraphs included in the policy, make it even more difficult for users. Hence despite it is clearly stated on websites how private data is treated, people often don't understand what these terms actually signify.

Because people may not clearly understand the privacy policy stated on a webpage and hence don't really know how their disclosed private information will be treated, people may not find the site trustworthy and may decide to don't use the service at all. People may be even afraid of going online, because of information disclosure, that could occur without their knowledge. Complicated written policies are a main reason why some people are still afraid of using online services. To strengthen trust again in online services, negotiation engines like P3P (Platform for Privacy Preferences) try to make things easier for people. Privacy policies are offered in a clearly standardized format, so they can be read by machines. User agents automatically parse privacy policies disclosed by web services and compare them to user privacy preferences. If website policies violate privacy preferences stated by the user, the user is automatically alerted. User agents for P3P protocol are now included in the newest versions of wide-spread webbrowsers (e.g. Internet Explorer, Firefox, ...). Therefore no separately installed user agent is needed anymore to make P3P work.

Below are P3P and TRUSTe (seal program to enable legal consequences for websites/companies that violate their own privacy policies) described in more detail:

### 4.4.1 P3P (Platform for Privacy Preferences)

To raise trust of people in online services (again), P3P tries to assist during communication/transactions with online webservices and websites. Standardized privacy policies are readable automatically by machine (user agents like web browsers, plugins, proxy server) and parsed, so that users are alerted when policies violate privacy preferences stated by the user. The file (proposal) includes the identity of the organization/company communicating with and of course its privacy practices (cf. [Rea99]).

Proposals can be accepted by user agents automatically in case they do not violate privacy preferences. When accepting, the user agent sends a *propID*. Then when a website is visited again, no new negotiation is needed, sending the propID from previously reached agreement suffices. If the proposal violates privacy, the user agent can alert the user, reject the proposal or ask the service to send a alternative proposal. Because P3P supports

offering of multiple proposals, the system is more flexible. If one proposals is rejected, another proposal can be offered and it is more likely that *negotiations* of privacy policies between service and user (agent) succeed. Hence the probability is higher that the service is actually used. So P3P could help in strengthen/restore people's trust in online services.

In addition of offering a standard for privacy policies, P3P implementation could also offer data repositories, where people can store data, that they are willing to disclose. So data can be disclosed automatically, if a agreement is reached. Services are able to store data in the repository too, if the reached privacy agreement allows such action. Apart from private data, pseudonyms (identifiers) can be saved for websites, a agreement has been reached.

Because P3P forms just the protocol for transmitting standardized privacy policies, it can't ensure that companies and websites act accordingly to their stated privacy policies. In case privacy violation occurs a organization like *TRUSTe* (see below) comes into play, which is able to take legal actions against the service/company.

With the protocol and semantics (using XML and RDF) offered by P3P it is possible to describe information practices and data elements in a way, that enables machines to parse them automatically and compare it to privacy preferences set by users. Whereby the practices are described by a standard set of vocabulary and the information collected by base data elements. Practices as well as information are merged by XML into the P3P policy. Policies usually don't contain as much information as human readable policies, because they are often the result of several answers to multiple-choice questions (cf. [Cra02]). These preferences (consisting of many different options), which the sent policy are compared against, have not to be set manually. People can use configuration files provided or use recommended settings to get started, when they are not so familiar with P3P yet.

Back in 1999 P3P tried to provide an alternative to cookies, which aren't transparent to users. Nowadays however it is much easier to block certain cookies by changing preferences in webborwsers used to go online. P3P alternative to cookies includes two different IDs:

*PUID:* the pairwise unique ID is not only unique for every site, with which an agreement was reached, but also for every single agreement.

*TUID:* the temporary unique ID is only used for maintaining state during a single session. A new TUID is generated for every single session.

The data repository (already mentioned above) provided by P3P should ensure faster transactions. User agents should be able to release data automatically, if privacy preferences aren't violated. Additionally users have not to reenter information they had already disclosed during a previous visit, or retype information of the same kind at different web services. Hence it is also ensured that services receive consistent data from users. When able to retrieve user data from data repository, services may not need to store needed data in huge databases. Hence user privacy is better protected (if the repository is implemented in a secure way). The information stored is structured by data elements possessing a name and a datatype, known by every user agent.

P3P only provides basic standards and leaves many options for the implementation of user agents (e.g. design of GUI, additional data elements in repository, ...).

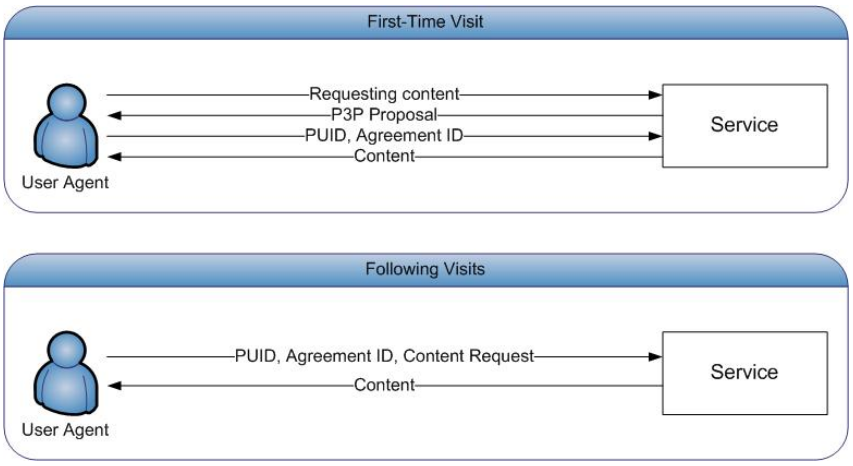


**Figure 4.9:** P3P Interaction between User Agent and Service

### P3P Specification

The specification provides the framework for privacy practices and information sent between service provider and user agent. It also includes a protocol for transmitting the policies. A list of vocabulary provided by the specification (1.0) is shown below (cf. [Cra02]):

*Entity:* Information on company/owner of the site/service

*Access:* People are able to determine what personal data is collected/kept about them

*Disputes:* How privacy-related disputes can be solved (e.g. privacy seals like TRUSTe)

*Data:* What kind of data is collected?

*Purpose:* Purpose of data collection. How is collected data used?

*Recipient:* Under which conditions collected data can be shared?

*Retention:* Defines policies for periodic purging of data

*Consequence:* human-readable description of data practices

## 4.4.2 TRUSTe

While P3P tried to raise trust in online transaction by providing easier access to privacy policies, organizations like TRUSTe (`www.truste.org`) ensure that companies don't violate provided privacy policies. In case a violation occurs, legal actions are taken.

In order to get a license issued by TRUSTe service providers and websites have to fulfill the following requirements (cf. [Ben99]), they are similar to the *Fair Information Practices* mentioned in ch. 2.

*Notice:* a link has to be provided on the homepage that includes the websites privacy statement. TRUSTe also provides support during creation of these statement to ensure that they are easy to read and understand.

*Choice:* the website has to at least provide an option to opt-out, to prohibit disclosure of sensible information to third parties.

*Security:* websites have to implement sufficient security measures to protect stored data from loss, misuse or unauthorized alteration.

*Data Quality and Access:* people should have access to the data stored about them, in order to correct errors or update the information.

*Verification and Oversight:* TRUSTe ensures that licensed websites follow the stated privacy practices by initial and periodic reviews.

Websites which a TRUSTe license is issued to display a clearly visible sign to ensure users and visitors, that fair privacy practices are used.


**Figure 4.10:** TRUSTe Seal for Web Privacy

TRUSTe provides several seals with different requirements, guarantees and seals for display (see `www.truste.org`), e.g. web privacy, email privacy. The childeren's Privacy Seal on the other hand additionally provides education for teacher and parents how to protect children surfing online.

## 4.5 Identity Management

Modern technologies made it much easier to collect and store data about people (in huge databases) for later use. Collection of sometimes private and sensitive information is often done without people's consent and knowledge. Many people are not even aware of the huge amount of data that is collected and stored. For example supermarkets in the US are allowed not only to monitor, save, but even sell information on purchase habits of their customers. Within the EU this kind of monitoring can't be done without the customer's consent. This needed consent however is given by entering loyalty programs, supermarkets start for that particular purpose (cf. [Cam05]).

Profiling of people's actions is possible by linking several different actions back to one particular user. Some technologies and measures used for linking of actions were already described in chapters and sections above.

To find a solution to this problem identity management architectures/tools are in development, which should help people to manage/use their "partial identities"/pseudonyms. These tools often combine techniques described above (anonymity, pseudonymity, encryption, ...).

Within the European Union a legal framework was provided to protect people private information (even after the changes 9/11 triggered). To accompany this framework a project in development called *PRIME* (Privacy and Identity Management for Europe). Objectives of this project are among others (cf. [Cam05]):

- limitation of private information disclosed. Only information needed for a particular information is disclosed, nothing more.
- actors in transactions can negotiate on (privacy) policies used during transactions.
- implement mechanisms to manage private information (for different purposes like pseudonyms/partial identities).

The PRIME-project will be described in detail later on. Tools like these are needed, because users often have no other choice, but disclose information, they not really want to give away. Otherwise they are not able to use a provided service. To make it easier for people and service providers, limitation of disclosed information on needed information only, is implemented. So people may be not afraid anymore to give away information, because additional (not needed) information isn't disclosed.

Private data is often stored at the service provider and if a security breach occurs and data is stolen, rather the customer is going to feel the consequences (e.g. identity theft) than the service provider (SP). One of the consequences for SPs may be damage of image and loss of customers (if there is a safe alternative customer can run off to). So as long as SPs are not really feel the consequences of security breaches, there will be no need for them to install identity management systems, which would not only raise security standards used, but usability for customers as well as service providers.

### 4.5.1 Identity Management Models

Below are identity management models listed (existing and in development) as described in [Cam05].

#### Existing Identity Management Models

Models are described concerning usability and scalability from user-perspective.

*Silo Model:* Users are identified by authentication token pairs (e.g. username and password) at the SP. These authentication tokens alongside with name spaces are managed by SP. Users are allowed to define their own identifiers (e.g. manually chosen pseudonyms), as long as they are unique within the name space (e.g. already registered usernames). Authentication tokens (e.g. password) can be chosen by users too. The model is easy to deploy and private data is only disclosed to the SP. On the other hand the user has to remember all the identities/pseudonyms he uses for communicating with different SPs. Hence usability of this model is rather low and could endanger security, because people could be tempted to use similar or identical usernames (and even passwords). Therefore unlinkability can probably no longer be guaranteed. Forgetting of passwords poses a problem for users as well as SPs. User have to request retrieve of the old password or creation of a new password. In cases were a high security level is required, these requests could turn out to be very expensive for the SP.

Identity Management Models like these are often used for online boards, discussion groups, etc. which require registration for publishing posts.

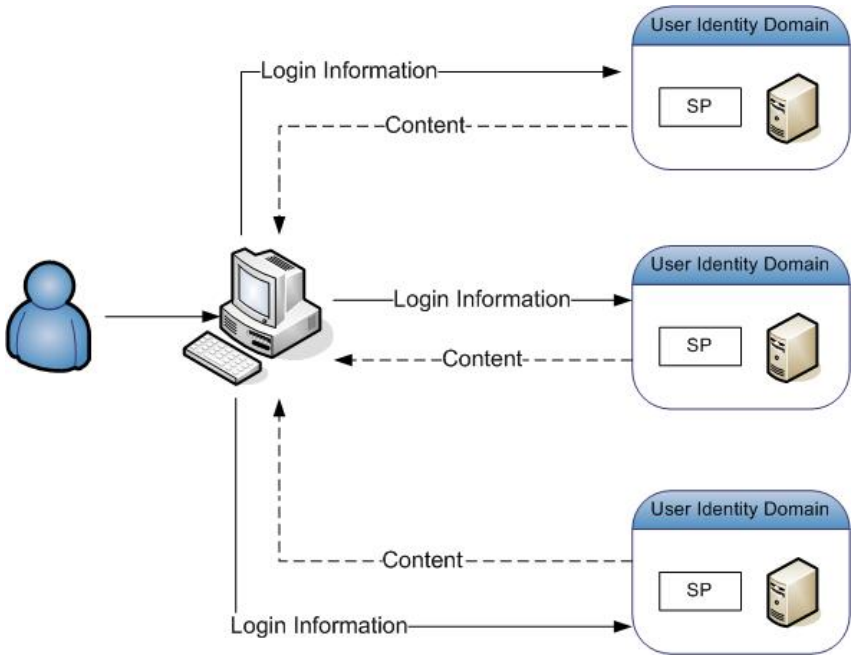


**Figure 4.11:** Silo Model Architecture

*Common Identity Domain Model:* In this model identifiers and authentication tokens are managed by a central authority, but users are still identified by the SP itself. The identifier is still unique within the (common) namespace (e.g. database of central authority) and the authentication tokens usually consist of some public certificate, issued by this central authority. So identities are simple to manage for user as well as SPs. The user has only to remember one token pair (managed by one central authority), but it is impossible to define a unique namespace for every user in the world. Social security numbers are defined differently in different countries or even aren't unique for one nation. Email addresses are unique, but most people have more than one address registered in use. Hence there will always be numerous authorities in place. Furthermore one identifier enables several SPs with connection to central authority to identify one particular user with one and the same identifier. Hence SPs are able to match personal information about one user known to them, because of the unique identifier, common among the SPs connected to the authority.

*Centralized SSO Identity Model:* Similar to the model introduced above authentication tokens and namespaces are managed by a central authority (CA). Unlike the Common Identity Domain user are also authenticated by the CA. The SP just receives a security assertion. This model provides high usability through single sign on (SSO) and works well within closed networks (e.g. in companies, universities, government agencies, ...). Users have to be authenticated once by the CA in order to use services connected to the CA. In open environments this model works not that well, because it can't be expected that many different SPs all around the world (or within one nation) work under the same policies (as issued by the CA). Once a user is authenticated he can be identified by every single SP within the CA-domain. This violates limitation

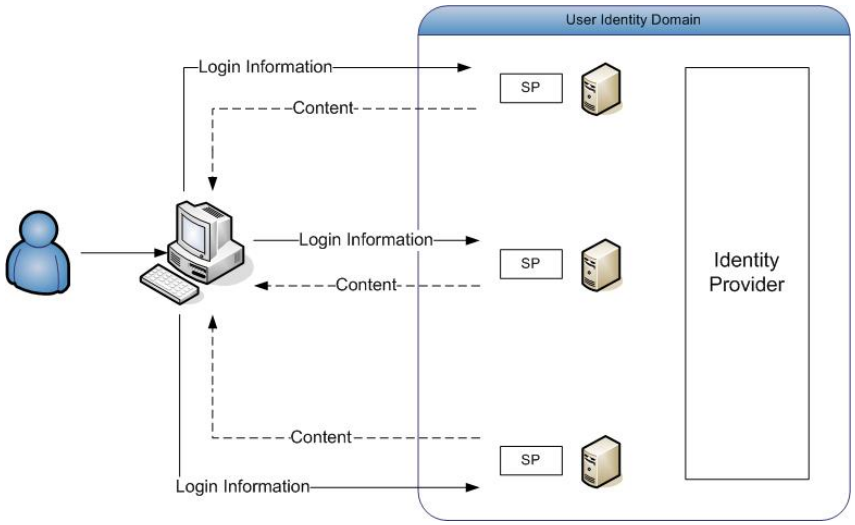


**Figure 4.12:** Common User Identity Domain

of information disclosure, because every SP is able to link information (stored by other SPs) back to the user.

In 1999 Microsoft tried to introduce a global SSO architecture named *Passport* (name changed into *Windows Live ID* back in 2006). Peoples (private) information should have been stored at one central location, so users have just to log in once, in order to use services. Naturally this service wasn't well accepted, because people were afraid of the fact, that one single authority holds all their private and sensitive information. Nowadays Windows Live ID is used for authentication in several services provided by Microsoft, like Hotmail EMail account or Windows Live Messenger.

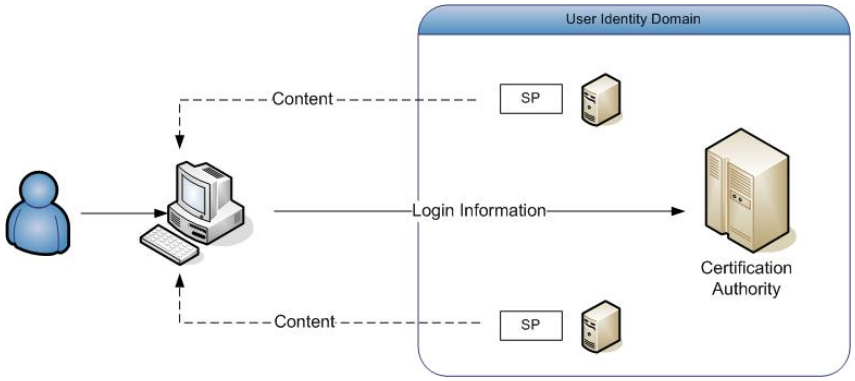


**Figure 4.13:** Centralized SSO Identity Model

*Centralized Model with Browser Support:* Because Passport wasn't well accepted by the public, Microsoft introduced a network and software architecture called *Card Space* (formerly known as *Info Card*). Unlike Windows Live ID, Card Space uses several CAs instead of just one single authority, which authenticates users. Hence Windows Live ID is just one of many different CAs.

One *InfoCard* represents one identity, which with a user is known by a SP. Data needed for identification isn't stored on a local machine (or within the InfoCard), instead each InfoCard contains a pointer with points to the CA, where the information is stored. CA is sending the security assertions to the CardSpace modules, located at the client, not directly to the SP. Hence there is no direct communication between SP and CA. Despite usage of several CAs this model remains a centralized architecture, which could pose a threat to peoples privacy, in case of abuse. So SPs still resist to accept that third parities authenticate users, who want to use provided services. Because several CAs are used and no real common domain is implemented, users have to be authenticated at each single CA. Although private information isn't stored on local clients, pointers to this information are saved on local harddrives. These markers to sensitive information can be abducted by hackers, similar to passwords. So attackers are able to use identities represented by stolen InfoCards, even without the users knowledge. Hence this kind of architecture has still huge security problems.

CardSpace is part of the Microsoft .NET Framework and is distributed with Microsoft Windows Vista (for Windows XP it can be installed separately). With this software users are able to use issued InfoCards or create new ones themselves for different situations, websites.

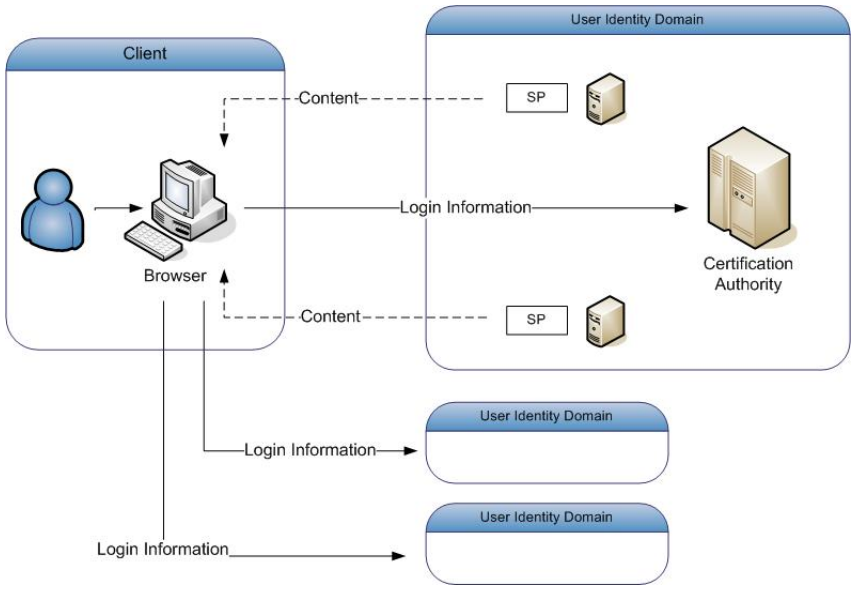


**Figure 4.14:** Centralized Model with Browser Support

*Federated SSO Identity Model:* Several SPs agree on common policies for security, authorization and create a so-called *circle of trust* to enable SSO. Each federation defines agreements, standards and technologies in use, so SPs are enabled to identify users from other SPs. Each SP manages users from its namespace and namespaces from different SPs are linked to form a *federated domain*. Identifiers unique for each namespace are mapped, so SPs are able to refer to one particular user, despite he is identified by several different identifiers among the namespaces and SPs are accepting security assertions issued by another SP within the cycle of trust.

This model is also suitable for open environments and is compatible with existing silo models - one SP with silo model can easily join a cycle of trust. It is however technically complex and SPs are not able to distinguish between requests made by users and other SPs. Furthermore SPs have to trust security assertion issued by other SPs and privacy of user data could be threatened because information stored by different SPs could be linked. On the other hand the user's true identity is only known by one SP, the one which conducted authentication. To all other SPs the user is only known by his identifier (pseudonym).

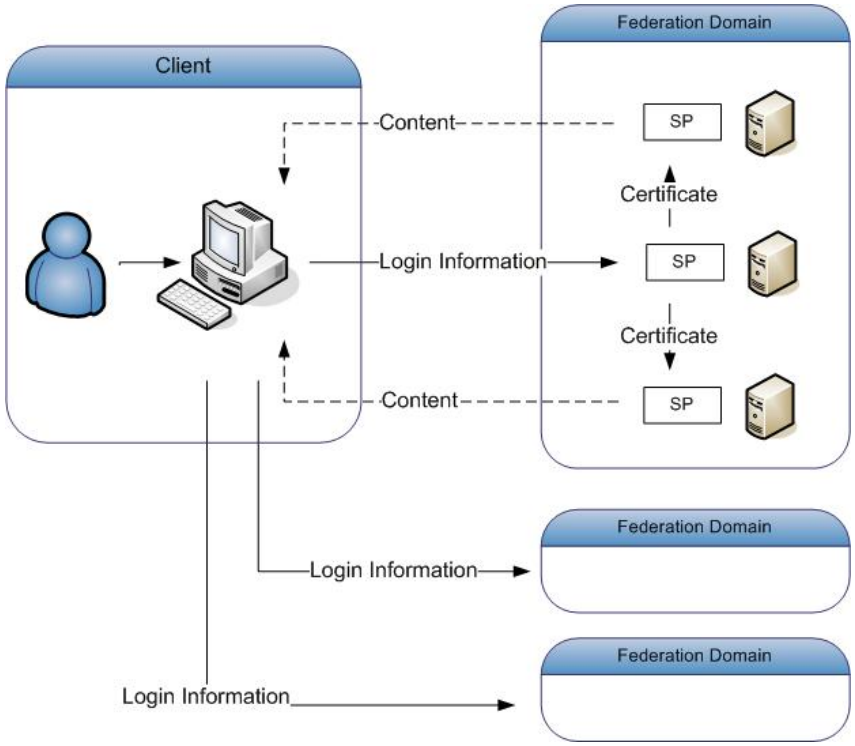


**Figure 4.15:** Federated SSO Model

### User-Centric Identity Management Models

In this section a approach for management of user identities will be introduced, developed by authors of [Cam05]. This approach should provide better security as the models described above by preventing session hijacking. The model includes usage of PADs (Personal Authentication Device), e.g. mobile phones, PDAs, for authentication of messages sent during transactions. But most important the PAD itself has to be secure, otherwise security of the model would be compromised again. Hence constant increasing number of features mobile phones, PDAs and similar devices offer, could offer increase in flexibility, but could also pose a threat to security, because of increasing number of system vulnerabilities.

*SMS Authorization Codes:* With introduction of online-banking keyloggers and similar tools (e.g. phishing) posed an even bigger threat, because hackers were able to determine not only account numbers, but login data. So hackers were able to

enter user's bank accounts and transfer (steal) money. To counter that threat bank institutes introduced so-called TAN-lists (list of transaction-numbers) which are issued to customers by mail. To conclude a transaction user have to enter exactly the TAN number requested (e.g. if the TAN number 5 is requested, users must enter the number listed in row 5). All security is solely based on this list, so it is important to keep the list in a safe place. Hackers are still able to change data like destination or amount without the customers knowledge.

To eliminate the threat of stolen lists some bank institutes already use SMS for message authentication and in this case the user's mobile phone works as the PAD. A TAN-list isn't necessary anymore, because all authentication is conducted over mobile network. When a customer wants to conclude a transaction, the bank institute is sending a SMS to the customers mobile phone. This SMS contains the account number (and name), destination, where the money is going to be sent to. Sometimes the amount of money to be sent is displayed too. The most important part of this message however is the authentication code consisting of several digits. The user has to enter the code into the bank accounts GUI to confirm that the information (receiver, amount of money, ...) is correct. Hackers may be able to change amount and destination of transaction (e.g. by trojan horses installed on the client machine), but this changed data will appear in the SMS and may be noticed by customers. So here security is based on the fact, that it is much more difficult for hackers to break into mobile phone networks and alter sent messages.

However authorization codes may be to complicated for secure manual copying from mobile phone display to bank account GUI or people simply fail to notice changes in data.

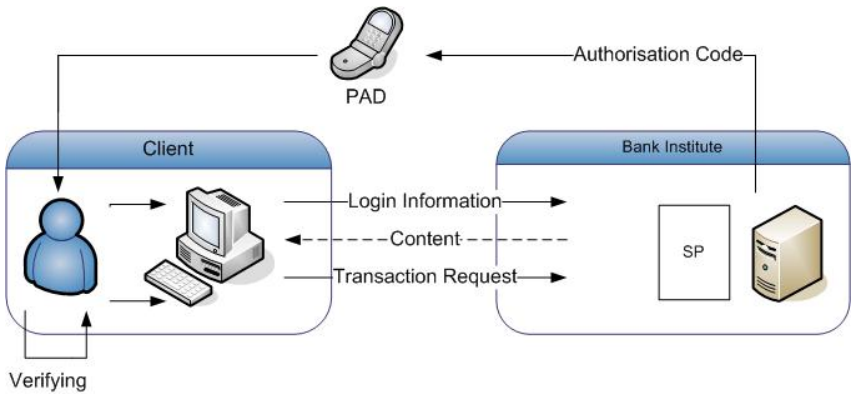


**Figure 4.16:** SMS Authorisation

*User-Centric Silo Model:* In this model the user has only to log in at the PAD once. Because the PAD is supposed to be used only by one user, identifiers aren't needed anymore. After successful login at the PAD, the user is able to access any service he wants, while the PAD is handling the login at the correspondent SPs. The PAD can also be configured to store and manage all identifiers and authentication tokens.

Level of privacy protection is similar to the level of silo model described above. Private information is only exposed to SPs users are communicating with (unlike

domain-models, where information or pseudonyms could be shared among the domain). Furthermore the user-centric silo-model can be combined with any model described above.

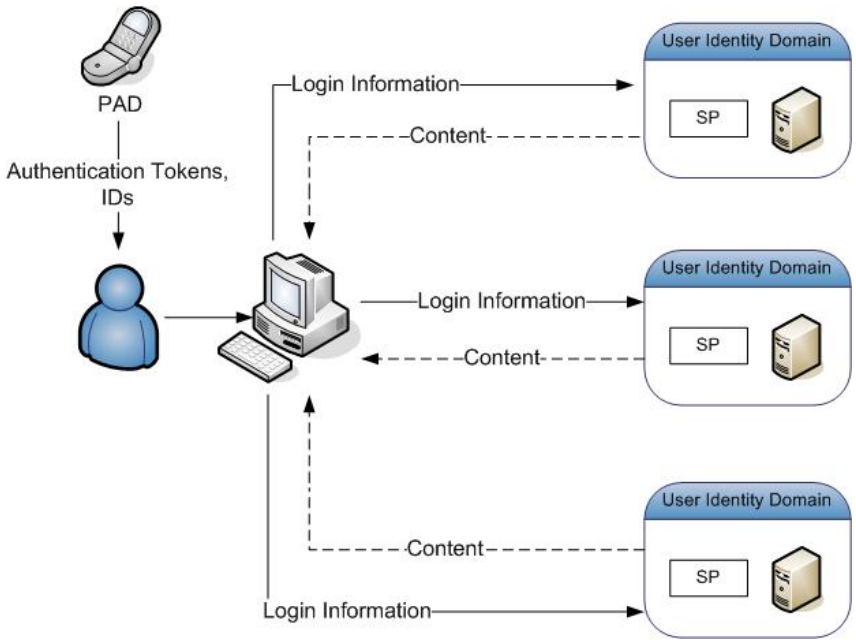


**Figure 4.17:** User-Centric Silo Model

### SP Identity Management Models

This section will focus on models to manage different SP identities and how different users can authenticate to the same SP.

*Common SP Identifier Domain:* This model focuses on the way a remote server is authenticated at the client. Usually server send a *server certificate*, that was issued by some *certificate authority* (CA). Web browser like Microsoft Internet Explorer receive that certificates and check whether they are valid or not. In case the received certificate is valid, connection to the server is established. Otherwise a warning is displayed and its up to the user to decide, whether he wants to proceed and establish the connection.

*User-Centric Management of SP Identities:* The Common SP Identification seen above has certain flaws in security. Like mentioned above web browser only check whether the certificates received are valid. This precaution doesn't guarantee that the certificated server is indeed the server the user had intended to connect with. This flaw in security is exploited by *phishing*.

People are lured to sites, which at first glance look like their official counterparts. The sites however are created by hackers to make people disclose sensitive data, like account numbers, passwords etc. The web browser establishes connection, because of the valid certificate received.

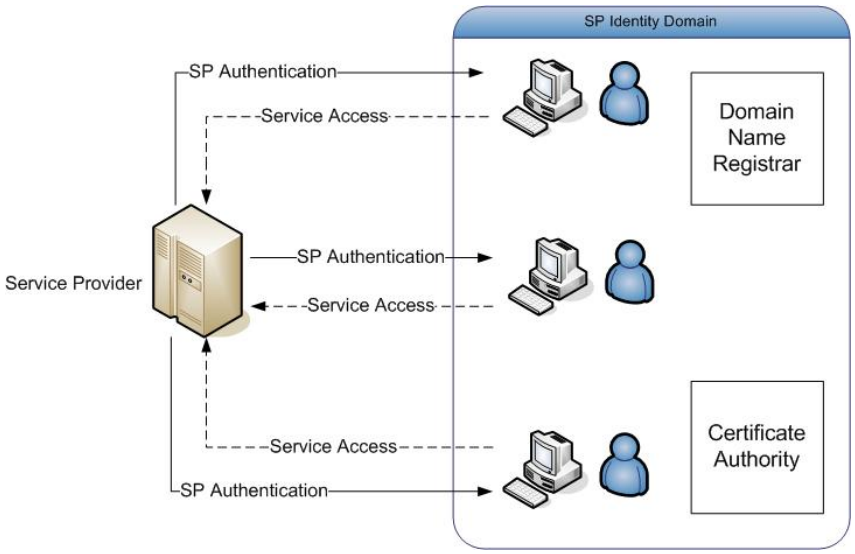


**Figure 4.18:** Common SP Identifier Domain

With user-centric management of SP identities user are able to map identifiers for SPs. For example they can map some kind of image to a certain domain name, for better recognition. To make this model work a PAD has to be used, for user-side confirmation (similar to the solution presented in *SMS Authentication Code*). When the web browser wants to connect to the server (to request a service), the user-chosen identifier (e.g. image, sound, ...) is sent to the PDA, where it is displayed for confirmation by the user. When confirmation was sent by the user, connection is established.

Tools like *TrustBar*[8] for Mozilla's FireFox try to integrate this kind of model directly into the browser. Like described above, people can map identifiers to certain servers. When the browser connects to the server, the associated identifier is displayed in the GUI. Hence users are able to confirm, whether this server is indeed the server they wanted to be connected to in the first place.

Microsoft on the other hand tried another approach and included the so-called *phishing filter* in its newest versions of Internet Explorer. The filter sends certain websites to Microsoft servers, in order to check, whether they are known phishing sites.

The newest version of FireFox (Version 3) tries to protect user from phishing sites in a similar way. If the site visited is secure (not a imposer used for phishing) the URL address bar is highlighted in green. Therefore people are easily able to confirm, whether a site is secure or not. Additionally the certificate is shown in the address bar and where the server was verified. So FireFox 3 provides similar features as *TrustBar* for protecting user from phishing attacks.

8 `www.cs.biu.ac.il/~herzbea//TrustBar/` , unfortunately the tool isn't compatible with FireFox versions 2 and 3

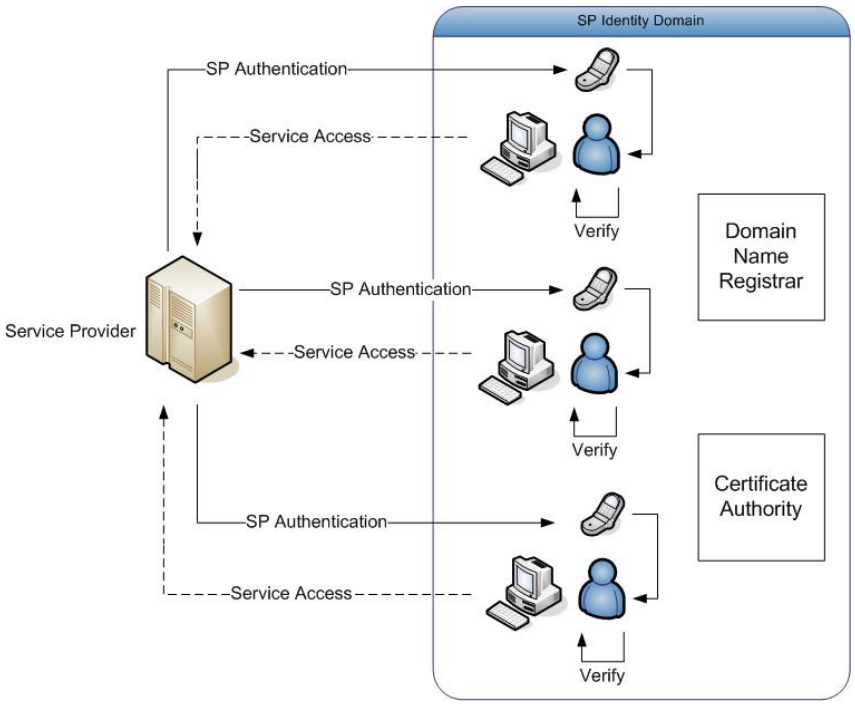


**Figure 4.19:** User-Centric Management of SP Identities

#### User and SP Identity Management combined

Authors of [Jø07] provide a model which reduces the amount of parties involved during transactions. Hence private and sensitive information is disclosed to a minimum of parties. This is achieved by using a PAD (e.g. a mobile phone).

The model presented is based on the user-centric silo identity model and the user-centric management of SP identities (see above). Individual steps (steps taken by members involved in the model, e.g. SP-Server, client machine, user, PDA) of the mutual authentication scenario are listed below.

The table (see below) shows the properties of presented models, concerning usability, privacy and security.

So to protect privacy usability of a model in use should be taken into account. If people have difficulties to use an application and change usage to a way that was unindented by the authors, security of private data could be endangered again.

The next section will focus on the most prominent project for management of private data and identities.

### 4.5.2 PRIME (Privacy and Identity Management for Europe)

As mentioned at the beginning of this section PRIME is a project partly funded by the European Union (cf. [RL08]) that intends to

- limit information disclosed to exactly the amount of data needed for a particular transaction

| Msg. Nr. | Message |
|---|---|
| 1 | Specify Service Request |
| 2 | Transmit Service Request |
| 3 | Transmit Server Certificate |
| 4 | Establish SSL[9] Connection |
| 5 | Transmit login page |
| 6 | present login page |
| 7 | Transmit Server certificate to PAD |
| 8 | Verify server certificate |
| 9 | Display mapped certificate info (e.g. play mapped sound) |
| 10 | Verify correctness of personalized info |
| 11 | Accept SP |
| 12 | Transmit identifier and authentication token (PAD –> Client) |
| 13 | Transmit identifier and authentication token (Client –> Server) |
| 14 | Verify identifier and authentication token |
| 15 | Submit service page |
| 16 | Present service page |

**Table 4.1:** Messages in mutual authentication scenario

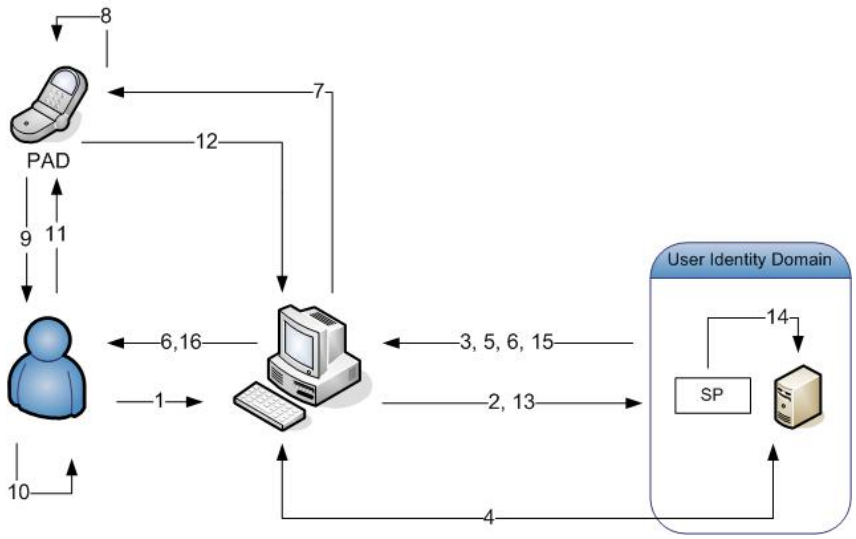


**Figure 4.20:** Manual Authentication Scenario with PAD

| Category | Name | Usability | Privacy | Security |
|---|---|---|---|---|
| Existing Models | Silo Model | low | high | NA |
| | Common Identity Domain | high | middle | NA |
| | Centr. SSO Id. Man. | high | middle | NA |
| | Centr. Model w. Brow. Supp. | middle | middle | low |
| | Federated SSO Model | high | middle | NA |
| User Centric IMM | SMS Auth. Codes | middle | high | high |
| | UC Silo Model | high | high | high |
| SP Identity Models | Common SP Identity M. | high | NA | low |
| | User-centric Man. of SP Id. | high | NA | high |
| UC/SP Mod. comb. | Mutual Auth. | middle | NA | high |

**Table 4.2:** Properties of Identity Management Models

- provide a opportunity for actors involved in transactions to negotiate privacy policies, which determine handling of (private) data (including possible further usage and security precautions).
- implement mechanisms to manage personal data.

People should be able to *use information services in a reliable and trustworthy way while keeping control over details of their private life* (cf. [Cam05]). PRIME's objectives are described in more detail below:

*User Informed Consent and Control:* Users are able to control which information is given to whom for which purpose and can vies the privacy policies negotiated with transaction partners.

*Privacy Negotiation:* Every time private data is disclosed, users can negotiate privacy policies with transaction partners and come to an agreement, to express how this data should be handled. These agreements are legal contracts, that have to be fulfilled.

*Data Minimization:* Transaction partners only collect data that is needed for the particular transaction. When collected data is no longer needed to fulfill the stated purpose, the data should be deleted. Credentials provided often disclose more data than needed for usage of service, e.g. driver's license used for age confirmation also discloses data like full name and address. Data that probably may not be needed by the service to work properly. PRIME provides a solution to this problem.

*Identity Management:* With modern technologies it is even easier to collect huge amounts of information about people (even without their knowledge). So users are able to release more or less private information depending on the trustworthiness of the transaction partner. Additionally PRIME lets users to monitor, record and manage information, which was released to different SPs. Identity management is provided to users as well as SPs, which must authenticate users, who want to use their services. With PRIME users can stay anonymous, but can still be authenticated to use services provided.

*Spectrum of Anonymity:* PRIME not only provides one single level of anonymity, but several levels. Depending on transaction type or negotiated policies, user can stay anonymous, use a pseudonym or be partly/fully identified to the service. By using of cryptography PRIME can provide a huge range of possibilities to authenticate a user without revealing too much information to the SP.

*Accountability:* Stay anonymous while surfing online is quite tempting. People are easily mistaken that they are able to do everything without heaving to fear consequences. Hence PRIME provides accountability (revoke of anonymity) in case of fraud or abuse. This is possible through *verifiable encryption*, where a third-party authentication authority can release (decrypt) information to the SP, with which the user in question can be identified.

Key components of the PRIME project, including parties involved during transactions, cryptographic tools used and the system architecture, will be described below. During a transaction performed with PRIME the following actors are involved:

*User:* To users are certificates (for authentication), data and policies (for handling of sensitive data) associated. Access control policies for example determine which data should be released during transactions.

*Service Provider:* This party provides services (to users). During transactions a SP may possibly have access to certificates, sensitive data associated to users and policies which determine handling of these information.

*Certification Authority:* A CA is a special type of SP, which issues *certificates* to users as well as SPs. With certificates the CA vouches for the truthfulness of the information given. Hence the CA has to be trustworthy itself, otherwise security would be compromised.

Within the PIME project numerous *cryptographic tools* are used. During secure communications for example, all transactions are performed over an encrypted, semi-anonymous[10] channel. Server is simply authenticated to the user by a standard certificate, while the user stays unauthenticated. Usually user disclose information like IP and MAC address when connecting to a server. To prevent disclosure of such information, user need to use anonymizing tools like *onion routing* or *mix nets*. Use of *pseudonyms* provides another level of privacy, by using different identities (with different parts of private information associated) for transactions.

The PRIME project uses so-called *credentials* to provide secure authentication, without revealing too much private information. A credential (also known as certificate or attribute certificate) is a piece of information that was verified by a third party and therefore can be used for authentication. These credentials have to be bound to the user by cryptography (e.g. use of secret key) to restrict access, ensure accountability and prevent sharing of data.

10 users can be identified, if necessary

Use of traditional certificates however can threaten privacy, because unlinkability isn't ensured anymore. CAs are easily able to link different uses of one and the same certificate. With *private certificates* on the other hand only information is disclosed that is needed for a particular transaction. These credentials are encrypted with a key provided by a third party, to ensure accountability. Additionally the user can bind a condition to the encryption, in which case the information could be decrypted and his identity be revealed.

### System Architecture

The paragraph below focuses on the *system architecture* of PRIME, which also includes the involved components. In general PRIME uses the same architecture for users as well as service providers. It makes use of the *Resource Referencing Scheme*, where pointers to information are used, rather than the actual data. To describe this information languages like RDF[11] and OWL[12] are used to enable *reasoning*[13].

The main component of PRIME is the database, which stores all certificates, policies, logs (record past transaction with parties in order to decide about further disclosure of information) and uncertified data. Access to the database is controlled by several components described below:

*Access Control:* controls and restricts access to data and enforces policies. A request to the AC usually consists of the URI identifying the resource to be accessed, purpose, operation to be performed and potential associated certificates and declarations. The AC responses either with "deny" including a list of conditions, which have to be fulfilled in order to gain access, or "grant" which also includes the requested resource. The decision whether or not to grant access is based on the declarations and certificates sent with the request.

*Identity Control:* manages interactive protocols with other parties (e.g. negotiations over privacy policies). It handles all requests to the AC, computes optimal ways to handle a request and manages user input over GUI.

*Obligation Manager:* A obligation describes a action that is triggered by a condition and is associated to data. When data is stored to the database, the obligation is activated. Each obligation is stored and managed by the OM. When a condition is fulfilled, the workflow described by the obligation, is triggered by the OM (e.g. data is deleted after some time or a certain date).

### Sample Transaction

In general a transaction involving PRIME consists of two phases (cf. [Cam05]): The *negotiation phase*, during which the policies are negotiated, used during the actual transaction, and contract execution phase.

11 Resource Description Language, to enable automatic parsing of (meta-) data provided per WWW. It consists of three parts: the subject is related to the object and the relationship is described by the predicate

12 Web Ontology Language, to create ontologies, hierarchically ordered information

13 enable/allow the system to make "decisions" on its own

*Negotiation Phase:* During the first phase the user sends a request, which the server receives and forwards to the AC. The AC responds with an offer (which also specifies how data is treated during transaction) and a list of requirements (credentials needed for successful transaction). This offer is received by the users IC, which forwards the requirements to the AC, and the AC decides which information is disclosed. The possible choices to fulfill the requirements specified by the SP are presented to the user through the GUI. Finally the user has to confirm the decision and the the SP in turn can accept or reject the offer sent by the user.

*Contract Execution Phase:* Contract execution phase starts with the SP sending the needed credentials to the user. With this credentials the user's IC is able to access the necessary information and sends it back to the SP. The SP's IC in turn decides, whether the received data satisfies the contract negotiated earlier. If contract is fulfilled the AC is requested to store the data according to the policies and save the obligations associated at the OM. The OM then activates these obligations as well as triggers any business process related to the ongoing transaction and handles triggered obligations.

For more complicated transactions however it may be necessary to perform several of two-phase sub-transactions (cf. [Cam05]).

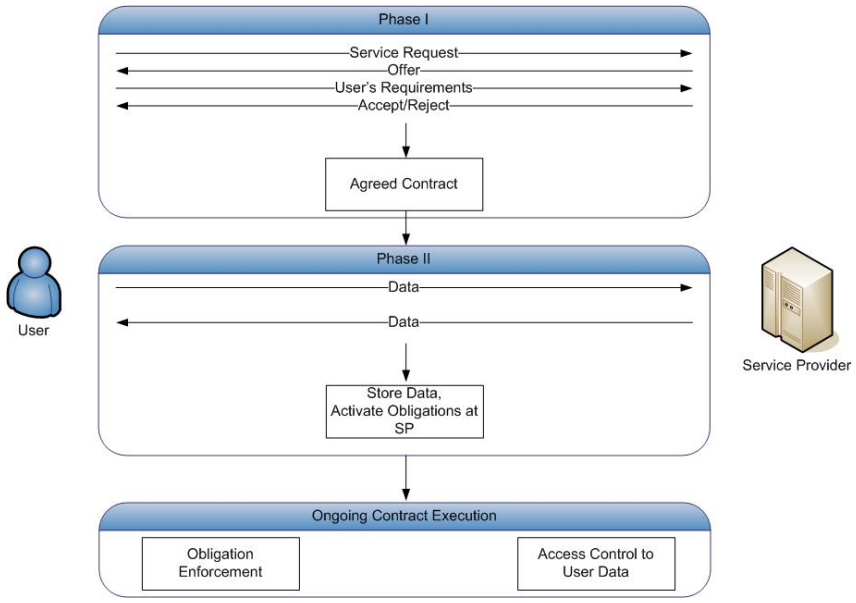


**Figure 4.21:** PRIME Transaction

### 4.5.3 Liberty Alliance Project

To date 150 organizations are part of the project (`http://www.projectliberty.org/`) which aims to address problems related to privacy, authentication and trust.

> Liberty Federation allows consumers and users of Internet-based services and e-commerce applications to authenticate and sign-on to a network or domain

> once from any device and then visit or take part in services from multiple Web sites

[LIB]

Hence Liberty Alliance deploys the Federated SSO Model, where a number of organizations form the federation (circle of trust).

### 4.5.4 WS-Federartion

This federation provides a specification (`http://download.boulder.ibm.com/ibmdl/pub/software/dw/specs/ws-fed/WS-Federation-V1-1B.pdf`) for web services, which was developed by IBM, Microsoft, Novel, VeriSign and others.

> The goal of federation is to allow security principal identities and attributes to be shared across trust boundaries according to established policies. The policies dictate, among other things, formats and options, as well as trusts and privacy/sharing requirements.

[WSF]

## 4.6 Responses to Threats

Above chapters and sections described risks occurring while taking part in online transactions, methods more ore less capable to protect peoples privacy while surfing online and tools related to them. This section will now focus on additional ways to battle risks. These include among others issuing of laws, which aim to protect people from harm caused by these risks or punish people responsible for harm inflicted. User initiatives are another approach, which aim to educate users. All these described nice and shiny tools are useless or even do more harm than good, when used in a wrong way or people are not even aware of risks online transactions and information disclosure (whether consensual or not) could pose.

### 4.6.1 Responses to Spyware

While most tools and methods introduced above aim to protect user privacy while transmitting data, they not cover spyware, which tries to obtain private information directly on local machines and transmit it to the creator of the malicious code. Similar to software, which scans for viruses on local harddrives, anti-spyware scanner (e.g. Ad-Aware `http://ad-aware.softonic.de/`) try to detect and remove malicious applications. This is achieved by comparing the files stored on a local computer to so-called *signature files*. These files have to be kept up to date, in order to be able to detect the newest spyware variants.

#### Legal Responses

One of the reasons for sluggish legal responses to spyware is related to the fact that most people are still unaware of the existence of spyware and therefore the use of malicious software and tactics didn't trigger enough tumult to force legislators to act in larger scale.

As of now the FTC[14] has the authority to take civil as well as criminal actions against installers of spyware.

Below are some laws and amendments listed that were issued until 2004 [Sip05]:

*I-SPY:* Internet Spyware Prevention Act of 2005 prohibits

> intentional access of a protected computer, without authorization, to install spyware to transmit personal information with the intent to defraud or injure an individual or cause damage to a protected computer

Penalties in case of violation range from fines to imprisonment up to 5 years.

*SPY-ACT:* Securely Protect yourself Against Cyber Trespass Act tries to

> protect Internet users from unknowingly transmitting personally identifiable information through spyware

Actions prohibited include taking control of a computer, modify settings, collection of private information, ...

### Industry Self-Regulation

In addition to legal responses adaware providers should follow the following principles [Sip05]:

- Users should be clearly notified that adaware is included, even before software is downloaded or installed. This notification should also be included in the EULA.
- Users should have the opportunity to (dis-)agree to terms associated with installation/download of software.
- Unwanted software should be easy to uninstall.
- Pop-ups should be clearly marked, to make it easy to identify the ad's source.
- Adherence to above stated principles and best-practices for Internet-business (in order to do not harm people's trust in online-businesses and transaction partners).

### Technical Approaches

Additionally to spyware scanners and spyware blockers (cf. [Sip05]) tools exist, that assist people in avoiding websites, that may contain malicious applications. McAfee's SiteAdvisor (bundled with McAfee VirusScan or as a plugin for your webbrowser of choice) test sites and marks them with icons of different colors. Green stands for a safe site, while yellow with exclamation mark signals that problems were discovered while testing (e.g. spam reached the inbox, after the email address was disclosed at this website). Red represents a unsafe site, which contains (downloads to) spyware. Sites marked with gray and interrogation mark were not tested yet by the SiteAdvisor team. Another example for

14 US Federal Trade Commission

such a tool is *Web of Trust* (`www.mywot.com`). It works similar to McAfee's SiteAdvisor, but is updated more regularly. Thanks to these supportive tools people may be protected from unintentionally accessing a site that contains spyware, distributes spam or even triggers "drive-by" downloads[15].

### 4.6.2 User Initiatives

More important than developing and adapting tools may be education of users. Many or even most users using the Internet aren't aware of the risks and threats the WWW can pose for not just the privacy of users. Hence users should be advised to read the often long and complicated EULAs bundled with software. While P3P (already introduced in the negotiation tools section) tries to assist users when dealing with privacy policies issued by websites, a similar tool for helping people with EULAs would surely be welcomed (at least by users).

Users should only download and install software, if they are really know what they are dealing with. Software provided for free download, could be used as a hide-out for spyware and because the software is free, it is more likely that it is downloaded (and installed) by users (cf. [TFS04]). Commercial software however isn't immune either. Creators of spyware often pay for the ability to bundle there malicious (probably unwanted) code with commercial software. The notification, that "spyware" is bundled with the to be installed software is often hidden deep inside the EULA. So bundling of spyware is just another reason for reading the EULAs thoroughly.

Additionally users could be made aware of alternatives to the currently most widespread webbrowser (cf. [Sip05]): Microsoft's Internet Explorer. Because this browser is still the most used, it is unfortunately also the most targeted by viruses, spyware and other malicious code. Although Microsoft is distributing patches regularly to get rid of vulnerabilities, new "holes" in the code, that could be exploited by attackers, are discovered nearly every day.

One alternative is Mozilla's Firefox. This browser is provided for free download (`http://www.mozilla-europe.org/de/firefox/`, available in current version 3) and patched regularly. It is not yet that widespread like Internet Explorer and therefore not the main target for attackers.

### 4.6.3 Protection against Social Engineering

For protection against social engineering user initiatives, the education of users are most important. Unlike spyware which poses (also) a technical threat, social engineering happens on human level, while human beings are communicating with each other. But there are also essential safety measures people/companies should be aware of [SOC]:

- educate people/employees thoroughly
- keep number of people authorized as small as possible (*principle of least privilege*[16])
- destroy/delete information, not needed anymore

15 downloads are not triggered manually, but by simply loading a site

16 every user/ module has access to the information he/it needs. no more, no less

- never use standardized usernames/passwords

But most important people need to be made aware that threats/risks related to social engineering exist. People should be aware that data is sensitive and should be protected. (cf. [SOC])

### 4.6.4 Possibilities to avoid Data (Identity) Theft

The growing problem of identity theft was already mentioned in this thesis. To address this issue measures are needed, which make it more difficult for criminals to access and combine identifying data.

#### Data Privacy Lab

The Data Privacy Laboratory (`http://privacy.cs.cmu.edu/`) together with *Journal of Privacy Technology* (JOPT) aim to provide a platform for sharing in order to communicate developments related to privacy. Apart from the list of projects conducted, the site provides white papers, publications on the subject and demonstrations (e.g. *CameraWatch* - a repository containing links to publicly available webcam feeds, *SSNWatch* - validation of SSN).

#### Database Privacy

With improvements in technology and data-storage getting cheaper, it is easy and economic to store data about people in huge databases (cf. [Swe05]), often without any particular purpose. This huge collection of information is very tempting for criminals. Hence the data needs to be secured, while the data can still be used for the intended purpose. To be exact there needs to be a *balance between confidentiality, integrity and availability* (cf. [Oli02]).

*Informed consent* is often used in guidelines concerning sensitive data stored in databases (e.g. who has access to data?, How data is secured?). People are able to give consent, hence they can choose whether they want to disclose the data or not. In fact, this choice is often illusory, because of lack of useful alternatives - hence people have no other choice than to disclose the data.

#### Selective Revelation

To protect privacy periodically through surveillance (in favor of "national security") *selective revelation* aims to still *allow data to be shared for surveillance*, while the shared data has still *provable assurances of privacy protection* (cf. [Swe05]). The (binary) decisions of US judges on the topic of private data sharing with law enforcement are replaced by technology (algorithms).

Identified privacy problems with data stored in databases are:

- many people stored in these databases have done nothing wrong
- Fair Information Practices (Notice, Choice, Security, ...) may not be implemented

- no juridical review or impartial oversight

To ensure privacy, while the data is still useful for surveillance, [Swe05] introduced the following *Privacy Conditions for Databases*:

- no person whose information is contained in the Database can be re-identified without permission
- investigators can access necessary information contained in the Database freely and easily
- results from qualified inquiries are equivalent to results found in the absence of privacy protection

The process of a (law enforcement) officer asking for data disclosure can be modeled in technology (cf. [Swe05]) by replacing

- the officer with data mining algorithm
- the informant with data (from various sources)
- the human judge with contracts and policy statements

The advantage of selective revelation lies in the fact that the level of anonymity of shared data depends on the situation it is used in, as opposed to a binary decision (disclose/stay anonymous). When more identifiable data is needed, then data, which more identifiable, can be accessed.

#### Identity Angel

Sensitive information more or less publicly available (e.g. resumés including SSN) make data theft or rather identity theft even more easy. To address this problem, Latanya Sweeney developed a tool called *Identity Angel* in 2004 (cf. [Swe06]). The software crawls the web for freely available information, which could be used for identity theft and tries to notify people of this data exposure by email, so suitable actions can be taken (e.g. deleting the data).

To achieve this Identity Angel uses filtered searching, every page in the results is scanned for e.g. a certain layout, phrase and pattern, to find so-called "roosters" (list of people's names and information). Together with *SSNWatch* (`http://privacy.cs.cmu.edu/dataprivacy/projects/ssnwatch/index.html`) the *Data Privacy Lab* aims to reduce the number of potential victims to identity theft.

## 4.7 Summary

Because of the rising threat to privacy it got more important than ever to use precautions. Approaches and methods introduced in this chapter provide different levels of anonymity, as services provided on the web need different information about the user (with different levels of sensitivity, respectively).

Anonymity provides the highest level of privacy. When using anonymity tools accurately, it is almost impossible to link requests back to a particular user. Requests made are routed through a network consisting of other users (e.g. CROWDS), before they are eventually sent to the targeted web server. Hence, the IP-address associated with the request is not necessarily the IP-address owned by the request's originator.

Unfortunately, many web servers and services need at least some information about the user (e.g. newsletter, discussion groups). Most of the time people are asked for login information like username/email address and password to link actions on a website back to a particular user in order to store profile information or layout preferences. To maintain unlinkability among different websites, login information should be different for each single website a user is registered for. Since remembering different login information is not an easy task, the pseudonymity tool *LPWA* aimed to help people in creating and managing these pseudonyms (partial identities).

On the other hand, for services like webshops pseudonyms are not suitable. Thereby, the user has to disclose far more sensitive information (after registering with a pseudonym). Data like postal address or credit card number has to be truthful and accurate, otherwise the service cannot be used.

To protect sensitive information, web services should take precautions. Sometimes encryption is used to make information unreadable, unless the associated encryption key is used for decryption. Thus the information is still protected, even if attackers manage to get hold of the encrypted data itself. Apart from protecting stored data, encryption is also used to secure information during transmissions.

Some kind of privacy is provided by pseudonymity, comprising numerous models of the implementation of pseudonymity. The most prominent principle among the introduced identity management models is called *Single sign on (SSO)*, where a user logs on at a central authority. Once authenticated, the user is able to use numerous services (which are connected to that particular CA). These models using SSO reveal disadvantages as well, since a user is authenticated with the same ID at numerous services. Hence unlinkability is not ensured anymore.

The next chapter will take a closer look at the advantages and disadvantages of these introduced approaches, while also taking a look at the benefit PET's offer in general.

# CHAPTER 5

## Discussion and Comparison of PETs

Usually after itemizing numerous approaches and methods, the following section will focus on the subjectively best alternative listed. During research, however, it soon became clear that there is no such thing as an "overall" solution for privacy protection. Different services provided on the web have different requirements. Some of them need more, others need less data about users, who are about to make use of a particular service.

Hence, there is a necessity for approaches that provide the data needed, while keeping the highest level of privacy possible. After describing different approaches and possibilities, this chapter will now focus on their pros and cons.

There seems to be no particular working way to sort these different approaches. The order in which the numerous sections were presented in the previous chapter was not set up after strict guidelines. Not all approaches work on the same level in favor of privacy.

While anonymity and pseudonymity provide different levels of anonymity, negotiation engines and encryption methods are rather auxiliary tools. Negotiation engines like P3P assist people when they are confronted with privacy policies of different websites. P3P helps users to decide whether a site's privacy policy is in line with the own personal privacy preferences. Different encryption methods, on the other hand, are used to secure sensitive data itself, regardless, if it is stored on a remote server, a local computer or if it is being transmitted. Apart from security, encryption is used for authentication (e.g. PGP) to determine whether the message/file was indeed sent by the stated originator.

### 5.1 Anonymity

Anonymity provides the highest level of privacy regarding the unlinkability of single requests made (to webservers). Tools like CROWDS help people to mask the actual IP-address of the request's originator. However, anonymity respectively the tools to provide anonymity have downsides too.

When regarding different types of anonymity, the tools nowadays do not provide an overall security. The capability of the networks to mask IP-addresses and the protection against traffic analysis attacks depends on the number of users taking part in the anonymity network (*environmental anonymity*).

Another issue concerning anonymity tools is the limited possibility to provide *content-based anonymity*. While the IP-address is masked, users are not protected from the

unwanted transfer of sensitive information (real name, postal address, email address or creditcard number) in messages posted in public chat rooms, discussion groups and on boards (regardless if this disclosure of information is intentional or not). The difficulty to maintain content anonymity is also an issue regarding industry espionage[1]. Employees could unintentionally post valuable information about developments, when asking for help on a public discussion group.

Although anonymity protects the identity connected to a IP-address, it fails to provide security against several other methods, which could threaten privacy. When a request is sent through an anonymity network, the response to that request is sent back to the originator. That response could include cookies, spyware and other kinds of malicious codes. With cookies it is possible to link several requests, even if these requests were made on different clients. Cookies are saved on the local hard drive. When a site is revisited, data saved with the cookie (e.g. preferences, automatic login, ...) is used, even if the cookie was not saved by the site in the first place. These so-called *third-party cookies* are usually used for monitoring which sites were visited by a user and derive profiles of interests (for direct marketing). For the protection against such threats, users should take a look at the preferences of their web browser. Current versions of wide-spread browsers include options to block third-party cookies or cookies in general. To secure a local machine against viruses, trojan horses and spyware (which could possible alter and/or steal sensitive data), the user is advised to install software like scanners and spyware-blockers. When used regularly to scan the system and kept up to date, it may be much more difficult for malicious codes to inflict harm on the system.

Eventually, anonymity assists users in keeping their identity secret. However, many services provided on the web need data to "recognize" users that cannot be gained by using cookies. In some cases this issue can be solved by *anonymous authentication*, where users need to be authenticated, before they are allowed to use the service/network. Authentication is achieved by identification data and can be revoked by authorized personal (e.g. administrators), in case of abuse. In this particular case, users could still be anonymous to other users, while being identifiable to administrators.

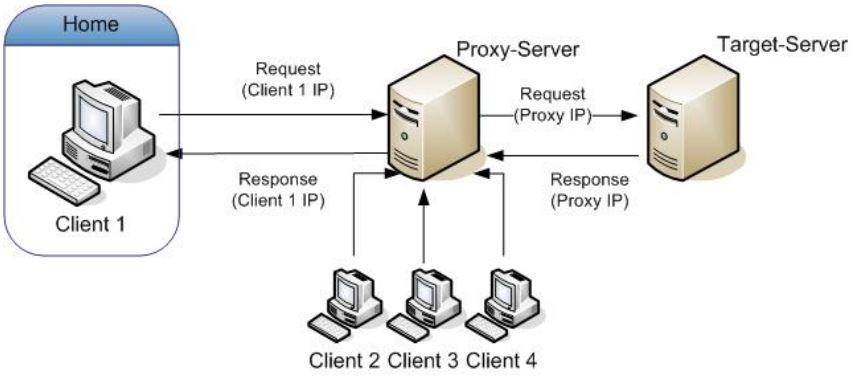


**Figure 5.1:** Anonymity

1 competitors try to get valuable information about (new) developments

## 5.2 Pseudonymity

To enable the usage of web services that need information about users in order to function properly, the principle of pseudonymity is used. With pseudonimity some information is disclosed, while a certain level of privacy is maintained. Unlike anonymity, the use of pseudonyms (partial identities) makes it possible to identify users (within a domain). Hence, data like profile information and preferences can be stored for each user. Sometimes username and password suffice as information to use a service, but in other cases, however, it is necessary to give away more sensitive data (e.g. postal address or creditcard number) for services like webshops, which need information that is accurate and truthful.

Normally login information (username and password) suffices to register for a public board or discussion group. Apart from entering a (at the time of registration) valid email address, other profile information is optional and initiated with default values. Sometimes users are asked for their date of birth, whereas this information is not validated at all.

Because of the chosen pseudonym users are not only identified to the webserver, where they are registered at, but also to other users, who are allowed to read the content. Users are not only recognized by a unique identifier (username), but sometimes by an avatar[2] and signatures[3] to make the recognition of users easier, as well. These images and signatures are not necessarily unique, though. Administrators of discussion groups and boards sometimes provide a gallery, from which users could pick an avatar to use. The amount of pictures provided is usually smaller, than the amount of registered users. Hence, users are allowed to pick any image/signature they want as long as the rules established (e.g. content, size, ...) by administrators are not violated.

When using a pseudonym unlinkability of actions (requests) is not ensured anymore. Linkability is even desired, when using discussion groups and similar services to enable a satisfying discussion. Like already mentioned in the anonymity section, people should be aware of the kind of information they disclose when posting messages online. Pseudonymity could pose even a greater threat to one's privacy, because of easier linking of messages posted by one particular user. Information included in these messages could be easily combined to derive profiles.

The domain where the user is identified must not be limited to just one single webserver (website, discussion group, ...). Depending on the model used, the domain could include numerous websites and services. In identity management models using some kind of single-sigh-on[4] approach, the user is usually authenticated by a central authority (CA). The CA holds the data to authenticate users, thus all services connected to that CA are able to identify (and link actions taken by) the user.

Despite SSO, people are still forced to use more than one CA, when surfing on the Internet. Yet there is no CA that takes care of authentication for services worldwide. Microsoft once tried to implement a central authority with *Windows Live ID*, but it was not well accepted by the public. People were intimidated by the fact, that sensitive data used for authentication (and other purposes) is managed by one single authority (c.f. [Kob03]).

2 small image usually placed directly beneath the username, when messages are displayed
3 text and/or image that is placed beneath the posted message
4 once authenticated by the server, users are able to use all services connected to that server

When unlinkability is no longer ensured within a domain, wise choosing of pseudonyms is important. Pseudonyms should be easy to remember and recognize, while pseudonyms chosen by the same user and used for different domains should not be linkable. In order to maintain unlinkability among different domains, [BP05] refers to identity management applications which create pseudonyms automatically for the user. These identifiers consisting of random characters are not easy to remember and recognize, hence the usability is very low. This is were *mnemonics* come into play, working as representations for automatically generated pseudonyms. Similar to the scope of domains, the security of pseudonyms depends on where the mnemonic is assigned. Global mnemonics are picked from one set (dictionary), while local mnemonics can be chosen by the user himself/herself.

Regardless of the pseudonymity model, actions taken by a user within a domain are linkable. Hence, it is still most important that requests made among different domains should not be linkable. Users should be careful when choosing a pseudonym (partial identity, login data, ...).

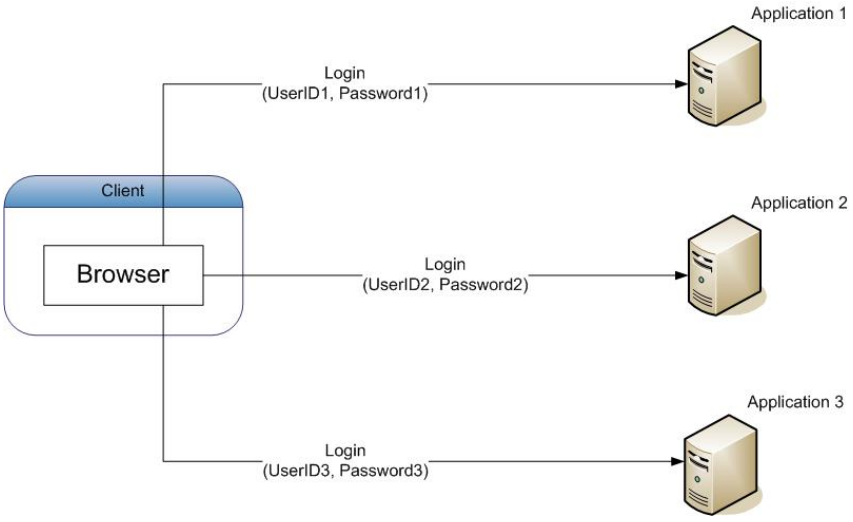


**Figure 5.2:** Pseudonymity

## 5.3 Negotiation and Trust Engines

In many countries it is mandatory for websites to include a so-called *privacy policy*. It contains practices how sensitive information disclosed by users is handled, stored or shared. These policies are usually long, rather complicated and written in a language typical users are not used to (cf. [Rea99]). With P3P (Platform for Privacy Preferences), privacy policies can be parsed automatically by user agents. After a user has determined his/her privacy preferences the user agent is able to compare the privacy policies sent by a website with the preferences chosen. If the policies do not violate privacy preferences connection is established, otherwise a new offer is requested by the user agent or connection is not established at all (and users are asked to take further actions).

When the standard was first introduced in 1999, it was still unclear if the protocol will be accepted by website owners and users alike (cf. [Rea99]). Nowadays, however, the

protocol is widely used. The fact that user agents are now included in the current versions of widespread web browsers boosted the usability of the protocol.

It is important that companies act according to the policies displayed on their website in order to facilitate the protection of people's privacy online and to maintain the function of P3P. Projects like *TRUSTe* aim to ensure that companies do not violate their own policies. By checking privacy practices regularly violations should be discovered quickly. If privacy violation occurs, legal steps are taken.

To show that a website's owner is a member of the TRUSTe project, a seal is displayed on the site. A link connected to the sign enables people to check for themselves whether the owner is indeed a member of the project or if the sign is illegally displayed.

One objective of P3P and TRUSTe is to strengthen the trust of people working in online companies as business partners. People should be assured that sensitive data sent and disclosed is being kept safe.

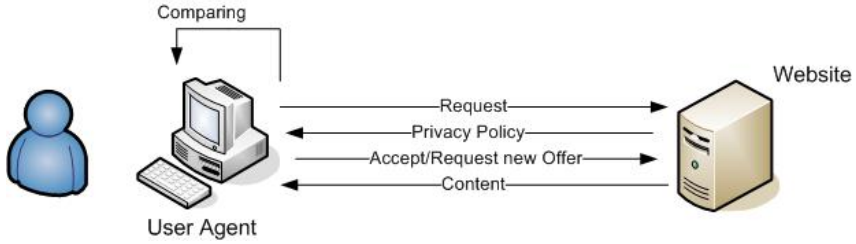


**Figure 5.3:** Negotiation Engine

## 5.4 Encryption

While anonymity and pseudonymity aim to protect one's privacy by limiting the (sensitive) data disclosed, encryption protects the data itself, by limiting the amount of people who are able to read it. At first, encryption was only used to encrypt stored data, later data was protected during transmission too. Many applications and (secure) protocols (e.g. SSH) use some kind of encryption.

Sensitive data saved on remote servers (e.g. bank institutes, online shops, ...) should be protected properly. Encryption of that information would not prevent theft of data itself, but decreases the value of data for unauthorized people. Even if attackers manage to gain access to the servers where the data is stored and steal the information, they still have to decrypt the stolen data with the right decryption key.

Apart from securing data, encryption can be used for authentication as well. With PGP (Pretty Good Privacy) a freeware solution exists, which makes it possible to determine if the file/message was sent indeed from the originator stated.

## 5.5 Discussion of PETs and Privacy

The introduced and described PETs currently available claim to limit access to sensitive information, increase the level of anonymity or protect one's privacy. Already in 2001, however, authors of [Tav01] argued that PETs do not necessarily provide privacy protection.

Privacy itself is often described as *a moral right or legal right* (cf. [Cla99]), *the wish to control the information that might be embarrassing or harmful*, *granting us control over our lives and enabling us to decide* or as the restriction of access to sensitive information (cf. [Tav01]). Control and limitation of access to information about oneself pose difficulties in today's online-world. Information is stored at many places, on servers and in databases all around the world and in many cases the user has, in fact, no control over it. Sensitive data is disclosed (medical information, religious confession) to government agencies and webshops. Right at the moment private data is sent to the server (after deciding which data to disclose), the user loses control over the sent information. However, the *loss of control should not entail the loss of the right to privacy*, where control is described in three different ways:

*Choice:* users can decide, which data to disclose, the level of risk (and who has access to it)

*Consent:* users can give permission for accessing their private data

*Correction:* users are able to access data (stored on remote servers) and are able to correct (or delete) it

Hence *new zones of privacy need to be created to protect individuals especially when individuals lack control of personal information and cannot protect themselves* (information, which is mandatory disclosed, should only be accessed by people, who are bound by confidentiality). According to *Fair Information Practice Principles*[5] adopted by the US Federal Trade Commission, access should be provided for users to correct or delete the information if needed. The principles also demand that owners of web servers implement adequate security measures to prohibit unauthorized access or the manipulation of data. Once the data is stored on a remote server users should be able to *trust* that the data is indeed handled like stated in the *Privacy Policies*. The PET *P3P* helps people to handle the often complicated and abstruse policies. Initiatives like *TRUSTe* ensure that information is handled, protected or shared according to published policies.

Privacy policies, however, are no guarantee that the privacy of information will be protected forever. The acquisition of companies could pose a risk to privacy of sensitive data. Although privacy had been maintained by the policies adopted by the former company, the new owner has access to that information and may use it while adopting new policies (e.g. Terms of Use and Privacy Policies of social networks like Facebook, cf. [EIN09]).

Another issue is related to the fact that many people are not aware of the risks they face, when disclosing sensitive data. Information like data saved in personal profiles in social networks, address-information and creditcard number at webshops or even comments made in discussion groups, could be easily used (by third parties) for purposes that differ from the aims stated in policies. Data mining, for example, represents a technique, which could pose a privacy problem for many people who are not that technical experienced. Data which is publicly available is used to find patterns, interests to derive future actions.

5 `http://www.ftc.gov/reports/privacy3/fairinfo.shtm`

Sometimes past actions of users are included by means of IP-address, cookies and even spyware. As already mentioned in earlier sections of this thesis, PETs claim to (partially) solve these issues. Anonymizers mask IP-addresses and settings of current web browsers make it possible to disable the use of cookies. The usability of these tools is not perfectly adequate (see chapter Future Challenges). Current PET's, as well, are not able to prohibit the adding of sensitive information to the message itself (content-based anonymity).

With *consent* it is possible to allow access to certain needed data (e.g. medical information, creditcard number). In most social networks (e.g. MySpace) it is possible to determine more or less detailed who (users, groups, inside and outside of the network) is able to view profile data, guestbook entries or comments made in discussion groups. When data is accessed without given consent beforehand, privacy is violated.

Apart from choice there is another aspect of control: *correction*. People are able to access (private) data related to them and may be allowed to edit it. In web shops like *Amazon* it is not only essential for providing the service to enter and view personal information (e.g. address and creditcard number), but also to update the data, in case something changed (e.g. creditcard subscription expired). On pages where people are able to leave anonymous or pseudonymous comments, however, it is usually not possible to edit the information after sending and storing (or just for a short period of time). In these cases it is also more difficult to track a comment back to one particular user. On sites that do not require a registered account to add comments, the most obvious instrument to link several comments to one individual - the pseudonym - may not be any help at all, because users are allowed to choose a pseudonym before adding every single comment. Possibilities to link actions anyway, offer "usual suspects" like IP-address and cookies. Web applications, which use registered accounts, normally allow to edit information (e.g. profile information, comments added).

*Deleting* of once stored information, which is related to editing, is usually the most difficult task. While editing usually allows the deletion of data, the most part of stored information can only be viewed (accessed), but not altered or erased. Discussion boards and boards, which do not require a registered account (e.g. comment feature provided at online-newspaper portals), usually do not provide a feature to edit or delete a once typed message. Information saved on servers and in databases all around the world could be harmful or at least embarrassing, when it is observed by unauthorized viewers (cf. [SDH08]). Regarding discussion boards with registered accounts and social networks it is not certain that data saved in profiles or the content created is erased after the registered account is deleted (cf. [EIN09]). Usually the messages can still be viewed (sometimes even with the author's pseudonym) and have to be manully deleted by the administrator. Today's search engines (e.g. Google) allow the access of information, which would be otherwise difficult to find. Sometimes it is even possible to view media, which should only be viewable from within a (social) network (cf. [PSN08]). Other search engines, which are especially designed to find information about people (e.g. *123people.com*), provide data (e.g. address, telephone number, profiles registered in several social networks) of the "targeted" individual. Not only family and friends may want to seek information about a person, but future partners in a relationship and more important future employers, who want to check on people applying for a job. Hence it is important to keep publicly available information related to oneself up-to-date and as accurate as possible. As already mentioned, the altering or deleting of embarrassing or improper data may be not easy

at all. PETs listed above do not assist users to achieve that goal. In fact, PETs do not ensure that privacy is not violated after sensitive information has been sent to a remote server (cf. [Tav01]). Usually people are confronted with the choice of *opt-in* or *opt-out* to allow or prohibit secondary use of sensitive data.

One of the most important PET's *Encryption* can help to protect sensitive (or embarrassing) information from unauthorized access and view - encrypted data is unreadable, unless the corresponding key is used to decrypt it. On the other hand, applications use encryption to protect data during transmission (e.g. SSL).

Regardless of which type of PET we are looking at - anonymizer, identity management tool or encryption tool - many people are not even aware of their existence. Currently distribution of PETs is not well developed either. Nowadays, people have to search distinctively for a PET (by name or type), download the file and install it. After the successful installation, configuration and everyday use of the chosen PET may not be that easy either (usability).

To sum it up, PETs are no guarantee for privacy protection (cf. [Tav01]). If people are aware of the capabilities of PETs, they can aid users to control which (sensitive) data is disclosed. For better protection privacy policies are needed to create zones which ensure the security of sensitive information.

The table below shows once again the PETs introduced in the previous sections:

| **PET** | **Connect. Anonym.** | **Data Anonym.** | **Personalization** |
|---|---|---|---|
| **Anonymizer** | low | high | n/a |
| **Pseudonymizer** | high | n/a | n/a |
| **Identity Manager** | n/a | medium | medium |
| **Encryption** | n/a | n/a | n/a |
| **Negotiation and Trust** | n/a | n/a | medium |

**Table 5.1:** Comparison of PETs

# CHAPTER 6

## Future Challenges

While researching the state of the art technologies which aim to protect people's privacy and to strengthen trust in online businesses and transactions, it was soon clear that there is still much need for improvements. While developers of security software need to stay up to date reagarding the improvement of technologies and methods to fight privacy abuse, current approaches to battle threats and risks still have their flaws.

The list below contains issues, which have to be solved in the future (cf. [Kob03] and [Lin07]):

## 6.1 Content-based Anonymity

Anonymity tools like JAP aim to protect the privacy of users by masking IP addresses (several users connect to the same proxy, whose IP address replaces all the others), simultaneously giving advice to prohibit methods which unmask identity (e.g. disabling Javascript, ActiveX). The use of pseudonyms helps people to take part in online discussions without the need to give away their full identity. All security precautions are useless, however, when private information (like postal addresses, contact information) is disclosed in the message posted itself.

Current tools and future versions aim to improve content-based anonymity. Improvements should take all subareas of these branch of privacy into account (cf. [Sch03]):

*Formal Anonymity:* all content should be removed or altered that (in combination) serves as a unique identifier for users (e.g. postal address)

*Contextual Anonymity:* no deanonymization should be possible by message content

Mechanisms to ensure content-based anonymity mainly depend on the applications used. From the secure choosing of pseudonyms (and mnemonics) (cf. [BP05]) to filtering of messages, every step should be covered.

### 6.1.1 Choosing of Identifiers

Pseudonyms could be

- chosen by the user

- picked from a limited set
- created automatically

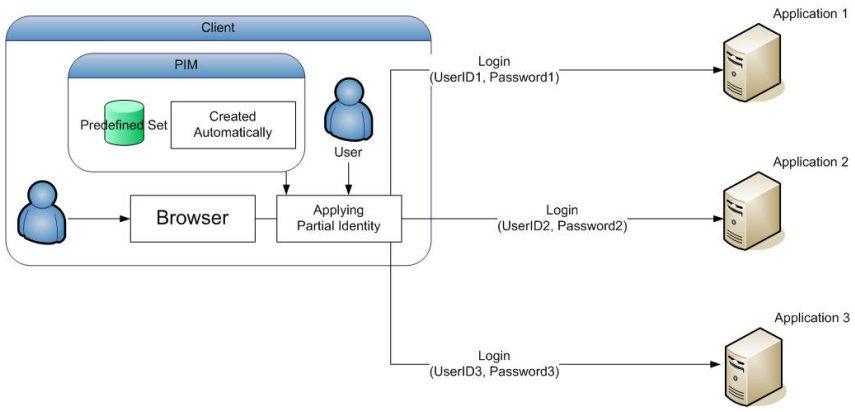


**Figure 6.1:** Creation of partial Identities

All three possibilities provide different levels of security as well as usability, whereas the ratio between the two attributes is unproportional.

When users are allowed to pick pseudonyms, the chosen identifier may be easy to remember and recognize, hence the usability is high. Leaving user-picked identifiers unchecked would decrease the security. Chosen pseudonyms or mnemonics could disclose information or could be linked to other similar identifiers representing the user at a different domain.

Discussion groups and similar applications may offer a so-called dictionary containing predefined identifier users can choose from. The set, however, has to be big enough to contain enough unique and secure identifiers. Furthermore, these identifiers should be easy to remember for users. This is no easy task, considering the users with different cultural backgrounds.

Pseudonyms created automatically by the system/application have a high security level, because they consist of random characters. The possibility of uniqueness is high, but the usability is rather low. Therefore, mnemonics are associated with these pseudonyms to make them easier to recognize and remember. Mnemonics and pseudonyms are subject to the same rules. Even when automatically created pseudonyms are secure, inconsiderately chosen mnemonics could still compromise the security of the system.

### 6.1.2 Checking of Message Content

Apart from the identifier associated with the message, the message's content itself could (unintentionally) compromise privacy by disclosing sensitive information (e.g. given name, city the user is living in); the combination could deanonymize the user's true identity.

The system could alert the user, when he/she is about to disclose data, which could compromise his privacy. Identifying data, which may be harmful to privacy is a difficult task. Firstly, the content has to be compared with sensitive data stored in the user's profile (cf. [Kob03]). Alternate spelling, typos, abbreviations or data which is not stored

in the profile yet would not be recognized. The system should distinguish between data that is safe to be sent and information that should be removed, replaced by pseudonyms or slightly altered. Implementations of spellcheckers face a similar problem. They are usually only able to compare entered words with the related dictionary. Checking whether the meaning of words is fitting in the sentence is not yet possible.

Additionally, users should be allowed to associate tags to sensitive data stored in profiles/databases, so the system registers the sensitivity level of each single piece of information (see profiles for social networks in the following chapter), so the system is able to decide, whether the user wants to disclose the information willingly.

## 6.2 Anonymous Authentication

Anonymous authentication is some kind of mixture between anonymity and pseudonymity. The webserver is still able to identify/authenticate a user, while he/she remains anonymous to other users. Digital libraries or webservers with premium content use this approach to restrict the access to non-public content. The user is only identified by the webserver itself, other users may only be able to see the number of users currently logged in.

The authentication process or storing of identification information still needs improvement. How is the user authenticated to the server? Should the server be able to identify the user at all? Or it is sufficient for the server to know that he is dealing with an authenticated user, who is allowed to access content on the server?

Additionally, services like webshops need sensitive data like postal address and creditcard number or bank account information for the transmission of products to customers. This data is usually stored on remote servers owned by the shop and needs to be protected. If users have reasons to mistrust the precautions, people may decide to stop business dealings with the shop.

Oftentimes, the storage of sensitive data at the webshop's server could be avoided (cf. [RL08]). The webshop gets confirmation from the creditcard company that the user in question owns a valid credit card. So the shop gets paid by the creditcard company without actually knowing the creditcard number.

When contacting the shipping company the webshops just send an identifier, so the shipping company is able to determine the user's postal address.

In both cases the businesses have to agree about the validation and the accuracy of the sent information.

### 6.2.1 Hypertext Access

One of the basic possibilities to protect single files/whole directories stored on a (Apache[1]-) webserver is *htaccess* Hypertext Access. To access a protected directory directly (without the authentication dialog) enter: *http://user:password@www.server.com/directory* (insecure, because password is visible in plaintext).

With htaccess (configuration file often named .htaccess) it is also possible to configure

1 `www.apache.org`, HTTP Webserver for numerous operating systems, often used with the script-language PHP. Easy to install package with other useful modules like MySQL: XAMPP - `http://www.apachefriends.org/en/xampp.html`

- redirections to errorhandling sites, local links
- limit access to certain computer (IP-addresses)
- limit access to certain users (groups)
- restrictions for different types to access a site (GET, POST, PUT)
- restrict access for certain web browsers, search engine robots

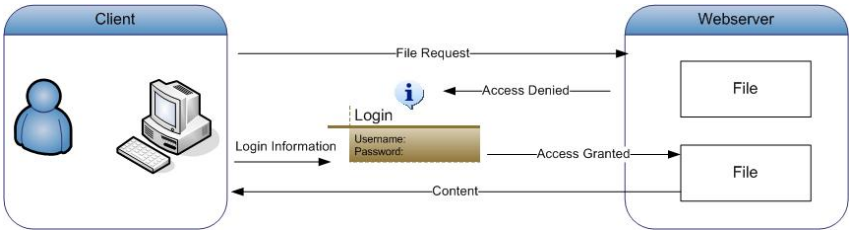


**Figure 6.2:** Hypertext Access Principle

Configurations are only valid for directories and subdirectories, where the htaccess-file is saved at. Users, who want to gain access to the stored content, have to enter a valid username and password.

The fact that a user is not identified to a server prohibits the revokement of a user's anonymity in case of abuse. Additionally, htaccess only limits access to content over the internet (cf. [VER]). When users are logging in directly at the server, full access is granted without authentication.

### 6.2.2 Personal Authentication Device (PAD)

PADs like cell phones, PDAs or chip cards can be used for authentication at a server, whereas ID information or just an identifier is sent to the server. This approach can also be used for sign forms or administrative activities digitally, whereas documents otherwise are being sent by mail and signed by hand.

One example for the use of chipcards as a personal authentication device is the so-called "Bürgerkarte" (Citizencard, `http://www.buergerkarte.at/en/index.html`) in Austria.

To make use of the citizencard concept some requirements have to be met with, which imply at least some experience with electronic devices (cf. [CIT]):

- computer (PC, notebook or similar)
- card reader (to connect chipcard with computer)
- application software (to decrypt information stored on the chipcard)
- chipcard

The chipcard stores private information and the digital signature of the user, while the user can only view the data stored on the card without the right to alter it. Hence, the

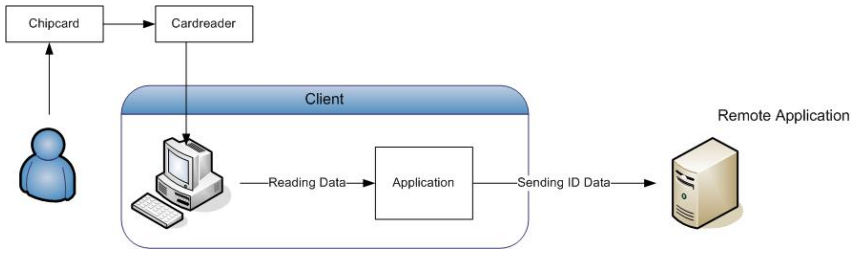


**Figure 6.3:** Identification via Chipcard

faking of identification data is prohibited. On the other hand, identification to the server (authority) enables the revocation of authorization if needed.

When using chipcards as PADs requirements to make the authentication process work, is rather demanding. People need certain software as well as hardware (PC, cardreader, application software, ...) in order to identify themselves via chipcard. Alternatively a device could be used as PAD, which many already have in their possession (cf. [MM07]). Cellphones, PDAs and similar devices are already used by several bank institutes for transaction verification (see Identity Management: SMS Authentication). Usage of short-term secrets instead long-term secrets (passwords), makes authentication process less prone to keylogger-attacks[2]. In combination with SMS, users are asked to compare data and enter (short) PIN-Codes (short-term secret: only valid for one particular transaction) displayed into the user interface on the (untrusted) client PC. To make this method work, people should be able to open and read SMS messages and take extra care while comparing and manually copying codes.

On the other hand *MP-Auth* (cf. [MM07]), a protocol designed for online authentication, uses mobile devices also for encrypt or de-cyphering processes. Mobile devices should be able to turn long-term secrets into short-term secrets in order to provide transaction integrity.

To make sensitive transactions more secure (anonymous) authentication processes (like MP-Auth) should provide protection against:

*Keylogging:* long-term secrets aren't stored on or entered at the client PC

*Phishing:* spoofed websites aren't able to encode encrypted passwords

*Pharming:* results of pharming (redirection to spoofed websites by e.g. manipulation of host-files) are similar to phishing, hence MP-Auth protects against these kind of attacks too

2 attackers log keystrokes to get eventually hold of valuable information like login data and passwords

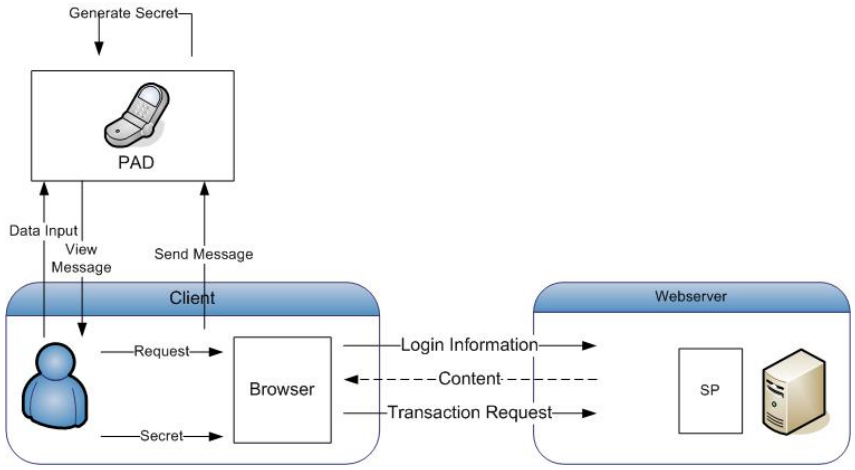


**Figure 6.4:** MP-Auth Protocol

## 6.3 Awareness of Risks and PETs

This thesis already stated that people are often not aware of the risks they face when engaged in online activities (cf. [Tav01]). Previous chapters and sections described risks and threats, which endanger sensitive data that is stored locally (e.g. spyware), on a remote server (e.g. hacking attacks) or during transmission (e.g. eavesdropping).

It is also important, however, that people are aware of measures and tools, which aim to help protecting privacy and sensitive data. Scanners created to detect viruses like worms and trojans are finally very popular these days - trial-versions of these tools are often bundled with computer systems already assembled at distributors. These scanners, however, provide only limited protection against all different kinds of spyware. To address this issue tools are available, which were especially created for detecting and blocking spyware. Software protecting against spyware is not as widespread as scanner tools dealing with viruses.

Websites like `www.anonymitychecker.com` provide information about anonymity in the WWW and offer links to useful tools like anonymizers and lists containing working proxy-servers. Other sites offer opportunities to check if the chosen anonymity tool is indeed working or show that the IP-address related to requests is not the only threat to ones's privacy (e.g. `www.gurusheaven.de`).

Anonymizers (or proxy servers) usually only mask the IP address, but script languages like JavaScript offer additional opportunities to gain information about users (cf. [GHS]).

- used screen-resolution, operating-system, web-browser can be determined with JavaScript
- with Java enabled even the internal IP-address, issued by a NAT-enabled router (e.g. 192.168.0.2) is visible and offers attackers the opportunity to conclude on the possible internal network-structure

- An exploit in CSS[3] makes it possible to determine whether a certain site has been visited
- it is possible to determine registered MIME-Types[4] and installed Plug-ins

Most of these risks could be avoided by deactivating or blocking cookies and script languages. Changing of web browser preferences would not only cause inconveniences while surfing, the particular options, which need to be changed, may not be easy to find, too. Clicking through several different menus may be too much for a casual or unexperienced internet-user (Usability - see below). Furthermore, many web applications and web servers nowadays need cookies or script languages activated in order to work properly.

Despite the increasing awareness of risks and threats, which endanger their privacy, people still need to know which tools exist for protection and where to find them. Above mentioned sites, which inform about risks may also provide links to adequate tools for protection. Casual internet-users may not search for information about distinct risks because they do not know of their existence. Hence it is rather unlikely that unexperienced users "stumble upon" supportive tools accidentally.

So how is the task of distributing PETs accomplished? Unexperienced users cannot be expected to find, download and install these tools. Hence bundling the tools with (widespread) software would be a solution, like it is already done with programs that may be considered spyware or other kinds of malicious software (cf. [TFS04]).

The integral part every computer needs is the operating system, hence bundling PETs with it would (partly) solve the distribution problem. The list below introduces programs, which may be considered as PET's and are bundled with *Microsoft's Windows* (the market-leader in operating systems):

*Internet Explorer:* Microsoft's web browser offers numerous options to adapt privacy and data-protection settings to one's preferences

*Firewall:* Since Windows XP SP2 a built-in firewall is included. The firewall, however, only monitors outgoing connections, but would help in detecting unwanted requests going out (e.g. trojans etc. trying to send data)

*Malicious Software Removal Tool:* Microsoft offers a tool to detect and remove malicious software (e.g. viruses). The tool is distributed over *Windows Update* and updated monthly

*Junk-Email Filter:* Incoming emails are parsed, if they match certain criteria. Mails considered spam are moved to the junk-folder instead of the inbox. The criteria used are updated regularly

PETs listed above provide "basic" protection of privacy, but there are still additional tools needed. Newer versions of MS Windows, for example, issue a warning if no virus scanner tool is detected on the system. Spyware detectors and blockers support privacy protection in addition to basic scanners.

3 Cascading Style Sheets - script-language makes it easier to create a consistent and easily changeable layout and design

4 Internet Media Type - defines content-type included in a website's body

## 6.4 Usability of PETs

Knowing about PETs, finding, downloading and installing them may not be sufficient to protect privacy. The tool you chose may need more attention in order to work properly. Even usage of standard web browsers require additional preferences to chose, if people do not want to endanger security of sensitive data. The ease with which settings related to security and privacy, however, depends on the web browser used.

The finding and understanding of security options may be rather difficult for unexperienced users, hence adequate default settings would help them to avoid risks to their privacy. Microsoft's Internet Explorer uses security levels (low, middle, high) to make it easier for people to adopt privacy options to their individually preferred level of security and data protection. Although the mean level is chosen by default, it is possible to pick another level as well as to change every detailed option offered (for sophisticated users). These detailed options are needed to provide an adequate level of protection, while at the different levels (profiles containing predetermined settings) privacy and data protection still provide usability.

Usability of PETs should not be neglected. Complicated tools with difficulty to understand options not just hinder unexperienced users from using the tool at all, but also increase the chances that incorrect set options endanger security even more. To make PETs usable for rather unexperienced users the tools should include:

- intuitive user interface
- thorough, but still easily understandable documentation
- help guide, which supports users during the first steps and the most essential options

A tool already introduced in this thesis - JAP - provides quite good usability. Installation is easy and the main interface is intuitive. After the start-up of the program the user is alerted that additional options need to be set. These settings, however, do not need to be changed in the tool-configuration itself, but can be manually set in the chosen web browser. An assistant provides guides and even screenshots for the most widespread browsers.

# CHAPTER 7

## Examples for current Issues with Privacy

The previous chapters focused on the risks to privacy and on approaches to protect sensitive data from these threats. This chapter will now take a closer look at definition of privacy in general and how the public's view on this issue may have changed after the incidents of September 11, 2001. Additionally, some examples of current issues with privacy will be presented.

The US-American constitution gave US-citizens the right to protect themselves and their property. In combination with the constitutional right to possess a firearm ("right to keep and bear arms", 2nd Amendment included in Bill of Rights[1]), US citizens have the means of protecting their *personal freedom*[2], property, homes and families. In some US federal states it is even legal to shoot an alleged trespasser. Apart from that radical example, many democratic countries have similar laws included in their constitutions, which aim to protect citizens against unauthorized infringements by other people or the government itself. Amendments included in the US constitution are for example (cf. [BOR]):

- *1st Amendment:* Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances.
- *4th Amendment:* The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.
- *8th Amendment:* Excessive bail shall not be required, nor excessive fines imposed, nor cruel and unusual punishments inflicted.

According to the 4th Amendment, a reasonable warrant is needed in order to conduct a search of premises. Additionally, US laws demand that the owner of premises has to

1 first ten Amendments of the US Constitution

2 ensured by civil liberties, as a legal and constitutional protection for individuals against the US-government

be present when the search is conducted. Yet the "Patriot Act" changed that attitude, at least in "certain" cases. If alleged terrorists are investigated the owner has not to be present anymore - the search can be conducted anytime.

Searches, however, still need to be legitimated by warrants, although it seems to be easier for law enforcement to get one issued by a judge. However, privacy of the home is still well respected. With the possibilities of modern technologies, the law enforcement may not need to set one foot into a criminals home/hideout to get hold of valuable information.

Philip Schaumann presents on his website (`www.sicherheitskultur.at`) a (regularly updated) list with threats to privacy people are confronted with (cf. [VAP08]). Some of the most important threats are presented below:

*Location:* when people hear *location tracking*, they probably think of storing traveling data (start, origin and additional very sensitive information), conducted in the US as well as European and other countries around the world. But location tracking is not limited to certain branches of life. Cellphone-providers routinely track the location of their customers in order to route calls and train or bus tickets bought online are usually not anonymous anymore.

In order to be able to buy a ticket at the ÖBB website (Austrian railroad company) you have to own a special card, which not only contains your name, but also a photo and your date of birth. The photo is stored digitally in order to provide a new card, every time the yearly subscription runs out.

*Location-based Services* help, for example, parents to keep track of their children or employers to monitor their employees. The tracking is usually realized by GPS-enabled cellphones or GPS attached to cars. Google's newest service for cellphones called *Latitude* is a means to track e.g. family and friends. Normally users are able to determine who is able to track them, but due to a security vulnerability it is possible to activate the tracking service without people's knowledge on certain cellphone models (cf. [DDD09]).

For road tolls and speed-limit checks (in Austria called "section control") technology is used to identify license plates and even automatically issue charges (payment forms) in cases of violation. The *Automated Plate Number Recognition* (APNR) is a system introduced in England to monitor tolls, so the collected data can be accessed by law enforcement.

*RFID-Chips*[3] were at first meant to tag products that can be used for several different purposes. Since 2004, passports include RFID chips which store (encrypted) ID information like biometric data. Schools in Japan and England are already using RFID chips in school uniforms. RFID chips could also be inserted in newer skis in order to determine the rightful owner at ski-lift stations. During the World Cup 2006 in Germany, the tickets were tagged with RFID in order to track "Hooligans".

*Communication:* with communication monitors (e.g. Echelon) it is possible to monitor telephone calls and other communication methods like e-mail and instant messaging.

3 Radio Frequency Identification, scanners are able to read out data stored on these chips within a distance of several meters

Encryption of content prohibits unauthorized reading, but still does not mask communication data (who communicates when with whom).

Due to bugs in cellphone software cellphones can easily be used as surveillance (recording) devices. With newer cellphones people may be able to record calls without the knowledge of the communication partner.

Call centers use voice stress analyzer to detect anger or stress in customer voices during calls.

In March 2009, the UK government will not only start to monitor email traffic coming from/meant for UK citizens, but will also save the messages themselves (cf.[GRO09]). This measure also includes foreign companies, which own a branch office in the UK.

*Surveillance:* Many inner cities (public places) in England are (almost completely) monitored by video surveillance, while in other countries rather privately owned buildings like offices and malls own a video surveillance system. In the US video surveillance and *face recognition*[4] is used at big events (e.g. Super Bowl) to identify a wanted delinquent within the crowd.

With *webcams* it is possible for people to monitor places (almost) in real-time. Because of improvements in technology cameras are getting smaller in size and are hence easier to hide. So-called "Nannycams", with cameras hidden in everyday items or furniture, help to monitor babysitters, while they are looking after your children.

*History:* Modern technologies provide numerous opportunities to keep track of actions taken in the past. Like mentioned before the *IP-address* attached to every request sent usually offers an easy way to link several action taken together. When the IP-address changes *cookies* stored on local harddrives still help to identify a certain user (e.g. a member in an online-community), unless cookies are disabled in web browser settings. Cookies can also be used to keep track of websites visited and to save additional information (e.g. keywords, preferences, ...). This information can be accessed when a website is re-visited.

Online-shops like Amazon do not only save the amount of products a certain user already bought, but also the product-sites he/she visited. This information is used to determine which products would probably be of interest for a particular customer. Sometimes users are even notified by email, if presumably interesting products are added or currently on sale.

Every search-query (keyword) sent to Google-servers is obviously stored. Searches from the past are not only used to make suggestions while users enter their search query, but also to derive the so-called *Google Trends* (`www.google.org/trends`). These trends are not just useful for everyday users to see what is currently in (online-) public interest, but the service also helps (marketing-, business-)companies through *data mining* (by finding search patterns)[5].

4 computer systems automatically identify individuals by digital images (e.g. photos, video surveillance)
5 `http://www.google.de/intl/en/trends/about.html`

A privacy issue which is rather regarding the "real world" revolves around chipcards, on which past actions are stored. Areas of implementation range from loyalty cards[6] to medical cards (e.g. the Austrian *eCard*)[7], which save the medical history of individuals, additionally to basic information like full name, date of birth and social-security number. The data includes every single hospitalization and consultation. In the future the information may not only be used for determining insurance rates, but also for indicating whether an individual is fit for a certain job.

Creditcard history, on the other hand, is not only used by law enforcement to determine what was bought in a certain period of time, but also for finding out when and where the card has been used (this is also true for ATM cards).

*Data Loss:* In 2000, the Federal Trade Commission composed the *Fair Information Practice*, which covenants companies to secure sensitive information they store[8]. In the past several years numerous incidents relating to data loss went public. One of the most recent incidents, which was covered by the media in October 2008, affected the cellphone-provider "T-Mobile" residing in Germany. The company admitted to have lost sensitive data of 17 Million costumers already in 2006.

## 7.1 Computer Surveillance

For the past couple of months and years computer surveillance (conducted by the government) was a hot topic in the US as well as Europe. Especially in Germany the issue is known by the public as the so-called "Bundestrojaner" ("federal trojan"). Discussions arose that the law enforcement should be allowed to gain remote access to a suspect's computer to transfer and monitor data saved on the harddrive. It is still unclear if this intrusive measure will only be conducted while prosecuting alleged terrorists or capital offenders. There is always a great potential for abuse, where *govware*[9] could be used for prosecuting even the smallest crimes.

Politicians and law consultants are still arguing whether the use of govware represents a violation of the fundamental rights, defined in the constitutions of the respective countries.

### 7.1.1 Limitations by constitutional Laws

A citizen's fundamental rights like

- privacy of home
- protection against unreasonable searches

are protected by the constitutions (of most democratically governed countries).

6 information about products bought in the past is used to derive interests and preferences of costumers

7 `www.chipkarte.at`

8 *Security* - data collectors must take reasonable steps to assure that information collected from consumers is accurate and secure from unauthorized use

9 (malicious) software used by governments to monitor remote computers

New technologies make it possible to gain access to private data, without law enforcement personal actually invading the suspect's home / personal space. Law consultants argue that the computer/electronic device is part of a person's private space. Even mobile devices (e.g. PDAs, laptops, ...), which can easily be used outside the boundaries of a person's private home, should be protected against such invasions by law enforcement (cf. [Bue07b]).

Apart from legal issues, there are also technical issues related to the implementation of govware.

### 7.1.2 Technical Limitations

Studies (cf. [Ste08]) came to the conclusion that the methods used by the law enforcement to monitor (Internet-) activities are rather inefficient.

*Identification by IP Address:* Usage of public terminals or HotSpots is sufficient to avoid identification by IP-address. Open (wireless-)networks used at airports or internet cafés assign IP-addresses only temporally, whereas the (public) IP-address seen by webservers when a request is sent is the same for the whole network. IP-addresses issued to private costumers are likely to be accessed by a rather small amount of people (one single person or people living in one household). On the other hand, IP-addresses issued to commercial users (e.g. internet café, bars with HotSpots) are not useful in case of identifying people. In huge open networks IP-addresses or rather terminals are used by numerous people per day.

*Monitoring of E-Mail:* Monitoring email messages turns out to be a difficult undertaking. When using email managing applications like MS Outlook, email messages are saved on local harddrives. These messages may be accessed, if a connection to the remote computer is established. But many people are using webmail services (like GMX, GMail) to send and receive email messages. Messages and contacts are not saved on the local machine, but on a remote server owned by the email service provider. To access that remote account may require additional authorization or a warrant by the service provider, respectively. Additional encryption of messages would complicate monitoring of email messages even further.

*Monitoring of Telecommunication:* Usage of Voice over IP[10] made it difficult to monitor telephone calls. One reason is the amount of different VoIP applications in use.

Regarding govware, it remains insecure how to infiltrate remote machines (cf. [Bue07a]):

*Exploit Vulnerabilities:* Vulnerabilities found in applications and operating systems are exploited. Disadvantages are the rather limited amount of vulnerabilities and the fact that developers aim to close vulnerabilities as soon as possible. Disposing developers to leave certain discovered vulnerabilities unattended in order to enable govware to infiltrate a machine, could turn out as a difficult as well as a dangerous task.

10 conduct telephone calls over internet protocols

The possibility that developers of software or virus scanners agree to do so is rather unlikely. An unattended vulnerability additionally poses a threat to the machine it was discovered on. The vulnerability could be exploited by govware as well as by attackers.

*Exploit in Operating System:* that approach includes the creation of a special operating system which already includes a purpose-built vulnerability (backdoor[11]), which enables govware to infiltrate the system. If this approach is used, govware could easily be repelled by switching to an alternative operating system. Similar to the attempt described above, this special designed backdoor could also be used by attackers, if information about this access possibility is not kept completely safe.

*Manipulate Internet-Infrastructure:* get users to download/install govware themselves. Requests are redirected to sites, which aim to impose the target site the user wanted to be connected with (similar to *phishing*). These sites could provide links to critical updates to software, while the file also includes govware. The user downloads and installs the file, while the included govware is installed without the user's knowledge. Infiltration by govware through some kind of phishing could easily be warded off by avoiding the download of executable files and constant checking of certificates provided by servers.

*Social Engineering:* This approach works similar to phishing, but govware is attached to emails sent to the user. The user is lured to open the attachment usually by the alleged sender of the email or the message's content. Here, the potential carelessness of the user is the crucial factor.

Every method described above can easily be repelled without the need of technical experience. Additionally, the capabilities of govware are limited even after a connection with the remote machine was established. Requirements for continuous monitoring are:

- (constant) internet connection
- high bandwidth

Users (still) utilizing dial-up connections are not likely to be targeted by prosecutors. Even with high bandwidth at disposal, however, it is not possible to copy whole harddrives (at least not in a reasonable amount of time). While the download bandwidth at user-side could be several MBits, the upload bandwidth is usually low.

Although the implementation of govware is currently not well-engineered, the idea behind is still rather troublesome. Infiltration by govware could easily be avoided without the need to be technically experienced. Thus, criminals and terrorists would probably be clever enough to take precautions. Other users ("normal" citizens), unware of the threats and possibilities offered by the world wide web, would be an easy prey for infiltrations of that kind.

11 part of computer program, often included by developers themselves to enable access to remote systems

After 9/11 the world was in shock. Governments tried to give people some sense of security back by the means of enhanced monitoring of activities (video surveillance, face recognition as well as electronic communications in every form possible). Official statements related to these measures included the prevention of further terror attacks. While some people were troubled by the fact that government and law enforcement seemed to increase the control over people's lives, others embraced the possibilities of new technologies, regarding public exposure. Social networks like MySpace `www.myspace.com`, FaceBook `www.facebook.com` and YouTube `www.youtube.com` help people to "appear in public" by sharing media with other users all over the world.

Several months back a new type of search engine surfaced. These search engines are specially designed to find (contact) information about people.

## 7.2 People Search Engines

At the beginning of 2008 search engines went online, which were especially designed to "find" other people or rather contact information and media related to them. Engines like 123people (`www.123people.com`) and yasni (`www.yasni.com`) work similar to search engines like Google or Yahoo, but 123people and yasni enable people to create own profiles, which work like calling cards.

So people are able to create and manage (private) data that can be viewed by other people. Additionally, engines like 123people search profiles of social networks (e.g. MySpace) for matching keywords.

When searching for a person (given name and/or surname) results from different sources are combined.

Yasni for example claims, that the project aims to strengthen people's right of self-determination (cf. [YAS]). Users are able to overview the extent of the information about their own person on the web and manage the data (to a certain extent). Just anonymous data is stored when a user visits the site:

- date, time
- requested site
- referrer[12]
- browser
- error messages

are saved for logging purposes only. The IP-address of a particular user is only used for delivering the request.

All search results displayed derive from publicly accessible sources. Results from these sources are combined and saved (cached) for a certain amount of time, in order to provide this service to as many users as possible. Resources used for getting results include

12 website (source), which contained the link that redirected the request

- search engines like google, yahoo
- online telephone directories
- social networks (profiles)

If some privacy violation occurs, yasni claims to delete or block that particular content at once.

For creating a profile at one of these search engines, users are asked to disclose information. Entering a valid email address and a password is mandatory, other additional information is optional. For every information in these profiles the user can decide himself, whether the data should be visible to other users. Furthermore, users can add new content to or associate already existing media with the created profile.

Data is only shared with third parties (not associated with the yasni-service), if authorized by the user himself.

## 7.3 (Mis-)use of Web 2.0 Applications

While the following issues may not apply to privacy directly, consequences could affect someone's right to privacy (e.g. law enforcement demands easier access to private data for prosecution purposes).

### 7.3.1 Sharing of private information, knowledge and media

With modern technologies it became much easier to get information about people or certain topics. Online dictionaries like wikipedia (`www.wikipedia.org`) provide easy access to information. These dictionaries, however, are built like social networks, thus many users are able to contribute information. Hence, the accuracy of displayed information may be not guaranteed. Contributions are checked regularly, but the huge amount of data makes it difficult to cover every new article (in a reasonable amount of time). The recent incident of disinformation occurred on January 20 2009 (cf. [WIK09]), when a user (already known for messing up with articles posted on Wikipedia) wrote that Senator Ted Kennedy had recently died. Although the statement was quickly exposed as untrue, the incident had grave consequences on the way new added information is handled in the future by the Wikipedia-Team. A so-called *Flagged Revision System* should decrease the amount of untrue information posted in Wikipedia articles to a minimum. Only information added by trustworthy authors is posted at once. Changes and additions made by "normal users" are checked, before they are accessible publicly. A similar system is already in use since May 2008 at the German domain of Wikipedia. Checking of new additions, however, can delay posting of changes up to three weeks. Jimmy Wales, the founder of the Wikipedia Online Dicitionary, promised that the delay triggered by updates will not be longer than just a few days.

The newest competitor for Wikipedia, Google's *Knol* (`knol.google.com`), on the other hand, only allows clearly identifiable authors to create content, while normal users are still able to contribute suggestions for modification.

Apart from online dictionaries they are also social platforms like YouTube, which enable users to share media with users all over the world. Registered users can easily upload

(self-created) videos while other people can watch and comment on them. With YouTube it is also possible to subscribe to channels in order to be notified when new videos are added to a channel of interest. Similar to Wikipedia, media posted on Youtube is checked, whether by YouTube personal or other users, who are able to "flag" media they perceive as improper or illegal. Hence people are not only able to add content, but can also help in keeping such networks "clean" of unsuitable content.

### 7.3.2 Social Networks and Data Protection

While YouTube is mainly used for sharing videos, networks like MySpace and FaceBook offer a wider spectrum of functionalities. People are able to create profiles and share information like hobbies and other personal information about them. Friends can be added to a list of friends in order to communicate and share media with them more easily. Apart from chatting, people can also play games together.

In 2005, the founder of *Facebook* Mark Zuckerberg, started the project at Harvard University (cf.[EIN09]). Only months later a great part of his fellow students had already created profiles within the network and were sharing information. Nowadays, Facebook counts over 150 million users worldwide and the number of created profiles is increasing weekly by a million members. MySpace, the main competitor of Facebook, which started out as an online data storage, currently counts 230 million members. The difference between those two networks lies mostly in the kind of crowds they attract. MySpace is more considered as a playground for teens and younger people, while Facebook rather appeals university graduants and other adults through a more "serious" theme[13].

The amount of personal information these profiles contain can be very detailed. Especially younger people tend to disclose every little detail about their lives (cf. [VAP08]). The more detailed the data about interests, hobbies, private information is, the easier it is for people to find users with similar interests. It is easy to add friends or be added by friends. The main purpose of Facebook (and MySpace) is to keep in contact with friends (even over great distances) and find new "friends" to chat with. But there are also networks that are designed strictly for purposes related to business. *Xing* (`www.xing.com`), for example, claims to be a *worldwide business platform for professionals, connecting know-how with know-who*. Companies and businesses are able to manage internal processes as well as connections to other businesses.

David Kirkpatrick, the author of "The Facebook Effect", even states that the feature included in most social networks of easy sending messages to all "friends" contained in the list will probably replace emails for distributing short memos and messages (cf. [EIN09])- it is easy to keep your friends and colleagues up to date with your current activities.

The potential to reach millions of people with "just one click" did not remain unnoticed by marketing companies. Profiles are created for the sole purpose of contacting as many users as possible in order to distribute advertisements. These advertisements are often tailored by considering the information (e.g. hobbies, interests, ...) disclosed in user profiles. Terms of use and privacy policies of social networks like Facebook seem to be built for the purpose of distributing tailored ads (cf. [EIN09]):

13 `http://www.webpronews.com/topnews/2007/03/02/youth-prefer-facebook-over-myspace`

- members are encouraged to use their actual name and to refrain from using a pseudonym. Networks like *SchülerVZ* (`www.schuelervz.net`) and *StudiVZ* (`www.studivz.net`), social networks meant to help people to get in contact with (former) fellow students, even prohibit the use of pseudonyms.
- user data stored in profiles is used to tailor advertisements, IP-addresses connected to users are shared with third parties (for following actions and deriving profiles)
- after a user decides to leave the network, user data stored in profiles is not deleted and is used further on (e.g. for marketing)
- if the network is taken over by another company (e.g. Google), the new owner is allowed to access user data
- terms of use can be changed without previous notice

Other risks are related to basic features of most social networks (cf. [PSN08]):

- adding of (detailed) private information into profiles to share with other users
- central data storage to enable easy access
- no or easy to fulfill constraints for the adding of new users to enable fast growth of the network
- relation of data by means of relationships between users in order to generate new data (data mining)

To protect private data operators of these networks should provide options that enable users to decide for themselves, which information is publicly available. Search engines like *123people.com*, however, display information publicly, which is blocked for access within the network itself (cf. SDH08).

In [PSN08] the *Fraunhofer Institut* (SIT) tested the most famous social networks like *Facebook*, *MySpace* for private purposes and the business platforms *XING* and *LinkedIn*. During the tests the testers took the role of a "normal" internet user (user and attacker) to conduct a proper *Black-Box Test*[14].

Criteria for the tests included:

- Data needed to be disclosed for registering an account. Amount of demanded mandatory data should be as little as possible
- Whether or not pseudonyms are allowed. Networks should provide the possibility to use pseudonyms, while avoiding services, which may enable depseudonymizing an individual user
- Is encryption used? Which communication/data is encrypted?

14 provide an input and monitor the resulting output, without knowing the detailed structure of the tested object

- Options to manage access to private data? Which elements (e.g. groups, data entities) are differentiated? Bypassing of security mechanisms possible?
- Default configuration - should provide satisfying security for private information
- External access to multimedia files? Access to multimedia files should only be possible within the platform, not from external destinations (e.g. results generated by search engines), unless user distinctly allowed external access.
- Search engine. Which criteria are available to influence search results. Privacy restrictions still comply with search results?
- Do access protocols exist and how comprehensive are they? Information stored in protocols should not violate privacy of users (registered members and guests).
- Whether and how cancellation of account is possible? Is (private) data stored in account profiles deleted after the cancellation?
- User navigation. Is access to distinct features (e.g. privacy preferences) rather complicated or easy? Does the platform offer adequate user support and documentation?

Test results showed that none of the tested platforms were exactly convincing when it comes to protecting privacy. Best in test was the social network *Facebook*, which offers the most comprehensive options for limiting access to private information. *MySpace* and *StudiVZ*, on the other hand, do offer possibilities to limit access, but they are either not as comprehensive or do not (entirely) affect search results. Access limitations implemented in business platforms are similar to the measures included in their "private" pendants, while *LinkedIn* received better overall test results than the competitor *XING*, because of more thorough deletion of data after account cancellation and possibilities of pseudonymization. Then again *XING* uses better encryption for communication between web browser and social network. In the end [PSN08] provided a list of typical flaws found in social networks:

- mandatory data needed for registering an account is too comprehensive
- pseudonyms are not allowed/supported (e.g. Facebook)
- no or inadequate encryption
- access limitations do comply with certain situations (e.g. search results)
- external access to multimedia files possible
- complicated registration process
- incomplete deletion of account data after cancellation
- complicated user interface or insufficient support

Additionally to the test results the study offered tips useful for users of social networks:

- Do not access social networks while using public WLAN (no or inadequate encryption)

- Structure of business platform is more open. Do not disclose unnecessary private information
- Do not use email addresses in profiles, which are used as pseudonym somewhere else (actions can be linked)
- Review privacy restrictions instantly after registration
- Deactivate public access to files
- Check privacy restrictions before entering new data

Apart from tailored marketing and copyright infringement (see below), new opportunities of communication social networks and similar technologies offer, so-called *Cyber-Mobbing* (cf. [COM09]) snowballed to an important issue. Arguments and name-callings, which were acted out in class or the schoolyard in the past, relocated now to social networks (often created by schools themselves) for everyone visible. EU-Commissioner Viviane Reding demands that it is prohibited for children younger than 13 to join social networks like MySpace. To battle cyber-mobbing SchülerVZ introduced a feature, which enables user to flag offending or aggressive postings. Furthermore, the European *InSafe Initiative* (`www.saferinternet.org`) aims to help *to empower citizens to use the internet, as well as other online technologies, positively, safely and effectively*[15]. InSafe also focuses on an information campaign *empowering teenagers to keep control when online.* It offers solutions to proplems teenagers (between 12 and 17 years old) may face while taking part in online sessions. The video accompanying the campaign aims to encourage teenagers to report abuses.

Owning an account in social networks, however, holds several risks. Although features enable users to control to whom information is visible, mishandling and faulty implementation of online applications pose serious threats to privacy (cf. [VAP08]). Attackers may be able to access information, which should not be publicly available. Information could probably be used against users (even after the data was deleted from the profile itself).

### 7.3.3 Copyright Infringement

With all the possibilities modern technology offers for enabling the contact between people, unfortunately, there is also a great potential for abuse (cf. [Mid05]). Almost everyone with access to the Internet is able to participate in data sharing. Although the owners of social networks and discussion groups aim to ban content and material, which violates the applicable set of rules, it is almost impossible to get rid of content once shared over the Internet. It is easy to copy material and upload it to several web servers at once, where videos or blog entries can simply be re-entered under a different account name or on a different sharing platform.

Activities like these are problematic particularly when illegal content is shared and copyright infringements occur. P2P sharing applications like Napster[16] already gained

15 `http://www.saferinternet.org/ww/en/pub/insafe/about.htm`
16 `http://www.pcworld.com/article/41327/antinapster_ruling_draws_mixed_reaction.html`

notoriety years ago in 2001. Even people who were not interested in technical topics heard about it in the news. Nowadays, video sharing platforms are also used to share (music-) videos and movies that may be protected by copyright. Links to illegal media and software are provided on websites. One site that is shut down because of copyright infringement makes way for several others. Peer to peer (P2P) applications are often used to distribute legal media as well as material protected by copyright. Several sites provide links (e.g. torrents used with the BitTorrent protocol) in structured databases and search features concerning material of interest. The shutdown of sites which provide direct links is equally difficult as the effective prohibition of sites which reprovide offerings.

Like already mentioned in this paper before, the question whether ISPs are responsible for content added by their customers (or third-parties) has not been clearly answered. Under current US and European law ISPs are not the originators of the content provided and thus are regarded as the "conduits". Hence, ISPs are currently protected against copyright infringement law suits. The laws, however, which provide juridical protection, were already introduced before platforms like MySpace, Facebook and YouTube made it even easier for people to upload and share content (which may violate copyright). Because of the increasing number of copyright violating material hosted on content provider networks, the question of liability became pressing again. But this time content like services provided or owned by Google are in the focus of interest. In 2007, the media company VIACOM sued Google (YouTube) because of "massive intentional copyright infringement" and YouTube had to disclose private information about their users (cf. [UOY08]). The decision also provided a more or less detailed insight in how YouTube is monitoring the users of their service. The most recent incident regarding copyright infringement and liability (cf. [PIS09]) could be precedent-setting for future cases (begin of trial assigned February 18 2009 in Milan, Italy). Italian officials do not intend to sue the Google company as a whole, but three individuals (executives) were picked out as responsible persons. The case relates to a short (offending) video, which was uploaded on the Italian domain of Google Video in 2006. Although Google deleted the video after receiving two complaints within 24 hours and the "stars" of the video were identified and punished (because of the video itself), Italian officials still argue that *the video should never have been allowed to be uploaded in the first place*[PIS09]. Because of the incident with VIACOM YouTube had already promised to implement means to monitor the content added, but obviously nothing changed to-date. The Milan case, on the other hand, could probably have even graver consequences, because prosecutors argue that companies like Google should be held responsible for all content made available by their services, even if the company (or individuals) is not aware of it (e.g. content that can be accessed through results found by search engines). If it is decided that companies (individuals) can indeed be held responsible for content added by third parties, then this could be regarded as *a warning shot for other companies' executives and their liabilities toward privacy violations* [PIS09].

### 7.3.4 Easy way to reach the general Public

The world wide web provides access to media and information as well as the possibility to voice opinions and make statements, thereby reaching thousands of people. Newspapers and news broadcasts are updated daily respectively several times a day. Online news

portals (e.g. online versions of news broadcasting stations and news papers) are updated almost at minute intervals, whereas all users are able to access the portal. The ability to reach far more people, however, implies much more responsibility to distribute well researched information and not to misuse the portal for distributing rumors or propaganda.

On the other hand, people with access to the internet are able to create their own textual content and share it with others. So-called blogs work similar to old fashioned diaries and enable users to write about certain topics or comment on things. Interfaces with other networks provide the possibility to add media. People can write about any topic as long as the content does not violate applicable laws and rules.

The election campaign for the US election 2008 was one of the first elections where the possibilities of social networks were used to a great extent. Younger generations, who are allowed to vote for the first time, are easier to reach with modern technologies like Internet and cell phones than with "old" medias like TV or radio. Both political parties used social networks like YouTube to distribute advertisement spots and to voice their manifestos. The potential for misuse is always present. Opposing parties could use the opportunity to discredit or influence public opinion.

Because the Internet does not only contain information about certain topics, but also individuals it is not uncommon anymore that people "google" each other (cf. [DSS08]). Nearly 30% of all search queries are sent to find out more about a certain individual. Simply by entering a name in the search field of a search engine results in different kinds of information about a person (of the same name). Some search engines are strictly designed for the easy access to information relating to individuals. Engines like *123people.com* make use of other search engines (e.g. Google) and combine the information with data found in social networks (e.g. MySpace). Although owners of these engines argue that only public available data is shown, the fact that the information can be accessed at one glance may pose a problem. With *123people.com* it is even possible to tag and attach (possibly offending) terms to individuals (cf. [DSS08]). Even though people are able to delete unrequested tags (after registering an account), some information may remain on the database.

An imprudent comment added years ago, a photo taken in a disadvantageous moment or even outdated information could inflict serious harm on a person's reputation. Not only future date partners are "googeled" nowadays, but also people who apply for a job (cf. [DSS08]). To avoid such nasty surprises, which could seriously harm one's reputation, certain companies provide services to delete unwanted information (cf. [SDH08]). Companies like *Reputation Defender* (`www.reputationdefender.com`) and *Datenwachschutz* (`www.datenwachschutz.de`) claim to erase undesired information and charge a fee per deleted entry.

### 7.3.5 Al Qaeda's Online Offensive

The possibilities new technologies offer could also be used to influence public opinion, if the content reaches enough people. Since 9/11, technology advanced. It did not take long until illegal and terror organizations got aware of the new possibilities social networks and blogs offer. Due to the ease and the swiftness of website creation, it becomes difficult to supervise the content of these sites. Similar to other illegal and criminal activities, which may harm economy or society itself, the distribution of propaganda over the Internet

became a serious issue.

Terror organizations like Al Qaeda

> are skillfully using the Internet to raise money, recruit and train members, and to spread a message of hatred

(cf. [Smi06]). Terrorists, especially trained to use modern technologies, do not only create media (e.g. images, videos of assaults), but manipulate already existing media (cf. [Smi06]).

Excluding propaganda videos from social networks like YouTube is only a temporally effective measure. Videos are quickly published and could easily be re-entered or uploaded on another server/site.

Distribution of propaganda media gets better organized with time (cf. [Whi08]) and is currently a quite secure network. Al Qaeda uses (password protected) webforums to announce new media (propaganda videos containing e.g. speeches of high ranked Al Qaeda executives or Osama Bin Laden himself).

The terror network communicates securely, whereas new videos are released regularly (in average every three or four days). The main node for propaganda distributions is called *as-Sahab*. Productions include videos suitable for displaying on cell phones and iPods. This kind of propaganda is intended for potential recruits residing in western countries, who are easily influenced by the distributed messages. New mobile media like cell phones, PDA's and iPods/iPhones with the option of playing music/videos and a connection to the WWW, are the current targets of choice, when trying to reach young people. New technologies, besides being a status symbol, made media content even more mobile compared to notebooks.

In 2004, when Al Qaeda started the online offensive, the US government underestimated the power of the propaganda distribution network. At the beginning the network did not function properly, but has evolved in the course of time. Nowadays, the network is nearly invulnerable [Whi08].

Al Qaeda has access to high end technology. A video is recorded directly at the "hideout". Modern technologies and mobile notebooks are used to create the file, which is then usually transported on a memory stick and protected with PGP. Some time later the file is uploaded to a (password protected) webserver - usually at an Internet café, to leave the least electronic traces possible. When using public networks for accessing the Internet, the IP-address, which may be logged with the requests from a terminal, cannot be related back to a particular person (see above). A manual called "Technical Mujahid" explains how to avoid electronic footprints. Later the videos are professionally edited and combined with subtitles in numerous languages (including Italian, German and Spanish).

Although Al Qaeda is more or less successful with recruiting new footsoldiers with propaganda, exposure "stars" (e.g. Ayman al-Zawahiri) in these videos could eventually be captured. Some terrorists had already been hunted down after they appeared in such videos. High executives appear more often in those videos than others. They are usually better suited to transport the propaganda message to the public and are regarded as leaders by new recruits.

# CHAPTER 8

## Conclusion

With easier and cheaper possibilities to access the data highway, not only the number of people surfing online increased, but also the opportunities for attackers, criminals (and after the incidents of 9/11, importantly) terrorists to abuse modern technologies. Many people (mostly casual users) accessing the internet are not even aware of the risks they face and the tools (PETs) that exist to improve privacy and security of sensitive data, whose disclosure could be embarrassing as well as harmful.

The topic of this thesis focused on the risks and PETs, and took a closer look at privacy, data protection in general and how things changed over time. The definition of privacy included in the *US Constitution* was and still is rather vague, but shortly after WW2 the *Declaration of Human Rights* confirmed the right that no one *should be subject to arbitrary interference between privacy.* Already in the 1960's first mechanisms existed to protect (sensitive) data stored on drives from unauthorized access.

At the end of the 20th century, the local networks connecting clients within one single company were replaced with one huge network spanning over the planet, making enormous amounts of information more or less publicly available. The incidents on 9/11 changed the attitude toward privacy and data protection in the US as well as Europe and the rest of the world. Laws, adopted and changed shortly after the attacks, claimed to improve (national) security in order to avert future attacks. Critiques argued that these changes limit the personal freedom without any chance to prevent future terrorism.

New adopted laws are not the only risks to privacy and data protection, because the WWW is not only used for the easy access of information. Thanks to online-dictionaries like *Wikipedia*, but every request made also discloses probably sensitive data, which may be abused. People need to enter more or less private information in order to use services provided online. Webshops (e.g. Amazon) need very sensitive data like residential address or creditcard number to conclude transactions and deliver ordered products properly. Other webservers need at least login-data and some information for profile and preferences.

Data that is entered by the user himself/herself is sent with every request made, via the *IP-address* issued by the ISP. Hence the IP-address can be used to identify a user and link several different actions (requests) back to one particular user. Another risk to one's privacy are *Cookies*, which are stored on local drives of users. They are used by many webservers, but for different reasons. One purpose for the use of cookies is to "recognize" revisiting users and to save individual preferences. But cookies are also used for more

dubious intentions. On the hard drive stored cookies are like a roadmap of websites visited in the past, where every cookie-file contains additional information. So-called *third-party* cookies are distinctly used for a purpose similar to IP-addresses, to trace back visited websites. These information is then used to derive profiles of interests (*data mining*) to issue tailored advertisements (controlled by marketing companies).

*Spyware* (e.g. trojans), on the other hand, endangers sensitive information not disclosed on the Internet, but data stored on local hard drives. Spyware attackers are not only able to access and alter sensitive information (e.g. bank account information), but send it wherever they want. Occasions during which malicious software could end up on a user's hard drive are versatile. The software can be hidden in occasional files, free or even commercial software or included in attachments sent via email.

Organizations like the *Internet Service Provider* themselves can pose a threat to the privacy of their customers, because they naturally store private information and banking information related to the IP-addresses visible in the WWW. *Google*, once the owner of a search-engine, quickly rose to a provider of numerous services. A big part of these services store data disclosed to them. The information includes keywords, private facts or codes uploaded on servers owned by Google (e.g. Google Code). *AdWords* or *AdSense* enables Google to display advertisement related to keywords or a websites content. *Gmail* parses the content of sent and received emails for exactly the same reason. Data collected with these numerous services owned by one single company could again be easily used for data mining. *Google Trends* makes statistics derived from keywords searched for with the Google search engine publicly accessible. With *social engineering* (SC) there exists a risk to privacy, which rather adresses the human component instead of taking the (fully) technological approach. Attackers using SC aim to trick people to disclose sensitive information, where *phishing* represents a more technological version.

With all the risks and threats to one's privacy it became necessary to develop tools, which aim to improve the protection of sensitive data. The *Privacy Enhancing Technologies* introduced in this thesis provide different levels of anonymity, privacy and security. Anonymizers mask IP-addresses to provide anonymous participating in online sessions and some tools also advise on additional measures to diminish other risks of identification (e.g. disabling of cookies, script-languages). Because many webservices cannot be used by anonymous users, *pseudonymity* is necessary. Partial identities (at least login-information like username and password) are used to create accounts and participate in online discussions with a chosen pseudonym. The most important PET is *encryption*. It secures the data itself by encrypting it in order make it unreadable unless the corresponding decryption-key is used. This measure secures local and remotely stored information as well as data during transmission (e.g. SSH).

All of the described PETs do not provide a complete protection of privacy, but rather help users to decide which data is disclosed automatically or which information they want to disclose themselves (cf. [Tav01]). There still exist dangers PETs are not able to protect against yet. One of these dangers is related to sensitive information disclosure in messages posted on (public accessible) discussion boards (content-based anonymity). Hence these boards allow people to conduct industrial espionage, if members happen to talk about current developments in their company, (virtual) stalking or profiling (carried out by marketing companies).

Mentioned "malicious" software and methods like data mining may not only be used

by hackers or marketing companies, but also (legally) utilized by law enforcement and government agencies for prosecution and even monitoring (e.g. computer surveillance and video surveillance with included face recognition). *Web 2.0 networks*, which claim that everyone can add to the (publicly available) content, raise privacy issues too. It is difficult, almost impossible to monitor all information that is made available and shared with others. Data shared can include illegal material like propaganda, copyright-protected media and systematic disinformation. Questionable material is especially problematic for networks, which have to be legitimate and comprehensive. To address the problem of disinformation and the adding of improper material Wikipedia plans to implement tools, which should make it easier to check added information. Another new trend introduced by Web 2.0, called *social networks*, raise additional privacy issues. Networks like MySpace enable members to create rather comprehensive and detailed profiles about themselves. The information included in profiles and the public accessible communication are a target for advertising companies, who aim to derive dossiers from found data.

This thesis took a closer look at the past and current situation of privacy and data protection and especially focused on the changes after 9/11. The introduction of threats to privacy and sensitive data related to the WWW, hopefully will help to raise awareness of the dangers the online world contains (especially for unexperienced users). Sections concerning PETs showed that these tools do not provide complete protection against privacy loss, but rather helps people to keep control of what kind of data they disclose.

# Bibliography


[ANA] Analysis of Provisions of the Proposed Anti-Terrorism Act of 2001, Internet, http://www.epic.org/privacy/terrorism/usapatriot/RL31377.pdf (Cited on page 13)

[AUS06] Wehrhafte Demokratie oder Gesinnungsterror?, Internet (2006), http://web.archive.org/web/20071008125126/http://www.politische-bildung-brandenburg.de/extrem/wehrhaft3.htm (Cited on page 11)

[Ban08] Banisar, David: Speaking of Terror (2008), http://www.privacyinternational.org/issues/terrorism/speakingofterror.pdf (Cited on pages 9 and 10)

[BEI04] Bei jeder Mail wird mitgelesen. *heise.de* (2004), http://www.heise.de/tp/r4/artikel/17/17108/1.html (Cited on page 36)

[Ben99] Benassi, Paola: TRUSTe: an online privacy seal program. *Commun. ACM* (1999), Vol. 42(2):p. 56–59 (Cited on page 73)

[BOR] Bill of Rights Transcript, Internet, http://www.archives.gov/exhibits/charters/bill_of_rights_transcript.html (Cited on page 111)

[BP05] Borcea-Pfitzmann, Katrin; Franz, Elke and Pfitzmann, Andreas: Usable presentation of secure pseudonyms, in: *DIM '05: Proceedings of the 2005 workshop on Digital identity management*, ACM, New York, NY, USA, p. 70–76 (Cited on pages 61, 62, 64, 98, and 103)

[Bue07a] Buermeyer, U.: Die "Online-Durchsuchung". Technischer Hintergrund des verdeckten hoheitlichen Zugriffs auf Computersysteme, Internet (2007), http://www.hrr-strafrecht.de/hrr/archiv/07-04/index.php?sz=8 (Cited on page 115)

[Bue07b] Buermeyer, U.: Die "Online-Durchsuchung". Verfassungsrechtliche Grenzen des verdeckten hoheitlichen Zugriffs auf Computersysteme, Internet (2007), http://www.hrr-strafrecht.de/hrr/archiv/07-08/index.php?sz=7 (Cited on page 115)

[Cam05] Camenisch, Jan; abhi shelat; Sommer, Dieter; Fischer-Hübner, Simone; Hansen, Marit; Krasemann, Henry; Lacoste, Gérard; Leenes, Ronald and

Tseng, Jimmy: Privacy and identity management for everyone, in: *DIM '05: Proceedings of the 2005 workshop on Digital identity management*, ACM, New York, NY, USA, p. 20–27 (Cited on pages 3, 74, 75, 79, 85, 87, and 88)

[CIT] The Austrian Citizen Card, Internet, `http://www.buergerkarte.at/en/index.html` (Cited on page 106)

[Cla99] Clarke, Roger: Internet privacy concerns confirm the case for intervention. *Commun. ACM* (1999), Vol. 42(2):p. 60–67 (Cited on page 100)

[COM09] Computerspiele über dem Limit. *Kurier* (2009) (Cited on page 122)

[COO08] Cooke Central, Internet (2008), `http://www.cookiecentral.com/c_concept.htm` (Cited on pages 23 and 24)

[Cra99] Cranor, Lorrie Faith: Internet privacy. *Commun. ACM* (1999), Vol. 42(2):p. 28–38 (Cited on page 3)

[Cra02] Cranor, Lorrie Faith; Arjula, Manjula and Guduru, Praveen: Use of a P3P user agent by early adopters, in: *WPES '02: Proceedings of the 2002 ACM workshop on Privacy in the Electronic Society*, ACM, New York, NY, USA, p. 1–10 (Cited on pages 72 and 73)

[DDD09] Die Datenschnüffler, die wir lieben. *Kurier* (2009) (Cited on page 112)

[DRW] Dr. Web: Google, Internet, `http://www.drweb.de/magazin/tag/Google/` (Cited on page 34)

[DSS08] Die Schnüffel-Suchmaschine. *Kurier* (2008) (Cited on page 124)

[ECH] ECHELON @ GlobalSecurity.org, Internet, `http://www.globalsecurity.org/intell/systems/echelon.htm` (Cited on page 47)

[ECK] Eckpunkte zum Sicherheitspaket II, Internet, `http://www.neue-einheit.com/mixed/terror/schily-entwurf.htm` (Cited on pages 10 and 11)

[EIN09] Ein mächtiger Teil der Internet-Welt. *Kurier* (2009) (Cited on pages 100, 101, and 119)

[EUR01] REPORT on the existence of a global system for the interception of private and commercial communications (ECHELON interception system), Internet (2001), `http://www.fas.org/irp/program/process/rapport_echelon_en.pdf` (Cited on page 48)

[EXP04] Explanation of the web of trust of PGP, Internet (2004), `http://www.rubin.ch/pgp/weboftrust.en.html` (Cited on page 70)

[Fox00] Fox, S.: Trust and Privacy online: Why Americans want to rewrite the Rules (2000), `http://www.pewinternet.org/pdfs/PIP_Trust_Privacy_Report.pdf` (Cited on page 23)

[FP04] F. Poppa, S. Witschel and Buchwald, S.: Die Auswirkungen des 11. September 2001 auf die amerikanische und deutsche Gesetzgebung im Hinblick auf den Datenschutz, Internet (2004), http://www.rz.fh-ulm.de/projects/datsusi/R0304_5.html (Cited on pages 11, 13, 17, 18, and 19)

[Fre05] Freeman, Lee A. and Urbaczewski, Andrew: Why do people hate spyware? *Commun. ACM* (2005), Vol. 48(8):p. 50–53 (Cited on pages 24, 28, 29, and 136)

[Ful04] Fule, Peter and Roddick, John F.: Detecting privacy and ethical sensitivity in data mining results, in: *ACSC '04: Proceedings of the 27th Australasian conference on Computer science*, Australian Computer Society, Inc., Darlinghurst, Australia, Australia, p. 159–166 (Cited on pages 41 and 42)

[Gab99] Gabber, Eran; Gibbons, Phillip B.; Kristol, David M.; Matias, Yossi and Mayer, Alain: Consistent, yet anonymous, Web access with LPWA. *Commun. ACM* (1999), Vol. 42(2):p. 42–47 (Cited on pages 66 and 67)

[GES] Gesamte Rechtsvorschrift für Sicherheitspolizeigesetz, Fassung vom 17.01.2009, Internet, http://www.ris2.bka.gv.at/GeltendeFassung.wxe?QueryID=Bundesnormen&Gesetzesnummer=10005792 (Cited on page 12)

[GHS] Guru's Heaven Sicherheitscheck, Internet, http://www.gurusheaven.de/index_sicherheit.html (Cited on page 108)

[GMA] Gmail is too creepy, Internet, http://gmail-is-too-creepy.com/ (Cited on page 36)

[Gol99] Goldschlag, David; Reed, Michael and Syverson, Paul: Onion routing. *Commun. ACM* (1999), Vol. 42(2):p. 39–41 (Cited on page 54)

[GOO04] Google gets the Message, launches Gmail. *Google Press Center* (2004), Vol. April 1, http://www.google.com/press/pressrel/gmail.html (Cited on page 36)

[GRO09] Großbritannien führt totale E-Mail-Überwachung ein?, Internet (2009), http://www.pressetext.at/pte.mc?pte=090109019&source=rss_2 (Cited on page 113)

[Hen00] Henry, Kevin: Getting started with PGP. *Crossroads* (2000), Vol. 6(5):p. 8 (Cited on page 69)

[Hof69] Hoffman, Lance J.: Computers and Privacy: A Survey. *ACM Comput. Surv.* (1969), Vol. 1(2):p. 85–103 (Cited on pages 6, 7, 8, 20, and 138)

[Jø07] J, Audun; Zomai, Muhammed Al and Suriadi, Suriadi: Usability and privacy in identity management architectures, in: *ACSW '07: Proceedings of the fifth Australasian symposium on ACSW frontiers*, Australian Computer Society, Inc., Darlinghurst, Australia, Australia, p. 143–152 (Cited on pages 6 and 83)

[Kan04] Kantarcioğlu, Murat; Jin, Jiashun and Clifton, Chris: When do data mining results violate privacy?, in: *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, New York, NY, USA, p. 599–604 (Cited on pages 41 and 42)

[Kob03] Kobsa, Alfred and Schreck, Jörg: Privacy through pseudonymity in user-adaptive systems. *ACM Trans. Interet Technol.* (2003), Vol. 3(2):p. 149–183 (Cited on pages 51, 58, 59, 61, 97, 103, and 104)

[LIB] Liberty Alliance, Internet, `http://www.projectliberty.org/` (Cited on page 89)

[Lin07] Lindell, Andrew Y.: Authentication Online. *BH-US-07* (2007), `https://www.blackhat.com/presentations/bh-usa-07/Lindell/Whitepaper/bh-usa-07-lindell-WP.pdf` (Cited on pages 4, 5, 9, 52, and 103)

[MH05] Müller-Heidelberg, T.: Das Terrorismusbekämpfungsgesetz - Ein Erfolg der Terroristen (2005), `http://www.bpb.de/files/Q7E6IU.pdf` (Cited on pages 10, 11, 12, 17, 18, 19, 20, and 21)

[Mid05] Middleton, Gaye L.: Copyright conundrum: liability of ISPs for online copyright infringement, in: *ACSW Frontiers '05: Proceedings of the 2005 Australasian workshop on Grid computing and e-research*, Australian Computer Society, Inc., Darlinghurst, Australia, Australia, p. 109–118 (Cited on pages 32, 33, and 122)

[MM07] M. Mannan, Marit Hanson, P.C. von Oorshot: Using a Personal Device to Strengthen Password Authentication from an Untrusted Computer (2007), `http://www.ccsl.carleton.ca/paper-archive/mpauth.pdf` (Cited on page 107)

[Oli02] Olivier, M. S.: Database Privacy. *SIGKDD Explorations* (2002), Vol. 4(2):p. 20 – 27, `http://www.sigkdd.org/explorations/issue4-2/olivier.pdf` (Cited on page 92)

[OVE] The USA Patriot Act, Internet, `http://epic.org/privacy/terrorism/usapatriot/default.html` (Cited on page 13)

[Par68] Parker, Donn B.: Rules of ethics in information processing. *Commun. ACM* (1968), Vol. 11(3):p. 198–201 (Cited on pages 2 and 7)

[PCF] PC Fear Factor: The Ultimate PC Disaster Prevention Guide The Official Gator Spyware Page, Internet, `http://alanluber.com/pcfearfactor/officiagatorpage.htm` (Cited on page 30)

[Pes05] Peslak, Alan R.: Privacy policies of the largest privately held companies: a review and analysis of the forbes private 50, in: *SIGMIS CPR '05: Proceedings of the 2005 ACM SIGMIS CPR conference on Computer personnel research*, ACM, New York, NY, USA, p. 104–111 (Cited on pages 2, 4, 5, and 9)

[PEU05] Terrorism Profile EU, Internet (2005), `http://www.privacyinternational.org/article.shtml?cmd[347]=x-347-508362&als[theme]=Anti\%20Terrorism\%20Home\%20Page` (Cited on page 15)

[PIS09] Privacy, Italian Style, Internet (2009), http://www.spiegel.de/international/business/0,1518,606617,00.html (Cited on page 123)

[PSN08] Privatsphärenschutz in Soziale-Netzwerke-Plattformen (2008), http://www.sit.fraunhofer.de/fhg/Images/SocNetStudie_Deu_Final_tcm105-132111.pdf (Cited on pages 101, 120, and 121)

[PTA] THE PRIVACY ACT OF 1974 5 U.S.C. § 552a, Internet, http://www.usdoj.gov/oip/privstat.htm (Cited on page 7)

[PUS05] Terrorism Profile US, Internet (2005), http://www.privacyinternational.org/article.shtml?cmd[347]=x-347-508361&als[theme]=Anti\%20Terrorism\%20Home\%20Page (Cited on pages 15 and 17)

[Rea99] Reagle, Joseph and Cranor, Lorrie Faith: The platform for privacy preferences. *Commun. ACM* (1999), Vol. 42(2):p. 48–55 (Cited on pages 71 and 98)

[Rei99] Reiter, Michael K. and Rubin, Aviel D.: Anonymous Web transactions with Crowds. *Commun. ACM* (1999), Vol. 42(2):p. 32–48 (Cited on pages 22, 56, 57, and 136)

[RL08] R. Leens, Marit Hanson, J. Schallaböck: PRIME White Paper (2008), https://www.prime-project.eu/prime_products/whitepaper/PRIME-Whitepaper-V3.pdf (Cited on pages 83 and 105)

[Sch03] Schreck, J.: *Security and Privacy in User Modeling*, Springer (2003), http://books.google.at/books?id=ikh7jAcVkOMC&pg=PA68&lpg=PA68&dq=content+based+anonymity&source=bl&ots=y-YcJRcKMn&sig=mkHEvs3E7OxTg7r2i0jkXuxUAgs&hl=de&sa=X&oi=book_result&resnum=1&ct=result#PPA68,M1 (Cited on page 103)

[SDH08] Striptease am Daten-Highway. *Kurier* (2008) (Cited on pages 101 and 124)

[Sei03] Seidl, B.: *Data Mining und Datenschutz*, Masterthesis, RECHTSWISSENSCHAFTLICHE FAKULTÄT DER UNIVERSITÄT WIEN (2003), http://www.rechtsprobleme.at/doks/datamining_datenschutz_seidl.pdf (Cited on pages 39 and 40)

[Sip05] Sipior, Janice C.; Ward, Burke T. and Roselli, Georgina R.: A United States perspective on the ethical and legal issues of spyware, in: *ICEC '05: Proceedings of the 7th international conference on Electronic commerce*, ACM, New York, NY, USA, p. 738–743 (Cited on pages 24, 26, 27, 90, and 91)

[Smi06] Smith, S. D.: Terrorists Use Internet for Propaganda, Internet (2006), http://www.defenselink.mil/news/newsarticle.aspx?id=15854 (Cited on page 125)

[SOC] Social Engineering, Internet, http://social.christian-grafe.de/Social\%20Engineering\%20-\%20Christian\%20Grafe.doc (Cited on pages 43, 44, 45, 91, and 92)

[SOC01] Social Engineering Fundamentals, Part I: Hacker Tactics. *Security Focus* (2001), Vol. December, http://www.securityfocus.com/infocus/1527 (Cited on pages 43, 44, and 46)

[Ste08] Sterbenz, B.: Die Cybercops sind zahnlos. *Kurier* (2008):p. 25 (Cited on page 115)

[SUR08] Surfen auf der Google-Welle. *Kurier* (2008) (Cited on page 35)

[Swe05] Sweeney, L.: Privacy-Preserving Surveillance Using Selective Revelation. *IEEE Intelligent Systems* (2005), http://privacy.cs.cmu.edu/dataprivacy/projects/selectiverevelation/pps.pdf (Cited on pages 92 and 93)

[Swe06] Sweeney, L.: Protecting Job Seekers from Identity Theft. *IEEE Internet Computing* (2006), Vol. March-April, http://privacy.cs.cmu.edu/dataprivacy/projects/idangel/idangel3.pdf (Cited on pages 47 and 93)

[Tav01] Tavani, Herman T. and Moor, James H.: Privacy protection, control of information, and privacy-enhancing technologies. *SIGCAS Comput. Soc.* (2001), Vol. 31(1):p. 6–11 (Cited on pages 99, 100, 102, 108, and 127)

[TFS04] T. F. Stafford, A. Urbaczewski: Spyware: The Ghost in the Machine. *Communication of the Association for Information Systems* (2004), Vol. 14:p. 291–306 (Cited on pages 1, 26, 29, 91, and 109)

[Tho04] Thornburgh, Tim: Social engineering: the "Dark Art", in: *InfoSecCD '04: Proceedings of the 1st annual conference on Information security curriculum development*, ACM, New York, NY, USA, p. 133–135 (Cited on pages 43 and 45)

[TOR] TOR Project, Internet, https://www.torproject.org/ (Cited on page 54)

[UOY08] Urteil outet YouTube als Datenkraken, Internet (2008), http://www.spiegel.de/netzwelt/web/0,1518,563821,00.html (Cited on page 123)

[Vai02] Vaidya, Jaideep and Clifton, Chris: Privacy preserving association rule mining in vertically partitioned data, in: *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, New York, NY, USA, p. 639–644 (Cited on page 41)

[VAP08] Verlust an Privatsphäre durch die Social Network Sites, Internet (2008), http://sicherheitskultur.at/privacy_soc_networking.htm (Cited on pages 112, 119, and 122)

[VER] Verzeichnisschutz/Passwortschutz mit htaccess sowie SSI/PHP, Internet, http://www.kortstock.de/WWW-Kurs/verzeichnisschutz/htaccess.html (Cited on page 106)

[Wal01] Walters, Gregory J.: Privacy and security: an ethical analysis. *SIGCAS Comput. Soc.* (2001), Vol. 31(2):p. 8–23 (Cited on pages 4 and 50)

[WAS07] Was Google nicht zeigt: Das zensierte Weltauge. *Spiegel Online* (2007), http://www.spiegel.de/netzwelt/web/0,1518,464186,00.html (Cited on page 38)

[WEL08] Weltmacht Google. *Stern Online* (2008), http://www.stern.de/computer-technik/561422.html (Cited on pages 34 and 38)

[Whi08] Whitlock, C.: Al-Qaeda's Growing Online Offensive, Internet (2008), http://www.washingtonpost.com/wp-dyn/content/article/2008/06/23/AR2008062302135.html (Cited on page 125)

[WIK09] Wikipedia steigt auf die Bremse. *Kurier* (2009) (Cited on page 118)

[WSF] WS Federation Specification, Internet, http://download.boulder.ibm.com/ibmdl/pub/software/dw/specs/ws-fed/WS-Federation-V1-1B.pdf (Cited on page 89)

[YAS] Yasni Privacy Policy, Internet, http://www.yasni.com/index.php?action=yasni_privacy (Cited on page 117)

# List of Figures

# List of Tables

# Acknowledgements

An dieser Stelle möchte ich mich bei meinem Betreuer, Herrn Prof. Dipl.-Ing. Mag. Dr. Thomas Neubauer, für die Unterstützung und kompetente Betreuung während der Erstellung meiner Masterarbeit bedanken.
Er hat damit wesentlich zum Gelingen meiner wissenschaftlichen Arbeit beigetragen.

## Zusammenfassung

Neue Technologien ermöglichen immer mehr Menschen den Zugang zum Internet. Innerhalb kürzester Zeit können Informationen abgerufen werden. E-Mail oder Instant-Messenger ermöglichen es, Kontakt mit Familie, Freunden und Kollegen selbst bei sehr großen physischen Entfernungen nicht abreißen zu lassen. Selbst das Finden und Kaufen von Produkten wird durch Webshops erleichtert. Services wie diese, die über das WWW genutzt werden, erfordern allerdings auch Informationen. Webshops benötigen personenbezogene Daten, wie die postalische Adresse des Kunden oder dessen Kreditkartennummer, die auch entsprechend vor unerwünschtem Zugriff (z.B. durch Hacker) geschützt werden müssen.

Allerdings sind nicht nur private und personenbezogene Daten gespeichert auf Unternehmens-Servern in potentieller Gefahr, sondern auch Informationen, die "unsichtbar" bei jeder Anfrage mitgesendet werden (z.B. IP-Adresse) oder sich auf der Festplatte des Heimrechners befinden (z.B. Zugangsdaten für Online-Banking).

Der erste Abschnitt dieser Masterarbeit beschäftigt sich jedoch mit der allgemeinen Definition von Privatsphäre in Zusammenhang mit Datenschutz und wie sich der Zugang zu diesem Thema nach den Ereignissen des 11. September 2001 änderte. Die Hauptaufgabe dieser Arbeit liegt allerdings darin, auf zweierlei Dinge aufmerksam zu machen: Einerseits die Risiken, denen sensible Daten durch moderne Technologien (z.B. Internet) ausgesetzt sind und andererseits die Schutzmaßnahmen (Privacy Enhancing Technologies), die im Grunde darauf abzielen, eben diese Daten vor unerlaubtem Zugriff zu schützen.

Des Weiteren wird abschließend noch auf zu meisternde Hürden im Zusammenhang mit Datenschutz und Schutzmaßnahmen hingewiesen und aktuelle Probleme, die Privatsphäre (z.B. Soziale Netzwerke) betreffend, besprochen.

## Curriculum Vitae

**Persönliche Daten:**

**Name:** Barbara Lang, Bakk.
**Geburtsdatum, -ort:** 18. Juni 1983, St. Pölten
**Staatsangehörigkeit:** Österreich

**Ausbildung und Weiterbildung:**

TU Wien - Uni Wien
**Oktober 2002:**
Studium Bakkalaureat Wirtschaftsinformatik
Schwerpunkt Projekt- und Qualitätsmanagement
**Oktober 2006:**
Studium Master Wirtschaftsinformatik
Schwerpunkt Projekt- und Qualitätsmanagement
Wahlfach (Internet-) Security
Studium Master Informatikmanagement

**Besondere Kenntnisse:**

**Sprachkenntnisse:**
Deutsch: Muttersprache
Englisch
**EDV-Kenntnisse:**
ECDL-Computerführerschein
Programmiersprachen: Java, PHP