Why Do Privacy-Enhancement Mechanisms Fail, After All? A Survey of Both the User and the Provider Perspective

Thorben Burghardt
Universität Karlsruhe (TH)
P.O. Box 6980
76128 Karlsruhe
+49 (721) 608-3968
burgthor@ipd.uka.de

Erik Buchmann
Universität Karlsruhe (TH)
P.O. Box 6980
76128 Karlsruhe
+49 (721) 608-3968
buchmann@ipd.uka.de

Klemens Böhm
Universität Karlsruhe (TH)
P.O. Box 6980
76128 Karlsruhe
+49 (721) 608-3968
boehm@ipd.uka.de

ABSTRACT

Web2.0 applications have become an inherent part of everyday life, with unpredictable consequences for the privacy of individuals. People communicate via social network sites and participate in the life of others by commenting blogs or tagging web sites, images or videos. Specialized search engines can assemble this information to comprehensive personality profiles, and platform providers have in-depth knowledge of their customers. This is problematic, as data protection is an important factor for the users to establish trust in the platform provider. Without it, Web2.0 services will hardly attract an active community.

Our objective is to identify fundamental problems for both classical and Web2.0 providers when handling personal information of the users. Therefore, we explore the challenges that impact the efficiency of current legal and technical mechanisms for data protection. Today's Internet provides many reasons for such challenges, e.g., legislation issues, economic interests or the particular behavior of Internet users. Data protection depends on the data collectors as well as on the individuals who reveal personal information. Thus, we will analyze online interactions both from the perspective of the user and the provider. We motivate and validate various hypotheses which reflect that users are unaware of the impact of uncontrolled information disclosure. From the other perspective, platform providers often implement information collection and forwarding processes that are not transparent to the user. This even holds in the presence of a privacy policy. We validate our hypotheses using different methods, which we motivate and explain. Our findings are significant: For example, skilled Internet users typically forget about half of their registrations. Thus, the efficiency of all legal and technical means for data protection is questionable if they require the awareness of the individual. Finally, we derive requirements for future privacy-enhancement technologies.

1. INTRODUCTION

Currently, the Internet challenges the privacy of almost any member of the society. While Web1.0 and Web2.0 platforms offer convenient options for governmental services, shopping, entertainment, social life etc., all of these services include the disclosure of personal information, and the consequences for privacy are unpredictable. Traditional platforms, e.g., web shops, possess in-depth knowledge of their customers, which puts them at risk to be manipulated. On the other hand, anecdotic evidence shows that private information revealed in Web2.0 sites like Facebook or MySpace can support online predators or might result in identity theft and cyberstalking.

Trusting the platform provider to handle private information with care is essential to attract large communities of users. For example, shortly after StudiVZ (a Web2.0 platform similar to Facebook) had announced a new privacy policy which allowed selling profile information, many users canceled their accounts. When considering current data-protection practices, two aspects determine the trust relationship between platform provider and user: Legal regulations, and privacy enhancement technologies (PET). International data-protection acts of the UN, OECD, EU or FTC [13, 15, 25, 33] and national privacy laws exist. Even more, the EU constitutes data privacy as a fundamental human right. A prominent example for a PET is the Platform for Privacy Preferences (P3P) [23]. It allows a software agent to compare the user preferences to the privacy policy of a web site. In the case of conflicts, the agent notifies the user and blocks the privacy-threatening activity. However, observations [7, 14] show that these technical and legal mechanisms are insufficient to ensure privacy for many Internet users.

The objective of this paper is to find out if there are fundamental problems when applying current legal and technical data-protection mechanisms on Web1.0 and Web2.0 platforms: Instead of narrowing the focus of our research to a specific scenario, we explore general issues that have an impact on mechanisms for data protection. We concentrate on two key aspects: *awareness* and *transparency*. Many technical and legal privacy mechanisms have been devised under the assumption of an aware user. Only a user who is aware of privacy threats can operate appropriate PETs or assert his rights to access, update and delete his personal data as specified in EU directive 95/46/EC [13]. However, it is unclear if the awareness assumption applies in reality. In addition, laws impose transparency on the collectors of personal information, i.e., to make the flow of personal data explicit by declaring privacy policies and informing the individuals concerned. But it has to be investigated if the level of transparency which can be observed in reality is sufficient to support the decisions of the aware user.

In this paper, we motivate and validate hypotheses which address awareness and transparency for online interactions between the users and the platform providers. Therefore our work requires analyses from two perspectives. Regarding the perspective of the users, our hypotheses include that users forget about disclosed data, accidentally reveal personal information and are generally insensitive to the impact of disclosing data. The hypotheses regarding the service providers state that the information collection as well as the flow of personal data is often non-transparent to the users, even though many legal norms for privacy explicitly address this issue. Since we consider both the perspectives of the users and the providers, we require different methods to validate our hypotheses, as we will motivate and explain in the paper. In particular, we conduct a user study and an email survey, and we manually investigate privacy policies and Internet forums. The target group of our study consists of German students of computer science. Our findings show that even IT-skilled users are unable to manage their privacy on the Internet. Thus, we have provided strong evidence for fundamental problems that hold for other parts of the society as well. The EU harmonizes national data-protection law, and PETs are equally available throughout the EU. So our findings apply for other European countries and those companies that joined the Safe Harbor [34] or a similar agreement.

Our work offers deep insight in current privacy problems of the Internet. We show that many legal and technical privacy mechanisms suffer from two fundamental problems: First, they assume an aware and responsible user, which we disprove in this paper. Second, they depend on a level of transparency that can be rarely observed in reality. For example, our survey results reflect that users forget about half of their registrations and even the service providers investigated, which have a high market share, often use unspecific statements in their privacy policy. Based on our findings, we finally derive requirements for future privacy-enhancement technologies.

Paper outline: Section 0 is an overview of legal and technical privacy mechanisms and existing user studies. Section 3 describes observations of ours of privacy-related issues which have motivated our work. Based on these observations, Section 4 introduces and validates our hypotheses. Section 5 concludes.

2. Background

In this section we will provide background information on privacy enhancement technologies and national and international privacy norms. Furthermore, we will discuss user studies about the privacy awareness of the individuals, which are related to our work.

2.1 Privacy Enhancement Technologies

The most prominent example for privacy-enhancement technologies is P3P [23], a protocol standardized by the World Wide Web Consortium (W3C). With P3P, a provider can inform web users about his data-collection and data-use practices electronically. He publishes an XML document containing his privacy policy. On the client side, a software agent like the AT&T Privacy Bird [1] compares the privacy policy to the privacy preferences of the user [22]. If the privacy policy does not match the preferences, the software agent informs the user and blocks the requested operation. Although P3P strives for transparency, it faces a number of problems. For example the expressiveness of P3P is limited. It is not possible to map all information required by EU law to P3P [14]. Further, P3P requires a fundamental understanding of Internet technologies from the users.

Another prominent PET proposed by [3] is implemented by services like spamgourmet.com. The idea is to use unique email addresses like provider.myname@spamgourmet.com for Internet communications in order to identify service providers who trade contact addresses. However, we have carried out preliminary tests with more than 100 providers. These tests have shown that many service providers reject such email addresses in the registration process.

The TOR Project [31] provides a network of “onion routers” which anonymize the users’ IP address in Internet communications. If an Internet user participates in the TOR network, the service provider sees only the IP address of a randomly chosen onion router. Correctly applied, TOR can prevent that a provider learns the geographical position

of the user from the IP address or assigns web page visits to a certain individual. However, the usage of TOR is difficult for users without technical knowledge. TOR cannot prevent that browser plugins reveal private information. Active scripting, cookies, and iframes pose further privacy threats that TOR users have to consider. Furthermore, TOR does not encrypt the communication between the last router and the web server.

All technical mechanisms have in common that they require aware and responsible users. The users have to keep track of recent technological innovations and need a deep understanding of technology, i.e., they have to learn how Internet technologies like cookies or web site requests work and how they threaten the privacy of the individual. So "enhancing trust through security and privacy 'visibility' as well as PET simplicity may be the road to take for PET engineers" [29]. At the end of this paper we will propose four requirements for Internet PETs that would make privacy visible and keep PET complexity at a minimum.

2.2 Privacy Laws and Regulations

The OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (1980) [25] introduce a set of principles and recommendation to harmonize upcoming national legal regulation in the OECD countries. This has become necessary by the increasing amount of personal information that passes borders. The principle addresses the openness, collection limitation, data quality, purpose specification, use limitation, security safeguards, and the individual participation principle. Members of the OECD should respect these principles as a "minimum standard" and include them into national law. Therefore, the OECD principles are part of many current data-protection acts like the FTC report to the Congress (1998) [15], Canada's Personal Information Protection and Electronic Documents Act (2001) [26] or the Asia-Pacific Economic Cooperation (APEC) Privacy Framework (2004) [4]. While the OECD provides only recommendations, directives of the European Parliament must be transposed to national law by all EU member countries. The EU directive 95/46/EC [13] harmonizes data-protection legislation throughout Europe, and implements the OECD principles. We focus on Articles 6 and 10 of the directive that require transparent data flows (openness principle) and an "adequate, relevant and not excessive" data collection in relation to the purpose (collection limitation and data quality principle), and on Article 12 that specifies the rights of the individuals concerned (individual participation principle). Beyond that, the EU directive classifies privacy as a human right [8], i.e., the legislator protects the privacy of the individual. This provides a better protection of the privacy than the US approach. The US regulations define private data as a variant of property that can be traded, concealed or revealed by its owner [24].

Legal measures face a number of general problems. It is unclear if the specified level of transparency is sufficient for the individuals concerned to assert their rights. Transposing the directives and principles often results in an unmanageable number of regulations. For example, the German law contains approximately 1500 norms for data protection. There is a discrepancy of fast technological life cycles and the long period of time a legal norm needs to take effect [28]. Furthermore, it is difficult for privacy authorities to ensure that providers meet legal regulations in scenarios where the collection and processing of private data cannot be observed by the individuals concerned [20]. Examples of such scenarios on the Internet make use of cookies, web-bugs or iframes.

2.3 Related Work

There exist a large corpus of literature on privacy issues in the Internet. In the following, we will briefly introduce a number of prominent studies that are related to our work.

The perception of privacy on the Internet is biased in several aspects. This starts at the misleading name "privacy policy": Many people believe [32] that the presence of a privacy policy means that the website cannot sell data. This is in line with the finding that many users find it too time-consuming to read privacy policies [12]. Furthermore, many users cannot estimate the impact of disclosed information, and demand absurdly low compensation prices for private data [11]. As any kind of information revealed on the Internet can be "broadly disseminated, multiply stored in countless independent permanent storages, and can be retransmitted with the click on a button" [27], this is problematic. For example, specialized search engines like Spock.com or Yasni.de that create comprehensive user profiles from information disclosed at a broad range of Web1.0 and Web2.0 platforms, e.g., Amazon.com present lists, blogs or social networks. Other prominent examples show that linking anonymized health records with public sources lead to personalized medical histories [16, 30].

The biased perception of privacy issues leads to a contradiction: On the one hand, people explicitly state that privacy is very important [1, 28]. On the other hand, they do not behave according to their preferences when disclosing personal information. Studies have verified the discrepancy between privacy goals and user behavior for various scenarios. For example, the users of Web2.0 platforms tend to publish a large share of personal information [17]

without adequately considering the risks. In e-commerce, even users which esteem themselves as very privacy aware disclose much personal information to web-shop provider [29]. As an example from the offline world, study results [1] show that 87.5 percent of individuals which said to be concerned by privacy issues are willing to sign up for a loyalty card that allows to trace their shopping behavior. The consequence of this contradiction is that existing privacy enhancement technologies are not effectively used (e.g., [28]). Thus, it is of utmost importance to identify *fundamental* challenges that have to be considered when devising technical and legal privacy mechanisms.

3. OBSERVATIONS

In this section we will describe some real-world observations of ours that have motivated our work. Others can easily validate these observations by browsing the Internet, talking to experienced Internet users, or searching through empirical studies. In Section 4.2, we will use these observations to derive our hypotheses regarding the unawareness of the users and the non-transparency of the collection and forwarding of personal data.

3.1 Users

Our first observations consider the user perspective.

1. *Users tend to change their service providers frequently and forget about unused registrations.*

Many Internet users register at various service providers, but do not delete unused registrations, e.g., for the sake of a potential future use. For example, people disclose personal data to find classmates (www.friendsreunited.com), reveal their addresses to web shops for hot offers and register just for testing innovative services. However, attractive services change frequently, and the barriers for registering at new services on the Internet are low.

2. *Many users are confident when they see a privacy policy, but do not bother to read it carefully.*

National and international law forces the service providers to communicate their privacy policy. But these policies consist of many pages and often contain euphemistic formulations. Many users find reading privacy policies time consuming: They tend to ignore them [12] or feel confident as soon as they see such a policy [32] but do not watch out for suspicious paragraphs.

3. *Users often forget that friend lists, chatting protocols etc. can be publicly visible on the Internet for a long time.*

Social studies [6, 21] acknowledge that the upcoming generation uses the Internet for socializing with friends and as a replacement for personal interactions. Therefore, the distinction between private conversation and public disclosure becomes increasingly blurred. This is problematic, because web archives (www.archive.org) and the caches of numerous search engines may store revealed information for a long period of time.

4. *Many users are inconsequent when using pseudonymous forums, i.e., sometimes they reveal personal information.*

As the EU [13] (Article 6) implicitly engages the providers to collect as little personal information as possible, most social network sites, blogs or guestbooks offer pseudonymous access. However, the users often disclose real information about themselves and others, e.g., they reveal their addresses and telephone numbers to meet with friends in forums or include identifying characteristics in profile descriptions.

3.2 Providers

The following observations regard the service providers.

5. *Companies have an economic interest in collecting and storing large shares of personal information over a long period of time.*

The business model of various service providers includes collecting and storing detailed user information, e.g., for advertising or for personalized services. Google's 3.1 billion dollar takeover of DoubleClick serves as an example for the market value of such information. The accumulation of personal data over long time is particularly problematic in the case of takeovers, or when global enterprises share information with subsidiaries.

6. *The linkage of personal information accumulated allows service providers to create comprehensive user profiles.*

In the scientific community, the linkage of personal data is a well-known privacy threat. For example, providers have been able to combine registration data with public sources [16, 29] or join data collected by different subsidiaries. New challenges for privacy arise from services (Spock.com, Yasni.de) which link personal information from social network sites, blogs and other sources available through search engines [5, 18].

4. Fundamental Problems of Current Privacy-Protection Mechanisms

In this section we will formulate and validate hypotheses which in sum show that users are systematically unable to exploit the full potential of current privacy-protection acts. We start by introducing two definitions concerning the structure of disclosed data. Section 4.1 describes different methodologies necessary to validate our hypotheses. Section 4.2 motivates each hypothesis and presents our evaluation results.

Information disclosure on the Internet can be *structured* (Observation 1, 2, 4) and *unstructured* (Observation 3, 5, 6).

Definition 1 Structured information disclosure *means that users explicitly associate the sensitive data (attribute value) with an attribute name.*

Structurally disclosed personal information in registration forms makes the semantics of a disclosed value explicit. As an example, it is sometimes hard to distinguish first and last names in foreign languages. With structured information disclosure, the user provides this information by choosing the appropriate fields of the form. If the provider knows the semantics, he can directly process the data without complex techniques like natural language processing (NLP) or text mining.

Definition 2 Unstructured information disclosure *refers to situations where sensitive data can only be implicitly correlated to attribute names.*

Examples of unstructured information disclosure include blog entries, product reviews or guestbook comments where the user provides plain text in natural language. Without applying natural language processing (NLP) or text-mining techniques which transfer this text to structured data, the provider will not be able to process the information collected. The techniques mentioned often lead to equivocal or imprecise results.

4.1 Methodology

Since we distinguish Internet interactions from the perspective of the *user* and the one of the *provider*, we need different methods to validate our hypotheses. To address the user perspective, we set up a study among our colleagues and students, and we explore the user behavior in public Internet forums. We investigate the provider perspective with an email survey and by investigating privacy policies and registration processes by hand.

4.1.1 User Study

Since we are interested in obtaining fundamental results, we have decided to conduct a study with skilled Internet users. If we confirm that even these users cannot ensure their privacy, we have provided strong evidence that the problems apply to less IT-experienced individuals as well. Thus, we tested undergraduate students of computer science, who are very familiar with nowadays Internet practices and are using the Internet as a first-class resource for their educational and social life. In total we have invited 95 former participants of our lecture at a post-exam event to participate in our study. We did not provide money or other incentives for participation. 18 withdrew because of time or other issues, 5 after having read the first questions. Thus, our survey included 72 individuals (17% female, 83% male).

Before we have carried out this study, we had tested it with 18 staff members of our institute and friends, and we had updated our survey after this pre-experimental phase according to their comments. We have carried out the study as an offline survey in our lab. From the beginning every participant was informed about the privacy topic of the survey. Thus, we expect to observe the best privacy practices of our participants. Completing the questionnaire lasted about 20 minutes. During the survey our participants were separated from each other. In order to avoid mistakes due to unclear questions, we encouraged participants to approach the study supervisor, should problems occur. Survey questions that exceed the scope of this paper can be found on our website [19].

4.1.2 Exploration of Internet Forums

Unstructured information disclosure addresses the behavior of individuals in pseudonymous forums which are publicly accessible on the Internet. Thus, we manually examined the international forums of the New York Times and of the Washington Post. In both forums we looked at the most frequently used forum threads or those that had been referenced by popular articles on the main page in November 2007. The topics have been widely spread, e.g., elections, bloggers, health care, food. We excluded empty threads. For the New York Times we analyzed 440, for the Washington Post 200 postings and 54 user profiles.

4.1.3 Manual Investigation of Providers

Because we are interested in the data privacy practices of the providers, we investigated their registration processes and privacy policies by hand. In order to capture the providers which are most important for the participants of our survey, we examined the top-25 providers named by them. (There has been a respective question in the questionnaire.) We excluded that our participants systematically forget less visible but frequently used providers. We therefore included not mentioned providers of the top 10 German provider list reported by the international market researcher ComScore for October 2007 [9]. If a company is part of a group, ComScore only reports results for the group instead of a single company. In such cases we selected a representative from the group.

4.1.4 Email Survey

Legal regulations force the service providers to inform individuals about the handling of personal data. Many providers interpret this vaguely and use unspecific formulation in their privacy policy. However, this renders the flow of personal data non-transparent to the user, and violates the openness principle. Thus, we contacted the companies with unspecific privacy policies (Section 4.1.3) via email. The English translation of the email can be found in the hypothesis. The structure of the email message has been the same for all providers. For our survey, we waited two months for an answer of our email. We have furnished the provider with an email address of ours for further inquiries. When referencing some paragraphs of the privacy policy of the provider, we copied the respective text fragment into the email body. Thus, we sent individual emails that did not reveal the nature of our test. We investigated all responses manually.

4.2 Hypotheses and Evaluation

Our observations in Section 3 have provided anecdotic evidence for privacy challenges on the Internet. In this section we will methodically investigate these challenges by deriving and verifying five hypotheses. We keep the differentiation from the previous sections and distinguish between the perspective of the *users* and the one of the *providers*.

4.2.1 User Behavior

Hypothesis *Awareness*-1 *Users do not keep track of structured information disclosure.*

Motivation Observation 1 (Users tend to change their service providers frequently and forget about registrations at unused providers) and Observation 5 (Provider collect and store as much information as possible for an unknown period of time) have motivated our first hypothesis. Specifically, we want to know if the users are aware of information disclosures due to past registrations. If our hypothesis applies, this means that individuals are not willing to spend effort to maintain their privacy.

Evaluation We can validate our hypothesis by showing that participants remember a significantly smaller set of providers than the one where they are indeed registered. Therefore, we included the following question in our questionnaire:

Question 1: At which service providers have you disclosed personal information in the past? Explicitly list ALL providers with name and/or URL.

The participants of our study had to answer Question 1 two times. First, they were asked to just write down all service providers where they are registered. At the end of the questionnaire we asked the same question again but gave support by a list of provider categories, each one together with concrete examples, like (social network site, MySpace; Instant Messenger, MSN/Skype/ICQ; email, gmx/1&1/web.de).

Obviously, we cannot expect complete lists of registrations. Our provider examples are necessarily incomplete, i.e., there is an unknown number of providers which participants will not remember in the second round of questions. However, this even strengthens our hypothesis.

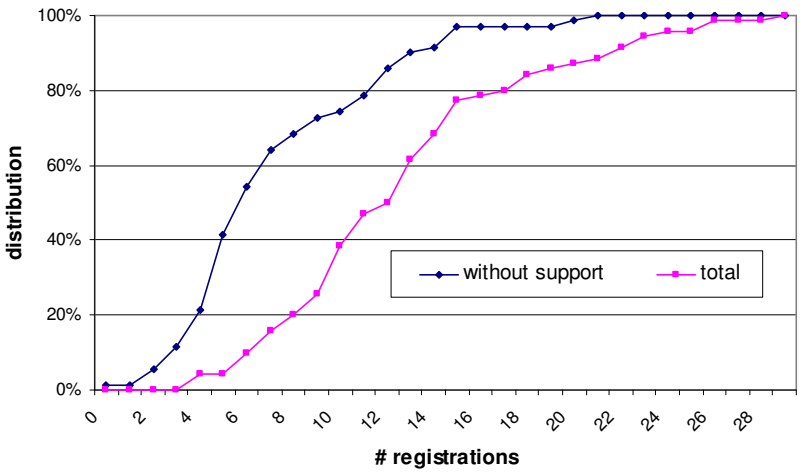

Figure 1 Distribution function of provider registrations

The survey shows that users are registered at a wide range of providers. We have obtained 259 different providers from 70 participants (two did not answer this question). Asking 18 staff members of our institute (pre-experimental phase) resulted in 375 providers.

The difference between the number of registrations participants could remember with and without supporting examples is relevant for our hypothesis. Figure 1 shows the resulting distribution of registrations. The horizontal axis is the average number of registrations; the vertical axis represents the fraction of participants. For example, the figure shows that 60% of our participants remember less than 8 registrations without support.

The difference between both curves indicates that the number of registrations users are aware of is significantly smaller than the real number of registrations. We observed an average of 12 registrations per user with supporting provider examples, and an average of 6 registrations without support. The maximum number of total registrations is 29, the one for registrations remembered without support 21. The minimum of total registrations is 4, respectively 0 for registrations remembered without support. Without support, 45 participants (64% of our study group) remembered less than half of their registrations. Summing up, the user study confirms that many Internet users are unaware of the disclosure of large quantities of personal data. At the end of our questionnaire, we asked all participants if they were aware of the fact that they forgot that many registrations, i.e., we included the following question:

Question 2: Were you surprised by the fact that you forgot about many registrations?

As result, 51 (73%) participants stated that they had expected such a huge difference. 17 (24%) participants said that they were surprised; two did not answer the question. The gap between registrations realized and real ones as well as the disinterest in this privacy aspect validates our hypothesis.

Hypothesis *Awareness*-2 *Users accidentally disclose unstructured personal information.*

Motivation As people use the Internet for socializing with friends and for other forms of private communication, fragments of private communication are visible at Internet forums or social network sites (Observation 3). Furthermore, Observation 4 implies that Internet users sometimes reveal their identity, even if they have the choice of using pseudonyms. This hypothesis reflects that many users are unaware of the fact that this information can be used to create comprehensive personality profiles.

Evaluation Our hypothesis holds if we find pseudonymous forums with a large community of users, where users accidentally disclose personal information that are not explicitly requested from the service provider. Therefore we investigated the supra-national newspaper forums of the New York Times (NYT) and the Washington Post (WP).

The NYT forum requires a registration, but allows its users to choose a new pseudonym for each forum submission. If consistently applied, other forum users can neither identify the person nor find all submissions from a single individual. The WP also requires a registration. But unlike the NYT, the WP links all submissions of a certain user to a publicly visible user profile. Each user of the WP forum can see the profile of all other users, including gender, age, home location, favorite citation and job description.

Investigating 440 forum entries at the NYT lead to 206 submissions (47%) which are signed with the first name of the author. 120 (27%) entries are quoted with the full name, some (16) even added their country, state or web site to their signature like "John Public, FL" (4%). Investigating the Washington Post leads to different results. Only three of 54 users we investigated had entered any further information in their profile. Of those three with a profile, two are forum administrators.

Although we cannot exclude that some registrations contain fake identities, we see the high numbers from the NYT as a validation of our hypothesis. But as the WP shows, the discipline of the users when using pseudonyms depends on aspects that require further investigation. For example, the social group, the user interface or the domain of interest might influence the willingness of the individuals to accidently or unnecessarily publish personal information.

Hypothesis *Awareness*-3 *User are not sensitive to identifying attributes or attribute combinations.*

Motivation According to Observations 5 and 6, many providers have an economic interest in collecting and linking personal information. Joining unique attributes or attribute combinations can identify the individual concerned. For example, it is known that the linkage of zip, gender and date of birth gives way to identification of 63% of the US population [16]. Other examples of identifying attributes are the passport id or an enrollment number. This hypothesis reflects that many Internet users are unaware of the fact that the disclosure of identifying attributes supports the linkability of personal information.

Evaluation We can validate our hypothesis by showing that many Internet users are not aware of the fact that the disclosure of particular attributes or attribute combinations can lead to an identification of the individual. Therefore, we include the following question into our questionnaire:

Question 3 Please name all attributes and attribute combinations whose disclosure in a registration form would let the provider identify you.

As the target group of our study consists of students of computer science, our participants are familiar with the concept of identifiers in the scope of databases (key) and programming languages (object references). Even though the questionnaire did not contain examples of identifiers in order to emulate a real registration scenario, we had expected the participants to identify some obvious attribute combinations such as first/last name, telephone number etc.

Table 1: Number of identifying attribute (combination)s

# attribute (combination)s	number of participants	
0	17	24%
1	22	31%
2	17	24%
3	8	11%
4	3	4%
5	4	6%
6	1	1%

Table 1 and 2 contain the results of this part of our study. Table 1 shows the number of different attributes and attribute combinations which the participants classify as identifiers. Table 2 names the most frequently mentioned attributes or attribute combinations. The participants had to name as many identifying attribute combinations as possible. But surprisingly, 17 persons (24%) have not been able to mention at least one identifier. 22 (31%) participants mentioned exactly one identifier, 17 (24%) mentioned two identifiers.

Table 2: Frequently used attribute (combination)s

Attribute Combination	# Participants
last name, first name	16
last name, first name, birthday	14
last name, first name, address	10
bank id, account number	8
email address	5
phone number	4
credit card number	4
Address	4
identity card	4
nickname, birthday	4
last name address	4
IP	3
Matriculation number	3

Table 2 shows which attributes combinations participants have classified as identifiers. The list is limited to combinations mentioned at least three times. 16 participants stated the combination of first and last name to be sufficient for identification, 14 the combination of first name, last name and date of birth.

We found surprising and inconsistent that only 4 participants considered forum pseudonyms and nicknames. In addition, only 4 persons listed the phone number or credit card number as identifiers. We admit that we cannot exclude that some participants have misunderstood the question. However, recall that we encouraged participants to approach the study supervisor, should problems occur. Even if our participants have actually misunderstood the question, this would also serve as evidence that many individuals have never thought about the impact of identifying attributes or attribute combinations. Thus, we regard the results as a validation of our hypothesis.

4.2.2 Provider

Hypothesis *Transparency*-1 *In many cases users do not know and are not in the position to find out, which information is required when registering.*

Motivation Observation 5 reflects the economic interest of providers in collecting personal information. To balance this interest, legal regulations call for transparency and caution, i.e., collect less personal information (cf. Section 0). However, there is some tolerance for interpreting these regulations. The hypothesis refers to the problem that many implementations of registration processes make it impossible for the user to foresee at registration time if a provider follows the principle to collect as less information as necessary.

Evaluation We can validate our hypothesis by identifying a significant fraction of providers which do not inform the users in advance regarding the data they have to disclose when registering or using the service. To accomplish this, we manually investigate the providers as described in Section 4.1.3 which our survey participants mention frequently and are listed by comScore.

We distinguish three types of registration processes: The basic variant is (1) a registration at one page (SPR) where the provider asks for all information at once. (2) Multiple page registration (MPR) refers to web sites where the user must register to access the site, and is asked for further information when using certain services, e.g., when downloading files or buying goods. A hybrid version (3) is wizard-driven registration (WDR) where all information is requested at the same time, but in a sequence of web pages that appear one after another. While SPR lets the individuals see all required information at once, MPR and WDR hide data to be collected until the user sees subsequent pages. Considering the transparency principle, providers who implement registration processes should give hints regarding the personal information collected at latter stages of the registration process.

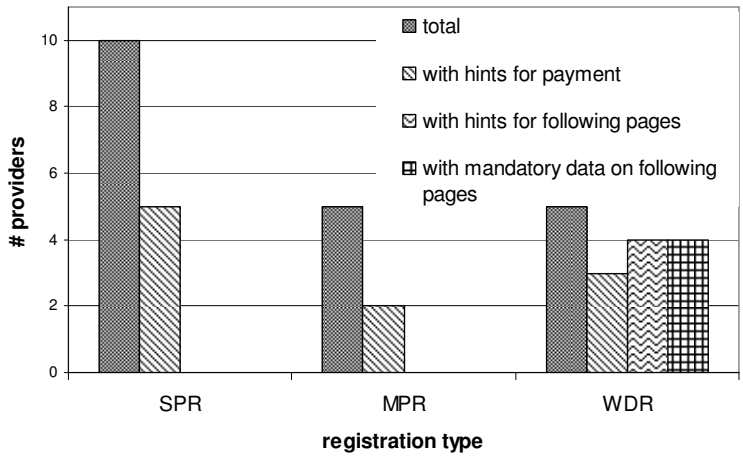

Figure 2: Registration process of the providers

Figure 2 graphs the registration processes we have examined. 10 of 20 providers collect the information necessary for registration at one page (SPR). 5 providers use registrations at multiple pages (MPR), and the same number of providers relies on WDR. The figure further shows that only two MPR providers specify in advance which payment methods are available. This is important, as methods like cash on delivery require less personal information than payments by credit card. Only three WDR providers give hints at the first registration page that further personal information will be requested later. Thus, a large share of providers does not explicitly inform the customers about the quality and quantity of personal information required to access the service. As the user is forced to disclose personal data without knowing which other attributes will be requested, this violates the openness principle. Because providers investigated are frequently mentioned by our participants or have a large market share, our findings confirm that users cannot anticipate information required at registrations.

Hypothesis *Transparency*-2 *Users cannot figure out the information flow between company groups.*

Motivation Large enterprises which provide many different services (e.g., Google) or have many subsidiaries (e.g., the Holtzbrinck group) often announce in their privacy policies to forward personal data to dependent companies only. According to Observation 6, this affects privacy: Individuals might disclose different information at various sites of the same corporation. Linking let them create comprehensive user profiles. The hypothesis reflects that privacy policies do not specify transparently the flow of personal data within the corporate structure.

Evaluation: To validate the hypothesis we decided to directly contact the providers in question. We exclude providers which state to not forward any personal information. We also exclude providers which claim to ask the individual concerned for each transmission of personal data. Table 3 shows the remaining providers with unspecific forwarding rules in their privacy policies.

For our email survey we focus on three kinds of unspecific statements: We ask the provider for their corporate architecture (A) and the related companies (B) mentioned in their privacy policy as a potential addressee of personal data. Many providers state to forward personal information under special circumstances. We found this formulation vague, so we asked the provider about such circumstances (C). An understanding of A, B and C is necessary to estimate the impact of the disclosure of personal data.

Table 3: Email survey results

Provider	(A/B/C)	Addr.	Response (d)
Amazon	A/B	no	No response
Otto	A/B	yes	2d, B open
eBay	A/B/C	no	13d
GMX	A/B/C	yes	7d
Web.de	A/C	no	No response
MSN	A/B	yes	1d, A/B open
Google	A/B	yes	No response
Skype	A/C	yes	2d
AOL.de	A/B	yes	No response
Yahoo.de	A	yes	16d, A open
Sat1	B	yes	7d

All providers in Table 3 state to forward information. Excluding Sat1, each provider states to forward information within its company group without explicitly outlining the recipients (A). Sat1 also forwards information within its group but names the recipients. Skype and Web.de state not to forward data to any non-company group member. Amazon, Otto, eBay, GMX, MSN, Google and AOL state to forward information to related companies like special service provider to fulfill the requested service (B). eBay, GMX, Web.de and Skype state to forward information under special circumstances (C) without more concrete information. We addressed these providers with a German translation of the following email:

Dear Sir or Madam,

I cannot estimate the impact of three formulations in your privacy policy. First, I do not know which companies are part of your group. Second, I do not know who are related companies. I would appreciate, if you can send me two lists, one of the companies of your group and one of your related companies. Third, you state to forward personal information under "special circumstances" or at "legitimate interest". I kindly ask you to clarify this statement.

Best regards (signature)

After 16 days, we obtained answers from 7 companies.

Web.de sent an auto reply email with a telephone hotline we could contact. When calling they told us that they would not forward our information to anybody. We replied that their privacy policy would state to forward personal information within their company group. This resulted in the request to send them their own privacy policy via email. We did not get any further reaction to this email message.

Forwarding within company group (A): MSN and Yahoo told us not to be allowed to forward a list of their company-group members. As described, Web.de gave us information which is obviously wrong. eBay asked us to read the imprint of each eBay market place to identify the responsible subsidiary. As the number of eBay market places is daunting, this reply did not sufficiently answer our request. Only Otto, GMX and Skype followed our request to list all company group members.

Forwarding to related companies (B): eBay states not to disclose their service partners (related companies) because of business secrets. GMX and Sat1 sent detailed lists of related companies. Otto requested our customer id to find the related companies they have already forwarded our personal information to. So far, they have not replied after we had sent them our customer id.

Forwarding under special circumstances (C): eBay explained that the term "legitimate interest" of the privacy policy includes forwarding data to all marketplace members if needed, e.g., in the case a marketplace member has lost contact to a customer. Skype told us that "special interests" means the "interests of Skype or another company group member", and excluded only non-group members. Web.de has not answered.

During our evaluation we made two further findings that support our hypothesis that users cannot figure out information flow between company groups.

Finding 5.1 *Many providers do no explicitly outline contact information for questions regarding data protection.*

To contact the data-protection commissioner of the companies, we searched for privacy-related email addresses, phone numbers etc. To this end, we browsed the privacy policies and the contact-information pages of all providers. MSN offered a contact form with a privacy category, but forced us to acknowledge their privacy policy before we could send our questions regarding the policy. Three providers, Amazon, eBay and Web.de, did not provide contact information for privacy issues. Some providers only provide contact methods bound to a certain (non-privacy) topic. In these cases, we have tried to find the most appropriate alternative (like contract problems) for our request.

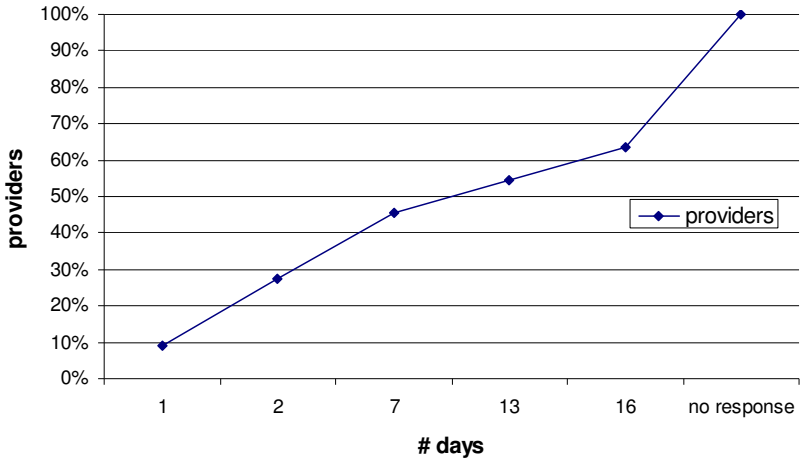

Figure 3: Response time of providers

Finding 5.2 *Many providers react lazily to privacy concerns.*

Figure 3 shows the number of labor days the providers needed to answer our email request. Amazon, Google and AOL did not answer at all. As Figure 3 shows, only 27% of all providers answered within one week.

Summing up, these results show that it is hard for the individual concerned to find out which subsidiaries or dependent companies get to know his personal data, for two reasons: First, privacy policies frequently lack important information. Second, providers make it difficult for the users to contact them regarding privacy issues, react slowly compared to the speed of online interactions, and provide useless responses at times.

4.3 Discussion

Having introduced and validated our hypotheses, we discuss the impact of our findings. Hypotheses Awareness-1-3 reflect that many individuals do not control the disclosure of sensitive information. They do not keep track of disclosed personal information, accidentally disclose data and are unaware of the impact of disclosed attributes. Thus any legal regulation for data protection reaches only a minority of Internet users if it depends on active participation of the individuals concerned.

An important consequence is that the concept of the EU regulations yield better privacy protection than the US approach. While the US legislator defines private data as a variant of property that can be traded and transferred by its owner, the EU directives let the legislator and governmental authorities protect the privacy of the individual. The concept of the EU data protection does not depend on a responsible person that protects his property. However, parts of the EU legislation also contain regulations for the aware user. Article 12 of the EU directive 95/46/EC [13] is an example for such a regulation. It gives individuals the right to access, update or delete personal data. Although this right is of utmost importance for data protection, it cannot be expected from most Internet users to make use of it.

But as Section 2.2 has shown, technical complexity, short innovation cycles and a huge number of laws limit the effectiveness of a regulatory approach. Thus, the legislator should develop new perspectives on data protection. For example, recent proposals for legislation suggest to regulate general issues only, and to create a legal framework for self-regulation in specific areas [10].

Hypotheses Transparency-1 and Transparency-2 tell us that the provider practices regarding the handling of personal data are often non-transparent. Hypothesis Transparency-1 reflects that many providers do not announce which personal information is required for registration. Hypothesis Transparency-2 considers the unclear information flow in a corporate structure due to unspecific privacy policies. This lack of transparency further reduces the efficiency of legal norms that define the rights of the individual: If a user does not know the whereabouts of his personal information, he cannot/will not assert his rights. To our knowledge, existing privacy-enhancement technologies do not propose practical solutions to this problem.

Based on our hypotheses, we can devise requirements for future PETs that hold for structured as well as unstructured disclosed information, traditional Internet applications and Web2.0 applications. These requirements complete and refine the abilities of current privacy mechanisms like P3P or TOR. In particular, a PET

R1: should keep a log at the computer of the user that protocols all disclosed data and let the user access this information at any time.

R2: has to operate in the background and keep the user effort at a minimum.

R3: must warn the user if he discloses personal information that can be linked to other data he has revealed before.

R4: should operate without the support of the data collector.

It will be part of our future research to find out how PETs can be structured in order to meet these requirements.

5. CONCLUSION

In this paper, we have identified fundamental challenges for privacy that impact the trust relationship between user and platform provider in today's Web1.0 and Web2.0 applications. We have introduced and validated various hypotheses on online interactions from the perspective of the users and the one of the providers. They acknowledge that (1) users are unaware of the impact of uncontrolled information disclosure. Particularly, the users do not consider the effects of private information they voluntarily publish on Web2.0 platforms. Furthermore, our hypotheses confirm that (2) even aware users cannot ensure privacy. This holds because current legal and technical measures do not provide the required level of transparency for the flow of personal data at the provider side.

The impact of our findings is high. Many laws and privacy enhancement technologies assume an aware user who asserts its rights. This paper shows that this is not realistic. Further, current regulations do not ensure transparent data flow. Future advances in the areas of data mining or search engines will make this issue more severe. Thus, our results call for the development of new trust increasing perspectives on data protection, both from the legislator and from the developers of privacy-enhancement technologies.

6. REFERENCES

[1] Acquisti, A. & Grossklags, J. Privacy and rationality in individual decision making Security & Privacy, IEEE, 3, 26-33, 2005

[2] AT&T, *Privacy bird*., www.privacybird.org

[3] Aggarwal, G.; Bawa, M.; Ganesan, P.; Garcia-Molina, H.; Kenthapadi, K.; Mishra, N.; Motwani, R.; Srivastava, U.; Thomas, D.; Widom, J. & Xu, Y. *Vision paper: enabling privacy for the paranoids*. Proceedings of the Thirtieth international conference on Very large data bases, VLDB Endowment, p.708-719, 2004.

[4] Asia-Pacific Economic Cooperation (APEC), *Privacy Framework*, 2004

[5] L. Backstrom, C. Dwork, and J. Kleinberg, *Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography*, in WWW '07: Proceedings of the 16th international conference on World Wide Web. New York, NY, USA: ACM Press, pp. 181–190, 2007.

[6] J. Bargh and K. McKenna, *The Internet and Social Life*, Annu. Rev. Psychol., New York University, 2004

[7] P. Beatty,I. Reay, S. Dick and J. Miller *P3P Adoption on E-Commerce Web sites: A Survey and Analysis* IEEE Internet Computing, IEEE Computer Society, 2007, 11, 65-71

[8] M. Caloyannides, P*rivacy vs. information technology.* Security & Privacy Magazine, IEEE, p100-103, 2003.

[9] comScore, *List of Germany's top websites October*, http://de.sys-con.com/read/467066_p.htm, 2007

[10] Council of the European Union, *European Policy Outlook RFID (draft version),* 2007

[11] D. Cvrcek, M. Kumpost, V. Matyas, and G, Danezis. *A study on the value of location privacy* WPES '06: Proceedings of the 5th ACM workshop on Privacy in electronic society, ACM Press, p. 109-118, 2006.

[12] EPTA, *ICT and Privacy in Europe*, Final Report, 2006.

[13] EU, *Directive 95/46/EC of the European Parliament*, 1995.

[14] The European Commission *Working Party on the protection of Individuals with regard to the processing of Personal Data, Platform for Privacy Preferences (P3P) and Open Profiling Standard (OPS),* 1998.
[15] FTC, *Fair information practice principles*., http://www.ftc.gov/reports/privacy3/fairinfo.shtm, 1998
[16] P. Golle, *Revisiting the uniqueness of simple demographics in the us population*. in WPES '06: Proceedings of the 5th ACM workshop on Privacy in electronic society. New York, NY, USA: ACM, pp. 77–80, 2006.
[17] R. Gross, A. Acquisti and H. John Heinz, I. Information *revelation and privacy in online social networks* WPES '05: Proceedings of the 2005 ACM workshop on Privacy in the electronic society, ACM, 71-80, 2005
[18] M. Hay, G. Miklau and S. Srivastava, *Anonymizing social networks*, Technical Report, University of Massachusetts Amherst, 2007.
[19] IPD Project Website, Uni Karlsruhe (TH) http://privacy.ipd.uni-karlsruhe.de/privacy
[20] Klüver, L., et al.: *ICT and Privacy in Europe – A report on different aspects of privacy based on studies made by EPTA members in 7 European countries*. Available at DOI: http://dx.doi.org/10.1553/ITA-pb-a44s, 2006.
[21] R. Kraut, T. Mukhopadhyay, J. Szczypula, S. Kiesler, W. Scherlis. *Communication and information: alternative uses of the Internet in households* CHI '98: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, 1998.
[22] M. Langheinrich (Ed.). *A P3P Preference Exchange Language 1.0, (APPEL1.0)*. W3C Working Draft, 2001.
[23] M. Marchiori (Ed.). *The Platform for Privacy Preferences 1.0 (P3P1.0) Specification*. W3C Proposed Recommendation, 2002.
[24] M. Markel, *Safe harbor and privacy protection: a looming issue for IT professionals Professional,* Communication, IEEE Transactions on, 49, 1-11, 2006
[25] OECD, *Oecd guidelines on the protection of privacy and transborder flows of personal data*, 1980.
[26] Office of the Privacy Commissioner of Canada, Privacy Legislation. *The Personal Information Protection and Electronic Documents Act (PIPEDA)*, 2001.
[27] D. Rosenblum, *What anyone can know: The privacy risks of social networking sites*. IEEE Security and Privacy, vol. 5, no. 3, pp. 40–49, 2007.
[28] S. Spiekermann, J. Grossklags, and B. Berendt, *E-privacy in 2nd generation e-commerce: privacy preferences versus actual behavior*, in EC '01: Proceedings of the 3rd ACM conference on Electronic Commerce. NY, USA: ACM Press, pp. 38–47, 2001.
[29] S. Spiekermann, J. Krumm, G. Abowd, A. Seneviratne, and T. Strang *Privacy Enhancing Technologies for RFID in Retail- An Empirical Investigation*. Ubicomp, Springer, 4717, 56-72, 2007
[30] Sweeney, L. *Uniqueness of simple demographics in the US population.* LIDAP-WP4, 2000.
[31] Tor Project, http://www.torproject.org/,
[32] J. Turow, D. K. Mulligan, and C. J. Hoofnagle, *Research report: Consumer fundamentally misunderstand the online advertising marketplace*. University of Pennsylvania Annenberg School for Communication and UC-Berkeley Samuelson Law Technology and Public Policy Clinic, 2007.
[33] UN, *United Nations guidelines concerning computerized personal data files*, 1990.
[34] US Department of Commerce *Safe Harbor Agreement*, 2000