To appear in: *User Modeling and User-Adapted Interaction: The Journal of Personalization Research*

A PLA-Based Privacy-Enhancing User Modeling Framework and its Evaluation*

Yang Wang · Alfred Kobsa

Received: 11 October 2010 / Accepted in revised form: 18 September 2011

Abstract Reconciling personalization with privacy has been a continuing interest in user modeling research. This aim has computational, legal and behavioral/attitudinal ramifications. We present a dynamic privacy-enhancing user modeling framework that supports compliance with users' personal privacy preferences and with the privacy laws and regulations that apply to them. The framework is based on a software product line architecture. It dynamically selects personalization methods during runtime that meet the current privacy constraints. Since dynamic architectural reconfiguration is typically resource-intensive, we conducted a performance evaluation with four implementations of our system that vary two factors. The results demonstrate that at least one implementation of our approach is technically feasible with comparatively modest additional resources, even for web sites with the highest traffic today. To gauge user reactions to privacy controls that our framework enables, we also conducted a controlled experiment that allowed one group of users to specify privacy preferences and view the resulting effects on employed personalization methods. We found that users in this treatment group utilized this feature, deemed it useful, and had fewer privacy concerns as measured by higher disclosure of their personal data.

Keywords user modeling, privacy laws, privacy preferences, compliance, product line architecture, performance evaluation, user experiment, disclosure behavior

* The managing editor of this paper was Sandra Carberry, University of Delaware.
The conceptual and technical research described herein has already been reported in earlier preliminary publications, as indicated by self-references. The user evaluation and its connection with the other research thrusts is completely unreported.

Yang Wang
Donald Bren School of Information and Computer Sciences, University of California, Irvine, U.S.A.
E-mail: yangwang@uci.edu
Present address: School of Computer Science, Carnegie Mellon University, Pittsburgh, PA

Alfred Kobsa
Donald Bren School of Information and Computer Sciences, University of California, Irvine, U.S.A.
E-mail: kobsa@uci.edu

1 Introduction and Overview

Since personalized websites collect personal data, they are subject to prevailing privacy laws and regulations if the respective individuals are in principle identifiable (see Kobsa (2007b) for a comprehensive review of privacy issues in web personalization). Internationally operating websites are particularly affected since a large number of countries extend the purview of their privacy laws to operators and personal data flows beyond their national boundaries. Such sites may therefore be subject to a multitude of privacy laws, each applying to a subset of users from a certain jurisdiction. In addition to divergent privacy laws and regulations, personalized sites should also cater to users' individual privacy preferences, to encourage them to interact with the site and thus benefit from the full potential of personalization. A user can have varying privacy preferences on different sites, and at different times on the same site, and thus each site should be able to accommodate dynamically changing privacy preferences without delay. In Kobsa (2002) and Wang and Kobsa (2007), we illustrated that these privacy constraints not only affect the data that may be collected by the personalized website, but also the admissibility of personalization methods for processing personal data.

The combinatorial complexity of these privacy constraints make them hard to cope with. We describe a novel approach based on the concept of software product line architecture (PLA) that models the variability in both the privacy and personalization domains. The configuration of the employed personalization methods is then dynamically tailored to each user at runtime, considering both the prevailing privacy norms and the user's current privacy preferences. This flexible approach not only helps address the complexity of building privacy-enhanced personalized systems, but also strongly supports their evolution: as new privacy and personalization concerns arise, they can be added to the product line architecture in a modular manner.

The remainder of the paper is organized as follows. In Section 2, we will review the various privacy constraints in the area of personalization and their impacts on personalized systems. Accommodating these constraints is the aim of our work. Section 3 gives an overview of the existing approaches towards privacy-enhanced personalization. Section 4 presents our PLA-based privacy-enhancing user modeling framework. We start with an illustrative example of its operation, introduce the concept of Product Line Architecture and discuss the aspects of a PLA that we utilize in our framework. We then describe the workings of our framework and show four different implementations of it. Thereafter, we present two evaluations of the framework: Section 5.2 reports a simulation study to assess the performance of the different implementations of our framework, and Section 6 a controlled experiment to examine from the user's perspective the effect of the privacy control that our framework enables. Taken together the two studies show that our framework is a viable solution for addressing users' privacy constraints in personalized systems, and specifically in internationally operating websites that are subject to many different privacy laws and very diverse users. Finally, Section 7 summarizes this work and its contributions.

2 Privacy Constraints in User Modeling

Privacy has been studied for decades, and many different definitions of privacy have been proposed. This disarray is largely due to the fact that privacy is "an overwhelmingly large and nebulous concept" (Boyle, 2003). Young (1978) wittily remarked that "privacy, like an elephant, is more readily recognized than described". In essence, privacy is personal, nuanced, dynamic, situated and contingent (Palen and Dourish, 2002; Dourish and Anderson, 2006), and privacy norms must respect contextual integrity (Nissenbaum, 2010).

If privacy considerations are taken into account in the design of computer systems, they restrain the possible design space for such systems. Solutions that violate privacy constraints cannot be considered any longer. Privacy constraints for computer systems stem primarily from two sources, namely (a) privacy laws and regulations and (b) the personal privacy expectations of computer users. Figure 1 shows the hierarchy of these constraints with a focus on privacy laws and regulations (Wang and Kobsa, 2009b). In the remainder of this section we describe these various privacy constraints and their potential impact on personalized systems.

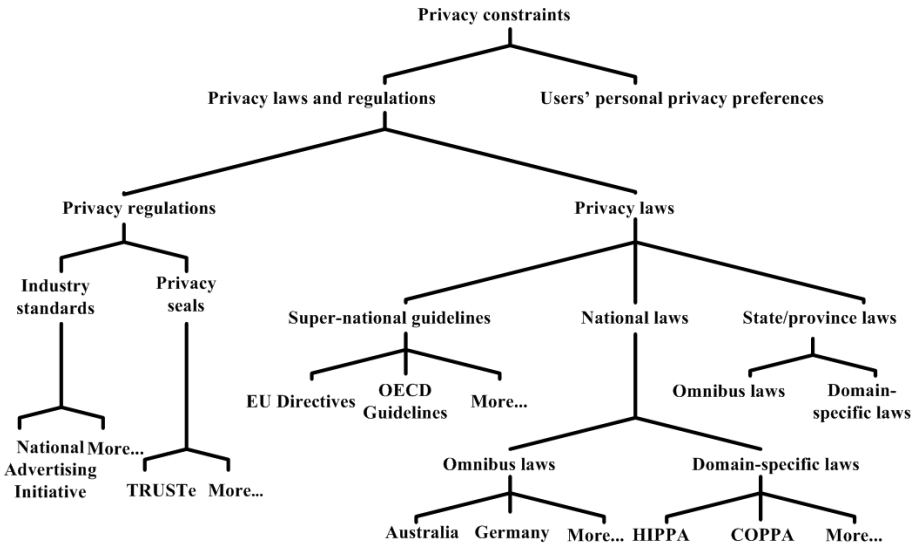

Fig. 1 The hierarchy of potential privacy constraints

2.1 Privacy Laws and Regulations

In the past 40 years, we have witnessed a proliferation of privacy laws and regulations. To date, around 50 countries worldwide and numerous states and provinces have privacy laws enacted. In addition, various types of privacy policies, privacy seal programs, and company or industry self-regulations have been established. These

laws and regulations generally apply only when users are identifiable, i.e. can be uniquely identified with a reasonable amount of effort.

Privacy laws and regulations usually lay out both organizational and technical requirements for ensuring the protection of personal data that is stored and/or processed in information systems. These requirements include, but are not limited to: proper data collection, notification about the purpose of use, permissible data transfer (e.g., to third parties and/or across borders), and permissible data processing (e.g., organization, modification and destruction). Other requirements lay down user opt-in (e.g., asking for users' consent before collecting their data), opt-out (e.g., of data collection and/or data processing), and access rights of data subjects (e.g., regarding what personal information was collected and how it was processed and used). Other provisions mandate adequate security mechanisms (e.g., access control for personal data), and the supervision and auditing of personal data processing.

2.1.1 Impacts of Privacy Laws and Regulations

Our early work (Chen and Kobsa, 2002; Wang and Kobsa, 2007) reviewed over 40 international privacy laws and found that if such laws apply to a personalized website, they often not only affect the personal data that is collected by the website and the way in which this data is shared, but also the *personalization methods* that may be used for processing them. Example provisions from various privacy codes include the following:

1. *Value-added* (e.g. personalized) *services based on traffic*[1] *or location data require the anonymization of such data or the user's consent* (EU, 2002). This clause clearly requires the user's consent for any personalization based on interaction logs if the user can be identified.
2. *The service provider must inform the user of the type of data which will be processed, of the purposes and duration of the processing and whether the data will be transmitted to a third party, prior to obtaining her consent* (EU, 2002). It is sometimes fairly difficult for personalized service providers to specify beforehand the particular personalized services that an individual user may receive. The common practice is to collect as much data as possible about the user, to lay them in stock, and then to apply those personalization methods that "fire" based on these data.
3. *Users must be able to withdraw their consent to the processing of traffic and location data at any time* (EU, 2002). In a strict interpretation, this stipulation requires personalized systems to terminate all traffic or location based personalization immediately after being asked, i.e. even during the current service. A case can probably be made that users should not only be able to make all-or-nothing decisions, but also more nuanced decisions regarding individual aspects of traffic or location based personalization.
4. *Personal data must be collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes* (EU, 1995). This limitation would impact central User Modeling Servers (UMS), which store

[1] The traffic data pertain to communication networks, such as cell phone networks or the Internet.

user information from, and supply the data to, different personalized applications (Kobsa, 2007a). A UMS must not supply data to personalized applications if they will use those data for purposes other than the one for which the data was originally collected.

5. *Usage data must be erased immediately after each session,* except for very limited purposes (DE-TML, 2007). This provision could affect the use of machine learning methods when the learning takes place over several sessions.
6. *The processing of personal data that is intended to appraise the user's personality, including his abilities, performance or conduct, is subject to examination prior to the beginning of processing ("prior checking")* (DE, 2009). *No fully automated individual decisions are allowed that produce legal effects concerning the data subject or significantly affect him and which are based solely on automated processing of data intended to evaluate certain personal aspects relating to him, such as his performance at work, creditworthiness, reliability, conduct, etc* (EU, 1995). These provisions could affect, for example, personalized tutoring applications if they score learners in a manner that significantly affects them.

We found that the privacy laws that impact personalized systems the most are the EU Directive 2002/58/EC concerning the Processing of Personal Data and the Protection of Privacy in the Electronic Communications Sector, and the German Telemedia Law. The reason is that these laws are particularly geared towards electronic communication while other privacy laws and regulations have a much broader scope. More countries are currently enacting such specialized privacy laws to regulate telecommunication, teleservices, e-commerce, and the usage of RFID tags.

2.1.2 Company and Industry Regulations

Many companies have internal policies in place for dealing with personal data. There also exist a number of voluntary privacy standards to which companies can subject themselves (e.g., of the Direct Marketing Association, the Online Privacy Alliance, the U.S. Network Advertising Initiative, the Personalization Consortium, and the TRUSTe privacy seal program). For instance, a website that seeks TRUSTe's certification must provide a detailed self-assessment of its privacy policy and practices (TRUSTe, 2010). The seal program issues a seal certificate only if the websites meets minimum standards that are based upon the U.S. Federal Trade Commission's Fair Information Principles (FTC, 2000). Once the certificate is granted, the seal program may monitor the website's privacy policies and practices, and will handle user complaints that were not resolved by the website providers, possibly leading to an onsite compliance review and seal revocation (there were a few cases in the past in which TRUSTe revoked the seals of some sites).

2.2 Users' Online Privacy Concerns

The second major privacy constraint for personalization are users' individual privacy concerns, specifically with regard to their online interaction. Numerous opinion polls

and empirical studies have revealed that Internet users harbor considerable privacy concerns regarding the disclosure of their personal data to websites, and the monitoring of their Internet activities. These studies were primarily conducted between 1998 and 2003 (and to some extent in 2008 and 2009), mostly in the United States. Below we summarize a number of important findings (see Teltzrow and Kobsa (2004) and Kobsa (2007b) for more details and references). The percentage figures indicate the ratio of respondents who endorsed the respective view, from various surveys.

2.2.1 Personal Data

1. Internet users who are concerned about the privacy or security of their personal information online: 70% - 89.5%;
2. People who refused to give personal information to a web site at one time or another: 82% - 95%;
3. People who refused to provide information to a business or company because they thought it was not really necessary or was too personal: 59% (Pew, 2008);
4. Internet users who would never provide personal information to a web site: 27%;
5. Internet users who supplied false or fictitious information to a web site when asked to register: 6% - 40% always, 7% often, 17% sometimes;
6. People who are concerned if a business uses their data for a purpose different from the one for which their data were originally collected: 89% - 90%.

Significant concern over the use of personal data is visible in these results, which may cause problems for all personalized systems that depend on users disclosing data about themselves. False or fictitious entries when asked to register at a website make all personalization based on such data dubious, and may also jeopardize cross-session identification of users as well as all personalization based thereon. The fact that 80-90% of respondents are concerned if a business shares their information for a different than the original purpose may have impacts on central user modeling servers (Kobsa, 2007a).

2.2.2 User Tracking and Cookies

1. People concerned about being tracked on the Internet: 54% - 63%;
2. People concerned that someone might know their browsing history: 31%;
3. Users who feel uncomfortable being tracked across multiple web sites: 91%;
4. Internet users who generally accept cookies: 62%;
5. Internet users who set their computers to reject cookies: 10% - 25%;
6. Internet users who delete cookies periodically: 53%.

According to a more recent study on tailored advertising (Turow et al, 2009), if given a choice, 68% of respondents "definitely would not" and 19% "probably would not" allow advertisers to track them online even if their online activities would remain anonymous. 63% feel that laws should require advertisers to delete information about their Internet activity immediately, and 69% would like to see a law giving them the right to access all of the information a Web site has collected about them. 86% of young adults reject advertisements that are tailored based on their activities across

multiple Web sites, and 90% of them reject advertisements that are tailored based on information gathered about their offline behavior.

All of these results reveal significant user concerns about tracking and cookies, which may have effects on the acceptance of personalization that is based on usage logs. Observations 4–6 directly affect machine-learning methods that operate on user log data since without cookies or registration, different sessions of the same user can no longer be linked. Observation 3 may again affect the acceptance of the central user modeling systems which collect user information from several websites.

A study on consumer online privacy concern (Lwin et al, 2007) shows that strong business policy is effective in reducing the concerns about collecting data with low sensitivity from users, and users' privacy concerns raise significantly when sensitive data is collected incongruent with the business context. These findings suggest that personalized websites that rely on users' data for provisioning personalization should also have a strong business policy, and should by all means explain why highly sensitive data is collected for their concrete business contexts. Our privacy-enhanced user modeling framework provides a technical foundation for a business policy that allows users control over the processing of their data in personalized systems.

3 Traditional Approaches for Privacy-Enhanced Personalization

In this section, we review a number of existing approaches for privacy enhancement, focused on personalized systems. The first two approaches regard primarily the compliance with privacy laws, while the remaining approaches address individual privacy concerns.

3.1 Largest Permissible Dominator

In the largest permissible denominator approach, websites only use those personalization methods that meet the privacy laws and regulations of *all* their visitors. The Disney website, for instance, meets both the European Union Data Protection Directive (EU, 1995) and the U.S. Children's Online Privacy Protection Act (COPPA) (Disney, 2002). This approach is likely to run into problems if more than a very few jurisdictions are involved, since the largest permissible denominator may then become very small. Individual user privacy concerns are also not taken into account.

3.2 Different Country/Region Versions

In this approach, personalized systems have different country versions, which only use those personalization methods that are admissible in the respective country. If countries have similar privacy laws, combined versions can be built for them using the above-described largest permissible denominator approach. For example, IBM's German-language pages meet the privacy laws of Germany, Austria and Switzerland (IBM, 2003), while IBM's U.S. site meets the legal constraints of the U.S. only. This approach is also likely to become infeasible once the number of countries/regions,

and hence the number of different versions of the personalized system, becomes larger. Individual user privacy concerns are also not taken into account.

3.3 Pseudonymous Personalization

Pseudonymous personalization allows users to remain anonymous with regard to the personalized system and the whole network infrastructure, whilst enabling the system to still recognize the same user in different sessions so that it can cater to her individually. Most of these techniques allow users to adopt more than one pseudonym, to keep apart different aspects of their online activities (e.g., work versus private life).

The Janus Personalized Web Anonymizer (Gabber et al, 1997) serves as a proxy between a user and a web site. For each distinct user-website pair, it utilizes a cryptographic function to automatically generate a different alias (typically a user name, password and email address) for establishing an anonymous account at the website. Janus also supports anonymous email exchanges from a website to a user, and filters potentially identifying information in the HTTP protocol to preserve users' privacy.

Ishitani et al (2003) implemented a system called Masks (Managing Anonymity while Sharing Knowledge to Servers). The system consists of server-side and client-side components. The Masks server, acting as a proxy between users and websites, manages masks (temporary group identifications that are associated with specific topics of interest) and assigns them to users. This enables user information to be collected under those masks, and users to receive group-based personalization. At the client side, privacy and security agents running in conjunction with users' web browsers allow users to configure the masks and provide other privacy functionalities such as blocking and filtering cookies and web bugs.

Kobsa and Schreck (2003) propose a reference architecture for pseudonymous yet fully personalized interaction. The architecture includes a MIX network between applications and user modeling servers, supports standard anonymization techniques between clients and applications, offers a choice of encryption at the application and the transport layers, and a hierarchical role-based access control model. One privacy enhancement of this architecture over other anonymization or pseudonymization techniques is that it hides both the identities of the users and the location of the user modeling servers in the network. The latter is important if a user modeling server is located on a user's local network or platform.

Hitchens et al (2005) present an architecture that allows users to easily create authenticated pseudonymous personas (a subset of a user model), and to selectively share them with certain service providers (via user defined preferences). Service providers can use the information contained in the personas to tailor their services to users.

At first sight, pseudonymous personalization appears to be a panacea for all privacy problems. It seems to protect identity and, in most cases, privacy laws do not apply any more when the interaction is anonymous. However, anonymity is currently difficult and/or tedious to preserve when payments, physical goods and non-electronic services are being exchanged. It harbors the risk of misuse, and it hinders vendors from cross-channel marketing (e.g. sending a product catalog to a web cus-

tomer by mail). Besides, users may still have additional privacy preferences such as not wanting to be profiled even when they are cloaked by a pseudonym, to which personalized systems need to adjust.

Even more troublesome are recent successes to re-identify anonymous data when similar data from the same individuals were available in identified form. Studies were carried out, e.g., in the areas of newsgroups postings (Rao and Rohatgi, 2000), database entries (Sweeney, 2000), web trails (Malin et al, 2003), and ratings (Frankowski et al, 2006). Several real-world incidents demonstrate that these results are not merely academic, but that users can indeed be re-identified from "anonymized" data in practice. In August 2006, AOL publicly released "anonymized" log files containing twenty million search queries from over 650,000 users over a 3-month period. The data included a unique identifier for each user, but otherwise nothing that would traditionally be considered as personally identifiable information. Nevertheless, Internet sleuths were easily able to identify individuals based on these "anonymous" records (Nakashima, 2006). Likewise, after the movie rental company Netflix published 100 million time-stamped ratings from 500,000 users in a contest aimed at improving their recommendation algorithm, researchers were able to associate with near certainty two individuals from a tiny sample of 50 users of the relatively public IMDB database with their "anonymous" Netflix ratings (Narayanan and Shmatikov, 2008). Empirical findings like these give rise to doubts about the value of anonymization (Ohm, 2010) and of the distinction between identifiable and non-identifiable information (Narayanan and Shmatikov, 2010).

3.4 Client-Side Personalization

A number of authors (Mulligan and Schwartz, 2000; Cassel and Wolz, 2001; Ceri et al, 2004; Coroama and Langheinrich, 2006; Fredrikson and Livshits, 2011) have worked on personalized systems in which users' data are located at the client rather than the server side. Likewise, all personalization processes that rely on this data are also carried out exclusively at the client side. From a privacy perspective, this approach has two major advantages:

1. Privacy becomes less of an issue since very few, if any, personal data of users will be stored on the server. In fact, if a website with client-side personalization does not have control over any data that would allow for the identification of users with reasonable means, the site would generally not be subject to privacy laws.
2. Users may possibly be more inclined to disclose their personal data if personalization is performed locally upon locally stored data rather than remotely on remotely stored data, since they may feel more in control of their local physical environment (no empirical verification for this assumption seems to exist as yet though). In times of global network connectivity, such a feeling of local control may be illusionary though. For instance, probably not many Skype users are aware that if they are not sitting behind a firewall or broadband gateway, and have good connectivity to the network, then they are likely to have other people's traffic flowing through their computers (and using their network bandwidth). The

pervasiveness of malware on people's computers also does not speak for higher safety of locally stored personal data.

Client-side personalization also poses a number of challenges though:

1. Popular user modeling and personalization methods that rely on an analysis of data from the whole user population, such as collaborative filtering and stereotype learning (see Kobsa et al, 2001), cannot be applied any more or will have to be radically redesigned.
2. Personalization processes will also have to operate at the client side since even a temporary or partial transmission of personal data to the server is likely to annul the above advantages of client-side personalization. However, program code that is used for personalization often incorporates confidential business rules or methods, and must be protected from reverse engineering. Trusted computing platforms are therefore needed to run such code, similar to the one that Coroama and Langheinrich (Coroama, 2006; Coroama and Langheinrich, 2006) envisage to ensure the integrity of their client-side collection of personal data.

3.5 Distribution, Aggregation, Perturbation and Obfuscation

A number of techniques have been proposed and partially also technically evaluated that can help protect the privacy of users of recommender systems that employ collaborative filtering (Schafer et al, 2007). Traditional collaborative filtering systems collect large amounts of information about their users in a central repository (e.g., users' product ratings, purchased products or visited web pages), to find regularities that allow for future recommendations. Such central repositories may not always be trustworthy though, and they are also likely to constitute an attractive target for unauthorized access. To some extent, central repositories may also be mined for individual user data by requesting recommendations using cleverly constructed profiles (Canny, 2002b). For instance, personal websites tend to be visited by their owners more frequently than by anyone else. In a recommender system that tracks users' website visits, websites that are highly correlated with personal websites are hence likely to have been visited by those owners as well. Requesting a recommendation for pages to visit using a profile that only contains this home page may therefore reveal frequently visited web pages of its owner. Another statistical vulnerability is that correlations between an item and others will disclose much information about the choices of its raters if this item has very few raters only.

Client-side personalization (see Section 3.4) alone is not a remedy against such privacy attacks in collaborative filtering systems. Even when all user profiles are stored at the clients' sides, a considerable number of them (or even all) must still be merged and compiled in order that recommendations can be generated. Below we describe several strategies that are currently investigated to thwart such risks.

3.5.1 Distribution

One possible strategy to better safeguard personal data is to abandon central repositories that contain the data of all users, in favor of distributed clusters that contain

information about some users only. Distribution may also improve the performance and availability of the personalized system. For instance, in the distributed matchmaking system Yenta (Foner, 1997), agents representing a user continuously form clusters of like-minded agents by exchanging information about their users and referring agents to potentially similar other agents. Yenta helps protect user privacy by virtue of the fact that at any given time, agents only maintain the data of a limited number of like-minded agents and that a pseudonymity scheme can by added to protect users' identity.

The distributed PocketLens collaborative filtering algorithm (Miller et al, 2004) goes even further in terms of data avoidance. For each user, PocketLens first searches for neighbors in a P2P network and then incrementally updates the user's individual item-item similarity model by incorporating their ratings one by one (ratings are immediately discarded thereafter). The recommendations produced by PocketLens were shown to be as good as those of the best "centralized" collaborative filtering algorithms.

3.5.2 Aggregation of Encrypted Data

Canny (Canny, 2002b,a) proposed the a secure multi-party computation scheme that allows users to privately maintain their own individual ratings, and a community of such users to compute an aggregate of their private data without disclosing them. The aggregate (a homomorphic encryption of a single-value decomposition of a user-item matrix) then allows personalized recommendations to be generated at the client side using one's own ratings. The scheme is however still prone to the above-mentioned statistical vulnerabilities. The PocketLens system (Miller et al, 2004) was also connected to a blackboard based on the same security schemes as those used by Canny, to allow a community of users to compute a similarity model without having to reveal their individual rankings.

3.5.3 Perturbation

In the perturbation approach, all collaborative filtering is performed by a central server. User ratings become systematically altered before submission though, to hide their true values from the server. Polat and Du (2003, 2005) show that adding random numbers to user ratings may still yield acceptable recommendations. The quality of recommendations based on perturbed data improves with increased number of items and users and decreased standard deviation of the perturbation function (the latter obviously reduces privacy). The authors conducted a series of experiments with two collections of user rankings, namely Jester (Gupta et al, 1999) and MovieLens (Movielens, 1997), using a privacy measure proposed by Agrawal and Aggarwal (2001) that is based on differential entropy between the unperturbed and the perturbed data. For the Jester database, the authors found that privacy levels of about 97% and 90% will introduce average errors of about 13% and 5%, respectively, compared with predictions based on unperturbed data. For MovieLens, the average relative errors due to perturbation at these privacy levels were 10% and 5%, respectively.

3.5.4 Obfuscation

In the obfuscation approach of Berkovsky et al (2005), a certain percentage of users' ratings becomes replaced by different values before the ratings are submitted to a central server for collaborative filtering. Users can freely choose which of their data should be obfuscated, and thus "plausibly deny" the accuracy of any of their data should they become compromised. In subsequent work, Berkovsky et al (2006) combined obfuscation with distributed recommendation generation by ad-hoc peers, which adds an additional layer of privacy protection through distribution (see Section 3.5.1).

The authors performed experiments on the user ratings of Jester (Gupta et al, 1999), MovieLens (Movielens, 1997) and EachMovie (McJones, 1997). They varied the ratio of obfuscated data in users' submitted rankings and compared the ensuing loss of prediction accuracy. They found that obfuscation of the true rating through replacement by the following values had the smallest impact on the prediction error (in the range of 5-7% at an obfuscation rate of 90%): the means of the ratings scale, a random value from the scale, and a random value from the scale taking the means and variance of the ratings in the data set into account. In contrast, uniform replacement by the highest or lowest scale value resulted in an about 300% increased prediction error at a 90% obfuscation rate.

In all these experiments, the data to be obfuscated were randomly selected for each individual user. This strategy disregards though that users are likely to prefer obfuscation for specific kinds of data rather than random data, namely specifically for extreme ratings. Follow-up experiments (Berkovsky et al, 2007) showed that obfuscating extreme ratings (both extremely positive and negative) unfortunately has a much stronger impact on the prediction error than obfuscating moderate ratings.

4 Our Privacy-Enhancing User Modeling Framework

Privacy constraints vary considerably across users and across jurisdictions (see Section 2). Catering to these diverse constraints becomes an enormous combinatorial problem, particularly in internationally operating websites. Moreover, users' privacy expectations are also strongly situation-dependent (Palen and Dourish, 2002; Dourish and Anderson, 2006). At a website, they may change their preferences between sessions or even within the same session. In this section, we present a privacy-enhancing user modeling framework that supports dynamic privacy adaptation in personalized websites. We start with an example of privacy adaptations that our framework is meant to support.

4.1 An Illustrative Example

4.1.1 MyHotel: Personalized Hotel Recommendation

Assume that MyHotel is a mobile application that provides hotel recommendations worldwide based on customers' current location and destination, their hotel prefer-

ences and demographics, as well as the presence of hotels nearby and the ratings of those hotels by other customers. Upon registration, users will be asked to disclose their identities and to optionally reveal some information about themselves (e.g., their hotel preferences). The system will then automatically retrieve their demographics from commercial databases and credit bureaus. The system also incentivizes users to rate businesses they have patronized, by offering discounts on hotels that will be recommended in the future. The processing of all personal data is described in a privacy statement, i.e. the disclosure duties of clause 2 in Section 2.1.1 are being met.

4.1.2 MyHotel's User Modeling Components

To infer information about users that can be utilized in recommendations to them, MyHotel employs a number of inference methods. Each method is encapsulated in a so-called "user modeling component" (UMC) and requires certain data about the user. Table 1 shows the UMC "Pool" of MyHotel, together with the required personal data and inference methods. UMC_1 can recommend hotels based on ratings of people

Table 1 The UMC Pool of MyHotel

UMC	Data Used	Method Used
UMC_1	Demographics and hotel ratings	Clustering
UMC_2	Demographics and hotel preferences	Rule-based reasoning
UMC_3	Hotel ratings and hotel preferences	Item-item collaborative filtering
UMC_4	Hotel preferences and current session log	Supervised machine learning
UMC_5	Hotel preferences and last *n* session logs	Supervised machine learning
UMC_6	Demographics, location data, and last *n* session logs	Supervised machine learning

in the same age range (e.g., 18-22). If a user indicates a preference for a specific type of hotels (e.g., vacation apartments), UMC_2 can recommend nearby hotels that have good ratings in this category. UMC_3 generates hotel recommendations by first running an item-to-item collaborative filtering algorithm on hotel ratings, and then further filtering the recommended hotel list based on the user's hotel preferences. UMC_4, UMC_5, and UMC_6 all use supervised machine learning (e.g., decision tree or support vector machine) to provide hotel recommendations (e.g., they rank hotels into categories with different presumed interest to the user). They differ in terms of the required personal data. UMC_4 uses the user's hotel preferences and current session log (i.e., the MyHotel pages that the user visited in the current log-in session). UMC_5 uses the user's hotel preferences and the *n* most recent session logs (i.e., the MyHotel pages that the user visited in the last *n* sessions). UMC_6 uses the user's demographic data, location data, and logs of the *n* most recent sessions.

4.1.3 Hypothetical Users and Their Privacy Constraints

We have three hypothetical adult users: Alice from Germany, Bob from the U.S., and Chen from China. Bob dislikes being tracked online (see the related results from privacy surveys discussed in Section 2.2.2), while Alice and Chen do not express any

privacy preferences. MyHotel can tailor its personalized hotel recommendations to the different privacy constraints of these users in the following manner:

1. When users log into the website, the system gathers their current privacy constraints, namely those imposed by applicable privacy laws and regulations as well as their personal privacy preferences (users can specify their privacy preferences and change them anytime during the interaction with the personalized system).
2. Our framework determines which UMCs may operate for each user given their privacy constraints.

Table 2 The hypothetical users and their privacy constraints

User	Country	Privacy Constraints
Alice	Germany	EU Directive on Electronic Communications (EU, 2002): - Personal data collected for one purpose may not be used for others - Location data require the anonymization of such data or the user's consent German Telemedia Law (DE-TML, 2007): - Usage data must be erased immediately after each session
Bob	U.S.A.	Personal privacy preference: - He does not want to be tracked online
Chen	China	None

Table 2 lists the privacy constraints of the three users, which bear the following implications:

- For Alice, the German Telemedia Law (DE-TML, 2007) and the EU Directive on Electronic Communications (EU, 2002) apply, with the following consequences:
  - In light of clause 4 in Section 2.1.1, UMC_1, UMC_2 and UMC_6 are illegal without Alice's consent because the demographic data that the website retrieves from commercial databases and credit bureaus had not been originally collected for personalization or recommendation purposes.
  - In light of clause 5, UMC_5 and UMC_6 are illegal because both use cross-session log data.
  - In light of clause 1, UMC_6 is illegal without Alice's consent because it uses location data without anonymizing it.

  Hence UMC_1, UMC_2, UMC_5 and UMC_6 cannot be used for Alice without her explicit consent.
- For Bob, the system can determine that UMC_4, UMC_5 and UMC_6 cannot be used because of his do-not-track preference.
- For Chen, no applicable privacy restrictions were found and thus all six UMCs can be used.

MyHotel will thus instantiate three different personal UMC pools for these three users, i.e. each user will have their own instance of the personalized system that meets their current privacy constraints.

4.2 Product Line Architectures

In order to enable personalized web-based systems to respect users' individual privacy constraints as illustrated above, Kobsa (2003) proposed a user modeling framework that encapsulates different personalization methods in individual components and, at any point during runtime, ensures that only those components that comply with current privacy constraints can be used. We adopted a Product Line Architecture (PLA) approach to implement this approach. PLAs have been successfully used in industrial software development (Bosch, 2000). A PLA represents the architectural structure of a set of related products by defining core elements that are present in all product architectures, and variations in which individual product architectures differ. Each variation point is guarded with a Boolean expression that represents the conditions under which an optional component should be included in a particular product instance (van der Hoek et al, 2001). A product instance can be selected from a product line architecture by resolving the Boolean guards of each variation point at design time, invocation time or execution time (van der Hoek, 2004).

4.3 Overview of the Framework

Figure 2 shows an overview of our framework. It consists of an LDAP-based user modeling server (UMS) (Kobsa and Fink, 2006), to which a user modeling component (UMC) manager, a Scheduler and a cache database were added (the additions are shaded in grey). External user-adaptive applications such as MyHotel can retrieve user information from the UMS so as to personalize services to their end users, and can submit additional user information to the UMS. The UMS includes a Directory Component and a pool of UMCs. The Directory Component hosts a repository of user models, storing users' characteristics and their individual privacy preferences. The UMC Pool contains a set of UMCs, each encapsulating one or more personalization methods (e.g., a specific collaborative filtering or data mining algorithm). UMCs draw inferences about users based on existing information in the user models, and then add the derived user information to the user models (Fink and Kobsa, 2002). In the case of MyHotel, the UMC pool includes the six UMCs presented in Section 4.1.

To enable PLA operations (e.g., the product architecture selections described in Section 4.2), the UMC Manager was added to the UMS. The enhanced UMS was then modeled as a PLA, in which the Directory Component and the UMC Manger are core components, and the UMCs optional components. Each UMC is guarded by a Boolean expression that represents privacy conditions under which the respective UMC may operate. Each privacy condition is expressed by a Boolean variable and relates to users' privacy preferences as well as applicable privacy regulations. For instance, the Boolean guard of MyHotel's UMC_6 reads "`Merging-profile & Tracking & Cross-session-log & (Location-anonymization | Location)`", representing the condition that the following must be legally permitted or approved by the user before UMC_6 may be used: merging profiles (demographic data and other user activity data), tracking users on the site, keeping cross-session usage logs, and using location data (alternatively, anonymized location

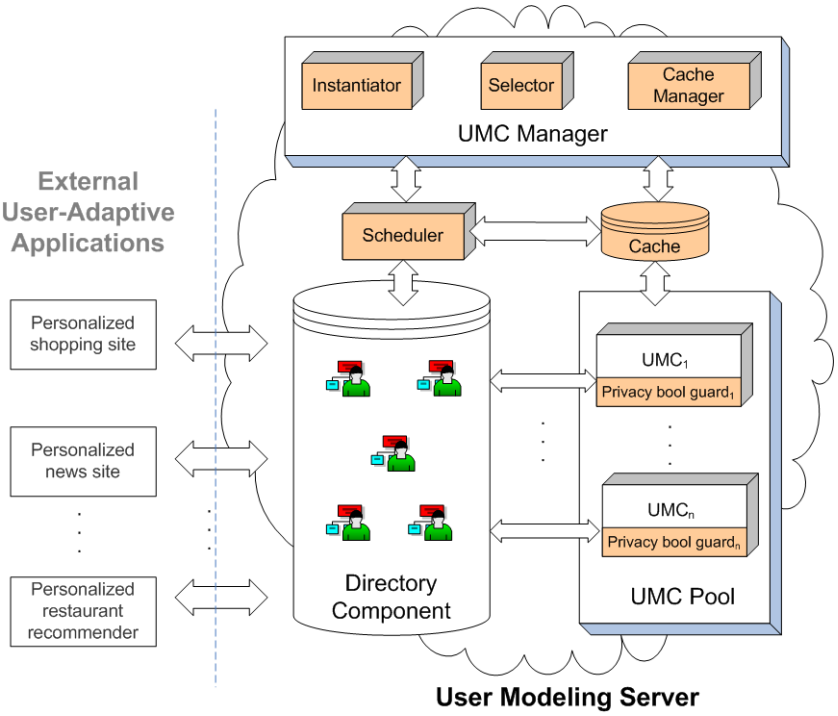

Fig. 2 Distributed dynamic privacy-enhancing user modeling framework

data may be used). The values ("bindings") of these Boolean variables can come from the evaluation of privacy conditions expressed in a privacy policy language.[2], and in the case of individual privacy preferences from the user model or a dialog with the user. For instance, the first binding for Alice is "`Merging-profile = FALSE`".

In the following, we describe in more detail the UMC Manager and its distribution over a network of hosts.

4.4 UMC Manager

The task of the UMC manager is to select the UMCs that may operate under the given privacy constraints of a user, and to instantiate a new architecture for the user that only contains these UMCs or assign the user to such an architecture if it already exists. The UMC Manager consists of the following components:

Selector: When a new user session begins, the Selector takes the PLA and the privacy bindings relating to the new session as inputs. Privacy bindings are name-

[2] See Wang and Kobsa (2009b) for a discussion of these languages, some of which were drafted by industry and by a W3C standardization working group. There also exists a growing industry of tool builders that help enterprises self-examine and self-enforce their compliance with privacy laws. Those tools rely on a formal representation of privacy provisions.

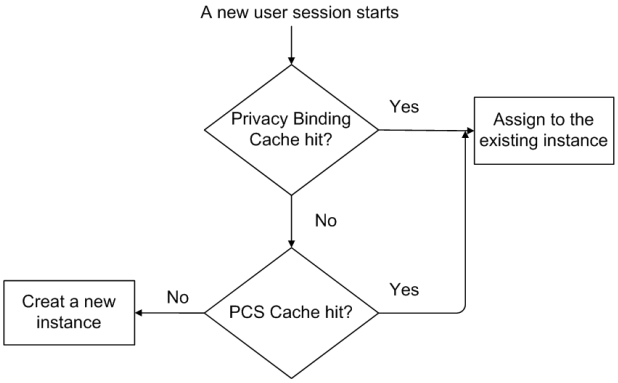

Fig. 3 Two-level caching mechanism

value pairs for the Boolean guards in the PLA. For instance, "Tracking = FALSE" would represent that the user (such as Bob in our example), or some privacy norm relating to the user, disallow user tracking. The Selector selects a particular product architecture from the PLA by resolving the Boolean guards associated with each optional component in the PLA using the current privacy bindings. It expresses the chosen architecture through a binary Privacy Constraint Satisfaction (PCS) vector (Wang et al, 2006) whose n^{th} element represents whether or not the n^{th} UMC may be included in the selected product architecture (1 for included, and 0 for excluded). For instance, since only UMC_3 and UMC_4 do not require Alice's consent, the PCS vector of Alice is (0, 0, 1, 1, 0, 0). Bob's PCS vector is (1, 1, 1, 0, 0, 0) and Chen's (1, 1, 1, 1, 1, 1).

Instantiator: The Instantiator takes a PCS vector as input and creates a runtime system instance for the product architecture. The total number of different PCS vectors ($2^{TotalUMCs}$) equals the theoretical maximum of instances that can be created.

Cache Manager: If two or more users have the same privacy bindings, or the same PCS vectors after selection, then they can share the same user modeling system instance (this is not the case for Alice, Bob and Chen in our example who have very different privacy bindings). This reuse saves the system from performing unnecessary architectural selections and instantiations. We designed a two-level caching strategy for this purpose, which will be described further below.

In order to cope with potentially millions of concurrent users, the privacy-enhanced UMS needs to be distributed. In Fig. 2, the cloud denotes the distribution of processing over a network of computers. The distribution of the LDAP-based Directory Component and the UMC Pool have already been discussed in Kobsa and Fink (2006). For performance reasons, we now also distribute the UMC Manager over a network of hosts, each having a stand-alone copy of the UMC Manager. In addition, we add a Scheduler to the framework to assign incoming user sessions to various hosts, and also a database to store the privacy bindings cache and the PCS vector cache.

5 Implementation and Performance Evaluation

In this section, we describe our implementations of the major components and operations of the proposed framework, the performance evaluations that we conducted with them, and the results from these evaluations.

5.1 Implementation of the Privacy-Enhancing User Modeling Framework

5.1.1 PLA Representation, Selection and Instantiation

We developed two different implementations of the above-described PLA representation, selection and instantiation:

ArchStudio-based Implementation: We adapted functionalities from ArchStudio 4[3] (Dashofy et al, 2007) and implemented it in the Myx architectural style (ArchStudio, 2008). We will call this the Myx version of our framework. ArchStudio 4 utilizes the XML-based architectural description language xADL 2.0 to describe a software architecture.

Our Customized Implementation: The standardization and extensibility of the XML-based PLA representation in ArchStudio 4 come at a price: XML processing can be expensive and can thus negatively affect the overall system performance. This is especially true when the number of components in a PLA is large. Therefore, we designed a customized lightweight alternative to the xADL 2.0 representation. It contains an array of component objects, and each optional component object stores its privacy Boolean guard as an array of privacy Boolean variables. Our customized implementation thus represents the PLA semantics in a succinct object notation and omits any XML processing.

5.1.2 Two-level Caching

As described earlier, if two users have the same privacy bindings, or the same PCS vectors during architecture selection, then they will share the same instance of the user modeling system to avoid unnecessary architectural instantiations. We designed a two-level caching strategy (Opler, 1965; Morenoff and McLean, 1967) for this purpose, which is shown in Fig. 3.

The Cache Manager controls caches of the current users' privacy bindings and of the PCS vectors of the currently instantiated user modeling architectures. When a new user session starts, the Cache Manager first searches the privacy binding cache for an existing user with the same privacy bindings (i.e., a user with identical privacy norms and individual privacy preferences). If one is found, the new session will be assigned to the system instance of this existing user. If no such binding can be found, the Cache Manager will further check the PCS vector cache since a PCS vector may meet the constraints of more than one privacy binding. Only if no such PCS vector can be found either, the Instantiator will start a new instance for this user session.

[3] See Wang et al (2006) for an earlier implementation using ArchStudio 3 (ArchStudio, 2005).

Assume for the example in Section 4.1 that Alice and Bob are the only current users, and that Chen to whom no privacy constraints apply starts a new session. The Cache Manager will then neither find an existing user whose privacy bindings match those of Chen, nor an instantiated architecture for Chen's PCB vector of (1, 1, 1, 1, 1, 1). More details about our dynamic runtime mechanism can be found in Wang et al (2006).

5.1.3 Resource-Aware Scheduling

Since hosts can have different hardware and networking characteristics in our distributed framework (e.g. different amounts of memory), the scheduler needs to take this heterogeneity into account, so as to optimize the overall system performance. When a host becomes available, it will connect and register itself with the Scheduler. The scheduler keeps track of all the registered hosts, their computing capabilities (right now we only consider the memory size), and the number of user sessions that each host is currently serving. When a new user session is initiated, the Scheduler first checks with the Cache Manager to see if any system instance can be reused for this session. If not, it selects the lightest-loaded host that can still handle this session with its resources.

5.2 Performance Evaluation

Performance is a major concern with our approach since architectural reconfiguration during runtime is usually resource-intensive. Will it be practically possible to deploy such a dynamic system at the scale of a contemporary internationally operating website? To answer this question, we conducted an in-depth performance evaluation of our system (details can be found in Wang and Kobsa (2009a)). Such an effort stands in the tradition of similar prior attempts to gauge the performance of user modeling tools through simulation experiments (e.g., Kobsa and Fink (2003); Carmichael et al (2005); Zadorozhny et al (2008)). It is however also substantially different from prior evaluations due to the fact that the workload is not induced by user requests (such as web page requests) or requests from software processes (such as user-adaptive applications or personalization methods), and that the aspired goal is not a user modeling tool that performs personalization tasks efficiently. Rather, the workload is induced by the initiation of new user sessions, and the goal is the efficient instantiation of user-modeling architectures that meet the privacy constraints of each individual user.

5.2.1 Controlled variables

We suspected that the XML-based Myx implementation described in Section 5.1.1 might perform poorly due to all its XML processing and its lack of caching. Therefore, we aimed at contrasting it with our customized implementation (also see Section 5.1.1), with and without caching (see Section 5.1.2). We thus chose the following 2-factorial design for our performance evaluation: (Myx vs. Customized) $\times$ (Non-caching vs. Caching). Resource-aware scheduling was used in all conditions of our experiment.

5.2.2 Simulation Parameters

Since we anticipated that a very large network of machines (which was unavailable to us) would be needed to handle real-world large-scale applications, we determined in pre-trials that 3000 users per host would be a reasonable maximum and simulated such a single host on a PC. The other parameters of our experiment were chosen based on our analysis of international privacy laws and their impacts on personalized systems (Wang and Kobsa, 2006, 2007), as well as the user modeling literature:[4]

- Total number of UMCs in our framework: 10.
- Total number of different privacy constraints: 100.
- Simulated number of user sessions per host: 3000.
- Average arrival rate of unique visitors per host per second: 0.5.
- Number of variables in the privacy Boolean guards of each UMC: 5.

We randomly chose 5 out of the total 100 privacy constraints for each UMC and randomly generated the privacy bindings (true or false) for each user session.

Previous work such as Bhole and Popescu (2005) and Chlebus and Brazier (2007) has shown empirically that the arrival of new user sessions at a website largely follows a Poisson process[5]. To compare the four conditions of our experiment on a common basis, we pre-generated Poisson-distributed session arrival times with a mean rate of 0.5 users per second, and used them in all experiments.

5.3 Testbed

Figure 4 depicts the overall testbed architecture. The performance evaluation of the LDAP-based Directory Component and the UMC Pool in Kobsa and Fink (2006) had already demonstrated that they scale well and can be deployed to high-workload commercial applications. To be able to measure the performance of the PLA selection and instantiation in isolation, we omitted the Directory Component and created functionless dummy implementations for all UMCs, thereby realistically assuming that those components would run on different hosts when deployed in practice. We added a Test Manager to control the experiments, a Request Generator to generate user sessions, and a MySQL database to store the test setup, logs and results. The whole testbed except for the database was implemented in Java, complied in Java 1.6, and run in the HotSpot Java Virtual Machine on a PC platform with two 3.2 GHz processors, 3 GB of RAM, and a 150 GB hard disk.

[4] The authors are not aware of hybrid personalization systems that include more than a handful of personalization methods, of more than a very few dozen identified individual or legal privacy constraints that affect the operation of personalized systems, and Boolean expressions combining those with a length greater than two or three. The simulation parameters are therefore very much on the cautious side and represent a "worst-case scenario".

[5] Chlebus and Brazier (2007) found two separate regions of time in a day, each lasting several hours and having a different average arrival rate. They therefore suggests that the arrival rate rather follows a non-stationary Poisson process, i.e. consists of more than one Poisson process, each with its own rate. Those results are not likely to apply to internationally operating sites though on which we largely focus.

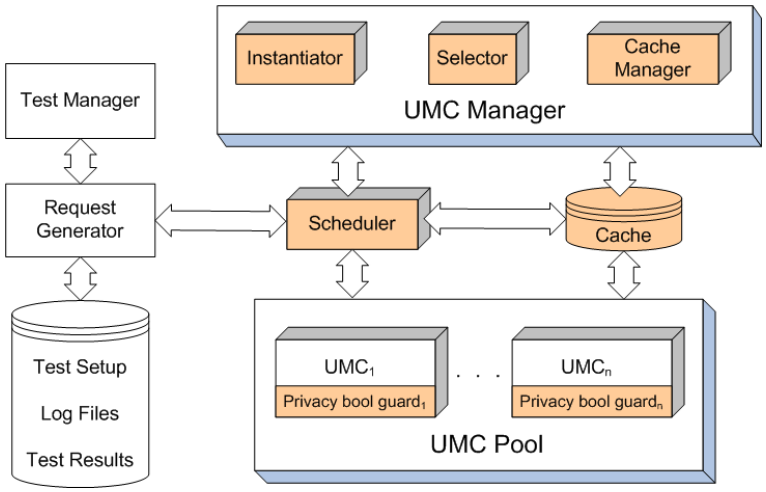

Fig. 4 Testbed architecture

5.3.1 Procedures

The Test Manager first reads the test setup from the database and informs the Request Generator to generate simulated user sessions and associated privacy bindings. The Request Generator reads the session arrival times from the database and starts sending user sessions to the Scheduler. The Scheduler chooses a host to handle the session. The host then performs the PLA selection and instantiation (in the Cache conditions, PLA selection and/or instantiation may be skipped, depending on the type of cache hit – see Section 5.1.2). Once the session has been assigned to a runtime system instance, the assignment is written into the cache if a cache is used. After all user sessions have been handled, log files and test results are written into the database.

For every user session, we measure three values:

Handling time, which is the period between the Request Generator sending the session to the Scheduler, and the session being assigned to a runtime instance.

Reuse rate of runtime instances, which considers the total number of user sessions and of instances currently in the system. It has a range of [0, 1) and is calculated as $\frac{Total\ Sessions - Total\ Instances}{Total\ Sessions}$

Relative performance improvement, which compares the system performance of the original implementation (Myx implementation without caching) with that of an enhanced implementation. For a given number of users handled, this value has a range of [0, 1) and is calculated as
$$\frac{\sum TotalHandlingTimeOriginalVersion - \sum TotalHandlingTimeEnhancedVersion}{\sum TotalHandlingTimeOriginalVersion}$$

5.3.2 Evaluation Results

Handling Time per User Session. Figure 5 plots the handling times for each user session in the four implementations, and indicates the means and standard deviations.

We can see that the customized versions perform better than the Myx versions, that our two-level caching mechanism improves both versions, and that the customized version with caching performs best. The average handling time per user session is less than 0.2 seconds for all versions except the Myx implementation without caching.

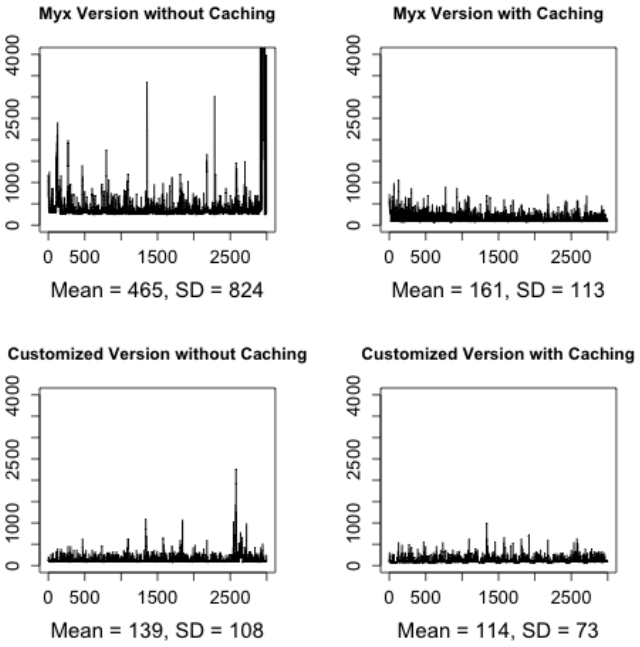

Fig. 5 Handling time for each user session (in milliseconds)

We also analyzed the spikes in the handling times in Figure 5, and disconfirmed that they were correlated with bursts in the arrival rate. Based on an analysis of the logs created by our experimental testbed we found that the main reason for the delays lies in Java's non-deterministic thread scheduling. Requests to handle a new session, select an architecture, and instantiate an architecture each create a new thread, and occasionally one of the threads gets switched out of processing and later switched back in. One can notice that in the Myx version without caching, the frequency of long handling times increases towards the end of the experiment. This is because the test machine almost ran out of heap space, and the Java Virtual Machine kept switching threads. A good remedy for these effects of non-deterministic thread switching is to shorten the processing time, which is confirmed by the substantial decrease of such delays in the conditions where the customized version and/or caching have been used.

Runtime Instance Reuse Rate. Figure 6 plots the runtime instance reuse rates for the two caching versions, which we define as the ratio of current instances with more

than one user to the total number of current instances (in the non-caching versions, no instances are being reused). The reuse rates for the caching versions increase degressively as the cumulative number of user sessions increases. The two curves are very similar because both versions use the same caching scheme; the small variations are due to the true randomness of privacy Boolean guard and privacy binding generation.

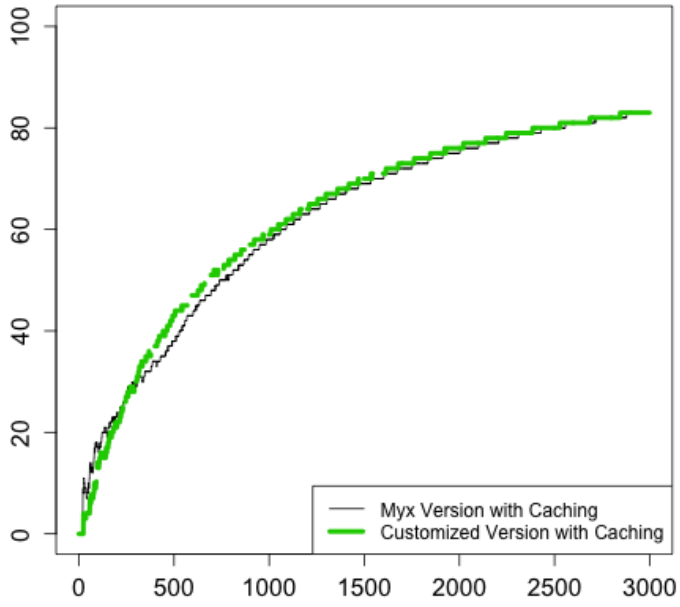

Fig. 6 Instance reuse (in %), by cumulative number of users

Performance Improvement. Figure 7 plots the performance gain of our three improved versions in comparison to the baseline Myx version without caching. The lowermost curve (gain from the Myx version with caching) goes up as expected: the cache size increases with an increased number of users, and hence the hit rate and thus the performance gain increase as well. The curve in the middle (gain from the customized version without caching) is always above the first curve, meaning that the gains through customization are larger than through caching. As expected, this difference becomes smaller with an increasing number of users and thus cache hits. The topmost curve shows the gains from both caching and customization. While the combined effect is always higher than each single effect, they are unfortunately not additive. While the gains through caching increase with an increased number of

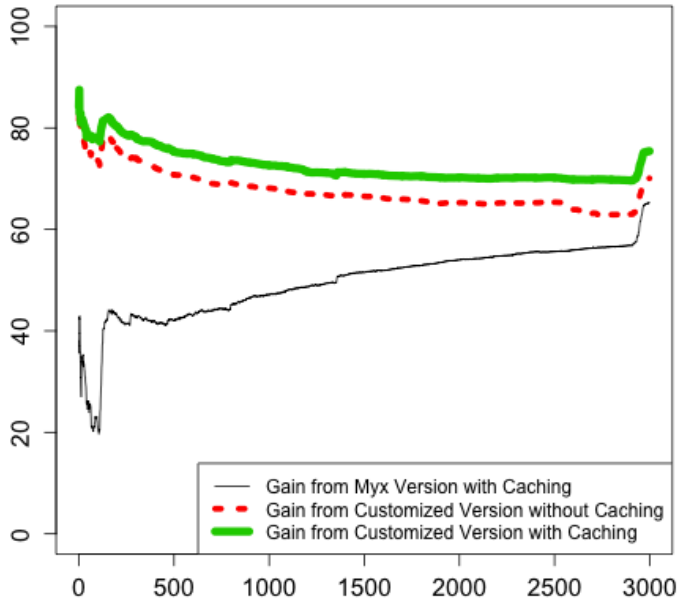

Fig. 7 Performance improvement (in %), by cumulative number of users

users, each hit "cancels out" the gains through customization which will not be invoked in such a case. A larger number of cache entries still leads to performance gains, as is demonstrated by the slightly increasing distance between the middle and upper curves. This differential however grows far less than the slope of the lowermost curve, which represents the gains through caching for the non-customized Myx version.

5.3.3 Discussion of the Performance Results

Performance Improvement. The evaluation demonstrates that caching and particularly customization both improve the performance, and that using neither engenders slow and erratic response behavior. We see two reasons for this result: The customized versions use a light-weight PLA representation, which consumes less memory and enables faster PLA selection and instantiation than the XML-based Myx versions. The two-level caching mechanism saves time and resources that would otherwise be spent on creating new runtime instances. Under the current completely random assignment of privacy guards and bindings, the probability of a privacy binding cache hit is $1/\,2^{TotalConstraints}$ (about $7.9e^{-31}$), while the probability of a PCS cache hit is

$1/\ 2^{TotalUMCs}$ (about $9.8e^{-4}$). Therefore, the vast majority of instance reuses came from the PCS cache hits.

Practical Implications. The average arrival rate of new visitors in the current experiment setup is 0.5 per second. In contrast, Google.com which Alexa and Compete currently (as of April 2011) rank No. 1 worldwide in terms of traffic seems to have a daily reach of about 3.24 billion visits per month[6]. This translates into an average arrival rate of 1250 visits per second. Because of its modular approach, our framework would be able to handle this workload in a cloud-computing paradigm (Buyya et al, 2008). Using our current average processing rate of 0.5 visitors per second per node we can handle Google-sized traffic with a cloud of 2,500 nodes. In comparison, Google uses hundreds of thousands of much better equipped servers (Shankland, 2009). Therefore we believe that with sufficient support from a cloud-computing environment, our approach can scale well to serve internationally operating websites, which would profit most from our privacy-enhancing framework. As a reminder though, this number does not include the nodes that would be required to run the Directory Component, the User Modeling Component, and of course the Web server.

Limitations of the Performance Evaluation. Privacy bindings are randomly assigned to sessions in our simulation, and hence their variations are evenly distributed across users. In reality though, users' individual privacy preferences are likely to gravitate towards typical constellations, countries may have limited combinations of privacy bindings, and visitors from certain countries may be more frequent than from others. In a more realistic scenario with uneven distributions, the hit rate in the privacy binding cache is therefore likely to be higher and hence the number of generated different instances lower than in our simulation, both of which reduces the memory load. Another limitation is that our experiments were conducted on a single PC platform. When the user modeling server is distributed in a cloud computing environment, the Scheduler and the cache database are likely to be overloaded, and therefore will need to be distributed as well. Finally, as mentioned above, our chosen simulation parameters in Section 5.2.2 are very much on the cautious side and represent a "worst-case scenario" to test the practical feasibility of our framework in the context of the largest internationally operating websites. Our simulation does not allow precise prediction of the performance of our framework if some of the parameters have to be changed for a concrete deployment (except that, everything else being equal, lowering a parameter will generally improve the performance). In such a case, another simulation run with revised parameters will be necessary.

6 User Evaluation

Our proposed privacy-enhancing user modeling framework selects personalization methods and creates user-specific personalization architectures, in accordance with the privacy regulations that apply to each user as well each user's individual privacy

[6] See http://siteanalytics.compete.com/google.com/

preferences. The adaptation to privacy regulations makes it easier for website operators to bring their websites in compliance with highly divergent privacy laws and regulations (at very reasonable cost, as we showed in our performance evaluation in Section 5.3.3). The adaptation to individual privacy preferences allows website operators to not only inform users about personalization methods that are being used (as they currently do in so-called website "privacy policies"), but also to give users *control* over which of the available methods may be employed.

The latter can be cautiously expected to have an impact both on the behavior of users who interact with a privacy-enhanced personalized system, and their attitudes toward such a system. Respondents in numerous privacy surveys demand knowledge of and control over the use of personal information in a website (see Kobsa (2007b) for a survey), and earlier behavioral experiments of ours indicate that improved disclosure of privacy practices by a website increases users' disclosure of personal data to the website (Kobsa and Teltzrow, 2005). Telling users how their personal data will be used and giving them control over this usage also has been shown to positively influence trust in a website (Hine and Eve, 1998; Jensen et al, 2005), and trust in return is positively related both to intended (Schoenbachler and Gordon, 2002) and actual (Metzger, 2004) disclosure of personal information. We can hence expect that our privacy enhancements will have similar effects on users.

We therefore developed a Privacy Control Panel (PCP) that allows users to specify their individual privacy constraints, view the personalization methods that can operate under these constraints, and obtain detailed information on the capabilities of these methods. Our privacy-enhancing user modeling framework can then turn these specifications into a personalized architecture for the specific user (and update it whenever the user changes those specifications, even during runtime). In the next few subsections, we will describe the design of this PCP, its evaluation in a controlled experiment, the results of this study, and will also discuss the implications of these results.

6.1 Privacy Control Panel

The PCP allows users to view the available privacy options and their consequences on the permissibility of personalization methods, and to specify their personal privacy preferences. In Fig. 8, this PCP is located on the right-hand side of the screen. The panel carries the title "Privacy Control" and has two parts. The top part contains a list of privacy preferences with regard to the operation of the system that the user can specify. For instance, by checking the third option "Track what you do on our site", a user gives her consent that the system can keep track of her interactions with the site. By default, all privacy options are unchecked. The bottom part contains a list of available (but fictitious) personalization methods. The second method "Rule-based reasoning I" is a minimum personalization method that the system always uses. Other methods will be marked as selected or de-selected depending on the privacy settings the user specified in the top component. For instance, if the user checks the third privacy option "Track what you do on our site", then the fourth personalization method "Incremental learning" will be marked as being turned on. If the user changes her pri-

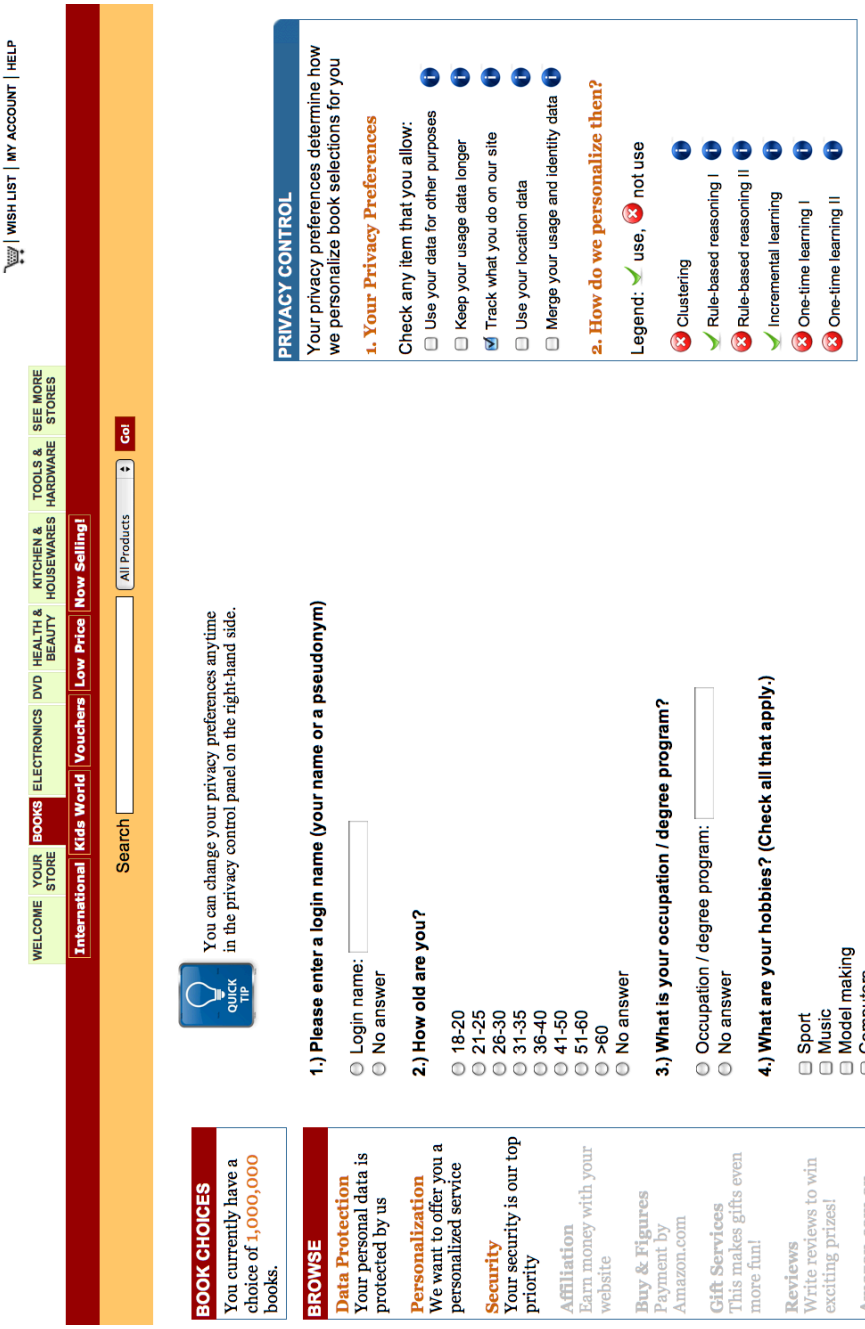

Fig. 8 Experiment website (treatment group)

vacy settings, the personalization methods will be re-evaluated for selection. The idea is that only the selected personalization methods will be used to provide personalized services to this user, which can be realized by our privacy-enhancing user modeling framework. When users click at one of the blue "i" icons next to each of these privacy options and personalization methods, a pop-up window will appear with more details thereon. For instance, clicking at the "i" of the third privacy preference ("Track what you do on our site") will yield the explanation "We will not track what you do on our site, unless you check the checkbox to allow us to do so." Clicking at the fourth personalization method ("Incremental learning") yields the following explanation:

> We use incremental machine learning with your book preferences and current session log on our site as input.
> If the checkbox of privacy setting 3 is not checked, then we will not use this method because users usually do not want to be tracked online.

The PCP is shown persistently on all pages of a web site. A "quick tip" that reminds users that they can change their privacy settings anytime using the PCP is persistently displayed on top of the page.

6.2 Methodological Background

Two approaches can be pursued to study users' reactions to our PCP. In an attitudinal approach, users would be asked about their opinions on the PCP, such as how it would influence their privacy-related behavior (to improve the external validity of the study, users can be provided with representations of the PCP design, from paper-based sketches to a fully functioning PCP). In an observational approach, the privacy-related behavior of users would be observed while carrying out some tasks using the PCP. Both approaches complement each other: while inquiries may reveal aspects of users' rationale that cannot be inferred from mere observation, observations allow one to see actual user behavior which may differ from self-predicted behavior.

This latter discrepancy seems to prevail in the area of privacy. Existing literature such as Spiekermann et al (2001) and Berendt et al (2005) found that users' stated privacy preferences deviate significantly from their actual behavior. Solely relying on interview-based techniques for analyzing privacy impacts on users, as is currently still often the case, must therefore be viewed with caution. Our empirical studies hence gravitated towards observation, but were complemented by questionnaires and brief informal interviews.

On a different note, behavioral studies with users interacting with a computer system often face the problem of lacking or incomplete system implementation (since this would be, e.g., too time-consuming, or even impossible given the current state of knowledge). Researchers then often use various forms of deception to give users the impression that they are working with a real system, which is likely to improve the external validity of the study. Ordered by decreasing degree of deception, typical techniques include (1) Wizard-of-Oz experiments, in which important operations of the system are carried out by a human; (2) pretend-studies (sometimes called "Robot-of-Oz" experiments), in which major "shortcuts" are introduced in the system to replace

missing system parts; and (3) studies in which the user is channeled past missing system functionality through clever interface design and task selection. Our experiment can be classed into the second and the third category.

6.3 Experimental Design

To test the effects of the proposed mechanism, we designed a between-subjects experiment with two conditions, to which subjects were randomly assigned:

- The control group used the standard version of the personalized website (i.e., without privacy enhancement).
- The "enhanced group" used the version with the above-described PCP. The PCP was verbally explained to subjects at the beginning of the study. It was constantly present at the website, and subjects could interact with it while carrying out the experimental tasks at the website shown in Fig. 8.

The experiment was designed to determine whether subjects would exhibit different data sharing and purchase behaviors and express different attitudes toward the system, depending on the condition to which they were assigned. In accordance with past findings mentioned at the beginning of this section, our hypothesis was that users in the enhanced condition would be more willing to share personal data and view sites more favorably than users in the control condition. We treat the condition as an independent variable and users' behaviors and attitudinal reports as dependent variables.

6.4 Material

We developed a fake book recommendation and sales website whose interface was designed to suggest an experimental future version of a popular online bookstore. Two variants of this system, with and without the PCP, were created for the two aforementioned experimental conditions. In both conditions, the standard privacy policy of the web retailer was used. The three left-hand links labeled "Data Protection", "Personalization" and "Security" led to the original company privacy statement (we split it into these three topics though and left out irrelevant text).

A counter was visibly placed in the upper left corner of each page that purported to represent the size of the currently recommended selection of books (see Figure 8). Initially the counter was set to 1 million books. Data entries in web forms (both via checkboxes or radio buttons and through textual input) decreased the counter after a page is submitted, by a random amount. The aim was to give study participants the feeling that the more information they provide, the more targeted will be the set of recommended books. The web forms asked a broad range of questions, most of them relating to books. A few sensitive questions on users' political interests, religious interests and affiliation, their literary sexual preferences, and their interest in certain medical subareas (including venereal diseases) were also present. For each question, users had the option of checking a "no answer" box or simply leaving the question

unanswered. The personal information that is solicited in the web forms was chosen in such a way that it may be relevant for book recommendations and/or general customer and market analysis. A total of 32 questions with 83 answer options were presented. Ten questions allowed multiple answers, and seven questions had several answer fields with open text entries (each of which we counted as one answer option).

After eight pages of data entry (with a decreased book selection count after each page), users were encouraged to review their entries. The website then displayed a list of fifty predetermined and invariant books that were selected based on their low price and their presumable attractiveness for students (book topics include popular fiction, politics, tourism, and sex and health advisories). The prices of all books were visibly marked down 70%, resulting in total out-of-pocket expenses between $3 and $12 for a book purchase. For each book, users can retrieve a page with bibliographic data, editorial reviews, and ratings and reviews by readers.

Users were given the opportunity to buy a single book from this set of 50 "recommended" discounted books. They were free to choose whether or not to make that purchase. Those who did were asked for their names, shipping addresses and payment data (a choice of debit or credit card charge was offered).

6.5 Subjects

Based on initial pilot tests we used the following eligibility criteria for participants: they must have previous online shopping experience and own a credit or debit card that can be used for online purchases. 65 subjects participated in the experiment. They were students from a large public university in the U.S. with a wide range of majors. The data of seven subjects who appeared familiar to the student experimenter was not used, since we suspected that these subjects might have behaved in a more privacy-conscious manner in case they also felt they were known to the experimenter.

6.6 Experimental Procedures

Study participants were recruited through posters on campus and through email announcements in various campus distribution lists. As an incentive for their participation, they were promised a $10 coupon for a nearby popular coffee shop and the option to purchase a book with a 70% discount. Scheduled participants were asked to bring their ID and a credit or debit card to the experiment. When subjects showed up for the experiment, they were reminded to check whether they had these credentials with them, but no data was registered at this time.

Subjects were given a study information sheet and were informed that they would test an experimental new version of an online bookstore with an intelligent book recommendation engine. They were told that the system would ask them a number of book-related questions and then generate 50 personalized book recommendations based on their answers to these questions. Users were also told that the more and the better answers they provided, the better would generally be the quality of the recommendations to them. They were made aware that their data would be given

to the book retailer after the experiment. It was explicitly pointed out though that they were not required to answer any question. Subjects were asked to work with the prototype to find books that suited their interests, and to optionally pick and purchase one of them at a 70% discount. They were instructed that payments could be made by credit or debit card.

Subjects were randomly assigned to either the control condition without the PCP or the enhanced condition with the PCP. After searching for books and possibly buying one, subjects filled in a post-questionnaire. In the debriefing phase, the IDs and payment cards of those users who had bought a book were compared with the address and payment data they had entered into the system.

6.7 Results

We obtained valid data from 58 subjects, half of them in the control condition and half in the enhanced condition.

6.7.1 Data Sharing Behavior

Number of questions answered: We first dichotomized subjects' responses by determining whether a question received at least one answer, or was not answered at all (i.e., no input was provided or the box "no answer" was checked). On average, 87% of questions were answered in the control condition, while this rose to 91% in the enhanced condition (see Table 3). A one-tailed t-test on the total number of questions answered by subjects in each condition showed that the difference between conditions was statistically significant (p=0.04).

Table 3 Data sharing behavior and results of one-tailed t tests

	control group	**enhanced group**	**df**	**t**	**p**	**N**
% Questions answered	87%	91%	45	1.7896	**0.04**	58
% Answers given	59%	63%	56	1.725	**0.045**	58

Number of answers given: The two conditions also differed with respect to the number of answers given (see Table 3). The maximum number of answers that subjects could reasonably give was 64, and we used this as the maximum number of possible answers. In the control condition, subjects gave 59% of all possible responses on average (counting all options for multiple answers), while this rose to 63% in the "enhanced" condition. A one-tailed t-test on the percentage of answers provided by subjects in each condition showed that the difference between conditions was statistically significant (p=0.045).

6.7.2 Purchases

Table 4 shows that the purchase ratio in the enhanced condition is higher than in the control condition, even though all subjects saw the same set of 50 books in both

conditions. A one-tailed t-test for proportions indicates that this result approaches significance ($p<0.09$).[7] The difference between the purchase rates is 0.21, which represents an increase of more than 60%.

Table 4 Purchase ratio and results of one-tailed t-test for proportions

	control group	**enhanced group**	**df**	**Chi-Square**	**p**	**N**
% Purchase ratio	0.34	0.55	1	1.74	*0.09*	58

Note that the decision to buy is a significant step in terms of privacy. In order to purchase a book, users must not only reveal their name, address and payment data but they also risk that data they had entered earlier pseudonymously may now be linked to their identities. The PCP, which allows users to set their privacy preferences and to view the resulting changes on what methods will be used to generate personalized recommendations, seemingly mitigates such privacy concerns.

6.7.3 Rating of Privacy Practices and of Perceived Data Disclosure Benefits

The post-questionnaire that was administered at the end of the study included a number of 5-item Likert questions. Table 5 presents these questions together with the average scores in each condition, and the results of a series of t-tests on the levels of response. The first five questions asked about subjects' privacy concern regarding the book site. The agreement with the statement "the new book website assigns high priority to data protection" was significantly higher in the "enhanced" condition than in the "control" condition ($p<0.02$). The difference for the item "the new book website uses my data in a responsible manner" approached significance ($p<0.12$). The two remaining questions asked about subjects' perceived quality of the recommendations. No significant difference was found between the two conditions.

6.7.4 Self-Reported Practices and Perceived Usefulness of the Privacy Control Panel

The post-questionnaire in the enhanced condition also contained six extra questions about the PCP. 83% of subjects said that they paid attention to the PCP during the study. 66% reported that they had set privacy options in the PCP in order to change their privacy preferences. 41% stated that they did so in order to try out what happens. 38% of subjects indicated that they clicked at information icon(s) in the PCP to obtain more information about the privacy preferences and/or personalization methods. Table 6 summarizes these results.

The remaining two extra questions in the enhanced condition refer to users' perception of the usefulness of the PCP, and their intent to use it in the future. Table 7 summarizes the responses.

[7] "When p fails to beat α by a small amount, researchers often say [. . .] that their findings *approached significance*" (Huck, 2012, p. 154).

Table 5 Users' perception of privacy practices and benefits from data disclosure. 1: strongly disagree, 2: disagree, 3: not sure, 4: agree, 5: strongly agree.

Item	**control** (mean)	**enhanced** (mean)	**df**	**t**	**p**	**N**
I felt that my data are in good hands at the new book website	3.59	3.69	54	0.42	0.34	58
I understood how the new book website used the data that I provided	3.24	3.41	56	0.63	0.26	58
I find the new book website reliable	3.52	3.52	56	0	0.5	58
The new book website assigns high priority to data protection	3.34	3.79	55	2.09	**0.02**	58
The new book website handles my data in a responsible manner	3.59	3.83	56	1.16	*0.12*	58
I was satisfied with the book recommendations that I received from the new book website	2.52	2.48	55	0.12	0.55	58
The data that I provided helped the website select interesting books for me	3.00	2.89	55	0.37	0.36	58

Table 6 Users' self-reported practices with regard to the PCP (in the "enhanced" group).

Item	**% of users**	**N**
I paid attention to the privacy control panel	83%	29
I clicked the "info" icon(s) to learn about privacy preferences or personalization methods	38%	29
I set options in the privacy PCP in order to change my privacy preferences	66%	29
I set options in the privacy PCP in order to try out what happens	41%	29

Table 7 Users' perception of the usefulness of the PCP (in the "enhanced" group). 1: strongly disagree, 2: disagree, 3: not sure, 4: agree, 5: strongly agree.

Item	**Average rank**	**N**
Privacy UI is useful in general	3.97	29
I would use a Privacy UI if a site offers one	4.03	29

6.7.5 Effect of Privacy Concern on Users' Attitudes and Behaviors

To gauge subjects' privacy concerns, we asked the 15 Likert scale questions of the Concern for Information Privacy (CFIP) scale by Smith et al (1996). We took the average of the answers to the 15 questions as a score of each subjects' privacy concern. The average score is 4.01 (SD = 0.65) for the control group, and 4.13 (SD = 0.58) for the treatment group. This indicates that our subjects were generally privacy concerned (1-strongly privacy unconcerned, 3-neutral, 5-strongly privacy concerned). We found no statistically significant difference between the two conditions using the Wilcoxon rank sum test ($p>0.46$).

We ran regression analyses to investigate the effect of privacy concern on users' data disclosure and purchase behavior. The experiment condition, the score of privacy concern, and gender were the independent variables. For the number of questions

answered, the score of privacy concern did not seem to be a significant predictor ($p>0.9$). However, for the number of answer options chosen, the score of privacy concern was approaching significance ($p=0.08$). We also performed a logistic regression analysis for the purchase behavior, which was a binary independent variable. We found that the score of privacy concern (regression coefficient = -1.2984, $p=0.03$) had an even stronger effect than the experiment condition (regression coefficient = 0.4177, $p=0.04$). This result suggests that people with higher level of privacy concern are less likely to make an online purchase.

We also examined the effect of privacy concern on how subjects in the treatment group perceived our PCP. In terms of the perceived usefulness of the PCP, we found that the score of privacy concern was approaching significance (coefficient = 0.5917, $p=0.068$). As for whether a subject would use such a PCP if a site offers one, we found that the score of privacy concern was a significant predictor (coefficient = 0.8159, $p=0.025$). These results seem to indicate that people with higher level of privacy concern value our PCP more.

6.7.6 Comments on the PCP in Informal Interviews

We also conducted brief informal interviews after the experiment, and solicited comments on the PCP from subjects in the enhanced condition as well as suggestions to improve it. In general, users liked the idea of a PCP. We heard many positive comments such as "I really like the privacy control. I wish companies can adopt it" and "Great feature! It's user-friendly. I like the fact that it stays all screens and you can change it anytime."

However, there is plenty of room for improvement. Several users complained that some of the textual descriptions of the privacy options and personalization methods, and also further explanations thereon from the "Info" icons, were difficult to understand. For instance, what does "other purposes" mean in the first privacy option "use your data for other purposes?", or what is the "clustering" personalization method? They suggested providing more concrete explanations of how their data will be used. One subject commented that "I didn't really understand all the practical implications, e.g., practically how the data will be used." Another subject suggested using some kind of metaphor like "calorie" information found on food packaging to explain what and how user data will be used.

Some users explicitly said that they trust Amazon and/or had positive experience with Amazon's recommendations before. They either ignored or paid little attention to the PCP, or quickly played with the privacy options to select the most powerful personalization. Others largely focused on the privacy options, and ignored the personalization methods since they did not quite understand them. One subject in the "enhanced" condition stated: "I liked the idea [of a PCP], but I didn't play with it. It reminds me of privacy concerns, I chose 'no answer' for many questions."

6.8 Discussion of the Experimental Results and Open Research Questions

Our experiment was designed so as to create an online shopping experience as realistic as possible, and thereby to increase its external validity. The incentive of a highly discounted book purchase and the deterrent of an initially extremely large selection set that visibly decreased with more answers given were devised to entice users to provide ample and truthful data about their interests. The claim that all data would be made available to the website operators meant that users faced potential privacy risks when providing answers anonymously, and even more so when deciding to buy a book and thereby to disclose their identities.

The results demonstrate that our proposed PCP that is enabled by our privacy-enhancing user modeling framework has a significant positive effect on users' willingness to divulge data to the website, and on their perception of the website's privacy practices. The additional finding that this mechanism also leads to more purchases (presumably due to less reluctance to disclose name, address and payment data) approached statistical significance. While the experiment does not allow for substantiated conclusions regarding the underlying reasons that link the two conditions with the observed effects, the results are largely in agreement with the literature (see Section 6). The adoption by web retailers of interface designs that contain such a PCP therefore seems clearly advisable. However, the concrete design of the Privacy Control Panel still needs further exploration and verification, and so does the question whether the degree of comprehension of the privacy options and their effects is a mediating factor in their effectiveness.

In an earlier experiment conducted in Germany, our collaborators analyzed what effects a contextualized disclosure of privacy practices at the interface would have on users' data sharing and purchase behavior in personalized websites (Kobsa and Teltzrow, 2005). In order to compare our results with theirs, our experiment largely reused the material from the German experiment including the overall website layout and structure as well as the questions asked.[8] In a nutshell, our results reveal similar trends as in the German experiment. The respective privacy enhancement in both studies showed positive effects on users' data sharing and purchase behavior, and on perceptions of the website's privacy practices. Our experiment has all the statistically significant results yielded in the German experiment except for the perceived usefulness of disclosing data. The subjects in the German experiment were predominately business students, whereas ours came from diverse disciplines such as engineering, mathematics, chemistry, biology, medicine, social sciences and law. Choosing 50 books that may potentially interest a heterogeneous group is seemingly more difficult than for a homogeneous group. Indeed, many of our subjects said they did not find the recommended books interesting. This may explain why we did not get a significant result on the perceived usefulness of data disclosure.

Hine and Eve (1998) found in their study of consumer privacy concerns that "in the absence of straightforward explanations on the purposes of data collection, people were able to produce their own versions of the organization's motivation that

[8] We had to choose a different set of 50 books though since many of the previously used books were in German and thus useless for our new study. However, we tried to keep the range of book prices and types of book topics the same.

were unlikely to be favorable. Clear and readily available explanations might alleviate some of the unfavorable speculation". One may conjecture that the opportunity offered by the PCP in illustrating the relationship between users' data disclosure and the underlying personalization alleviated some of the unfavorable speculation in our experiment. Culnan and Bies (2003) postulated that consumers will "continue to disclose personal information as long as they perceive that they receive benefits that exceed the current or future risks of disclosure. Implied here is an expectation that organizations not only need to offer benefits that consumers find attractive, but they also need to be open and honest about their information practices so that consumers ... can make an informed choice about whether or not to disclose." Again, the PCP makes the personalization process more transparent and better yet, more controllable for the end users. The PCP in our experiment aligns with the "openness" principle laid out in the above quotations, and the predicted effects were indeed observed in our experiment.

Having said this, we would however also like to point out that additional factors may also play a role in users' data disclosure behavior, which were kept constant in our experiment. One example is the reputation of a website. We chose a web store that enjoys a relatively high reputation in the US. It is well known that reputation increases users' willingness to share personal data with a website (see e.g. Earp and Baumer (2003) and Xie et al (2006)). Our high response rates of 87% without and 91% with the PCP suggest that we may have already experienced some ceiling effects (after all, some questions may have been completely irrelevant for the interests of some users so that they had no reason to answer them). This raises the possibility that websites with a lower reputation may experience an even stronger effect of our privacy enhancement mechanism.

In our experiment, the PCP was permanently visible in the enhanced condition. This uses up a considerable amount of screen real estate. Can the same effect be achieved in a less space-consuming manner, for instance, by replacing the PCP with a link or an icon that symbolizes the availability of such a panel? If so, how can the Privacy Control Panel be presented so that users can easily access it without being distracted by it? Should this be done through regular page links, links to pop-up windows, or rollover windows that pop up when users mouse over the link or icon?

From the informal interviews, we learned that many subjects were not really trying to understand or even paying attention to the PCP part that explains which personalization methods are being used. We may therefore not want to present that part permanently by default. Several subjects also asked for more concrete explanations of the privacy options and personalization methods. This remains a UI design challenge because the constrained size of the panel demands a succinct and yet clear representation. One idea is to explore the mentioned "calorie" metaphor that one of our subjects suggested. We can develop visual representations of privacy options and personalization methods illustrating what type of personal data (e.g., usage logs), how much of the data (e.g., from a single session or from one year of usage), and how the data will be used (e.g., one-year usage data combined with demographic data).[9] Some

[9] See the related attempt of designing "privacy nutrition labels" to represent corporate privacy policies (Ciocchetti, 2008; Kelley et al, 2009, 2010).

subjects also requested a clearer representation of the relationship between their data disclosure and the quality of the personalization they receive. Instead of just showing personalization methods being turned on or off based on users' privacy settings, we can potentially show examples of books that would not be recommended if certain privacy options are checked. In other words, the effect of choosing a privacy option is now reflected in terms of recommended books rather than (presumably less comprehensible) personalization methods. All of these design ideas will need to be tested with additional implementations and user studies.

7 Conclusions

Our work aims at reconciling privacy and personalization in personalized websites, in a manner that addresses both users' personal privacy preferences and the applicable privacy laws and regulations. We present a framework based on a software product line architecture that dynamically selects personalization methods during runtime that meet the current privacy constraints, and gives each user a privacy-tailored instance of the personalized system.

Since the PLA selection and instantiation process is quite resource-intensive, we developed four implementations of our approach and evaluated their performance in a simulation experiment. Our study shows that our light-weight customized implementation performs better than the original PLA implementation, and that our two-level caching mechanism improves both versions. Overall, our performance results demonstrate that with a reasonable number of networked hosts in a cloud computing environment, even the largest internationally operating website today can use our dynamic PLA-based privacy-enhancing approach to personalize their user services and at the same time respect the individual privacy desires of their users as well as the applicable privacy norms.

The privacy constraints that stem from privacy laws and regulations can be entered by the company that operates the website (this requires a very thorough one-time analysis, and infrequent updates when a law or regulation is being changed). The individual privacy constraints can come from a user model, but likely need to often be specified by the users themselves since research shows that privacy preferences often vary for the same individual depending on a variety of contextual factors. We therefore developed a Privacy Control Panel that allows users to indicate their privacy constraints, view the personalization methods that can operate under these constraints, and get detailed information on the capabilities of these methods.

In order to evaluate the prospective effects of our privacy-enhancing mechanisms on users' privacy-related attitudes and behaviors we developed a pretend shopping site equipped with this Privacy Control Panel that allows users to specify their privacy preferences and view the resulting changes to the activation status of the provided personalization methods. In a controlled experiment, subjects who had this control panel available exhibited less privacy concern than subjects who did not (as measured by the amount of personal data disclosure and of purchase decisions that bore privacy risks). These subjects also held favorable opinions of the privacy practices of this site and the usefulness of the PCP.

In summary, the performance study demonstrates that our privacy-enhancing mechanism can be scaled to meet the demands of high-traffic sites, and the user experiment shows that the presentation at the user interface of the controls and the information that this mechanism affords reduces users' privacy concerns. Taking the evidence from the two studies together, we believe that our privacy enhancement mechanism is a viable solution for addressing users' privacy constraints in personalized systems, and specifically in internationally operating websites that are subject to many different privacy laws and very diverse users. In future work (Knijnenburg et al, 2011), we will continue to explore possibilities of making personal data collection and processing practices more transparent to end users and enabling them to self-assess the benefits and risks associated with these practices and to control their data in a more informed manner.

Acknowledgements This research has been supported through NSF grants IIS-0308277, IIS-0808783 and CNS-0831526, and through a Google Research Award. We would like to thank Scott Hendrickson, Eric Dashofy and André van der Hoek as well as the four anonymous journal reviewers for their valuable comments.

References

Agrawal D, Aggarwal CC (2001) On the design and quantification of privacy preserving data mining algorithms. In: 20th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database System, Santa Barbara, CA, pp 247–255

ArchStudio (2005) Archstudio 3. `www.isr.uci.edu/projects/archstudio/`

ArchStudio (2008) Myx. `www.isr.uci.edu/projects/archstudio/myx.html`

Berendt B, Günther O, Spiekermann S (2005) Privacy in e-commerce: stated preferences vs. actual behavior. Commun ACM 48(4):101–106, DOI 10.1145/1053291.1053295

Berkovsky S, Eytani Y, Kuflik T, Ricci F (2005) Privacy-enhanced collaborative filtering. In: PEP05, UM05 Workshop on Privacy-Enhanced Personalization, Edinburgh, UK, pp 75–84

Berkovsky S, Eytani Y, Kuflik T, Ricci F (2006) Hierarchical neighborhood topology for privacy-enhanced collaborative filtering. In: PEP06, CHI06 Workshop on Privacy-Enhanced Personalization, Montreal, Canada, pp 6–13

Berkovsky S, Kuflik T, Ricci F (2007) Distributed collaborative filtering with domain specialization. In: Konstan JA, Riedl J, Smyth B (eds) RecSys: Proceedings of the 2007 ACM conference on Recommender systems, ACM, Minneapolis, MN, pp 33–40

Bhole Y, Popescu A (2005) Measurement and analysis of http traffic. Journal of Network and Systems Management 13(4):357–371, DOI 10.1007/s10922-005-9000-y

Bosch J (2000) Design and Use of Software Architectures: Adopting and Evolving a Product-Line Approach. Addison-Wesley Professional, Reading, MA

Boyle M (2003) A shared vocabulary for privacy. In: UbiComp 2003 Workshop on Ubicomp Communities: Privacy as Boundary Negotiation, Seattle, WA

Buyya R, Yeo CS, Venugopal S (2008) Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities. In: 10th IEEE Intl. Conf. on High Perf. Comp. and Comms., IEEE Computer Society, pp 5–13, DOI http://dx.doi.org/10.1109/HPCC.2008.172

Canny J (2002a) Collaborative filtering with privacy. In: Proceedings of the 2002 IEEE Symposium on Security and Privacy, IEEE Computer Society, pp 45–57, DOI 10.1109/SECPRI.2002.1004361

Canny J (2002b) Collaborative filtering with privacy via factor analysis. In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, Tampere, Finland, pp 238–245, DOI 10.1145/564376.564419

Carmichael DJ, Kay J, Kummerfeld B (2005) Consistent modeling of users, devices and sensors in a ubiquitous computing environment. User Modeling and User-Adapted Interaction 15(3-4):197–234

Cassel L, Wolz U (2001) Client side personalization. In: DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries, http://www.ercim.eu/publication/ws-proceedings/DelNoe02/CasselWolz.pdf

Ceri S, Dolog P, Matera M, Nejdl W (2004) Model-driven design of web applications with client-side adaptation. In: Proceedings of the International Conference on Web Engineering, pp 201–214, DOI 10.1007/978-3-540-27834-4_26

Chen Z, Kobsa A (2002) A collection and systematization of international privacy laws, with special consideration of internationally operating personalized websites. http://www.ics.uci.edu/ kobsa/privacy

Chlebus E, Brazier J (2007) Nonstationary poisson modeling of web browsing session arrivals. Information Processing Letters 102(5):187–190, DOI http://dx.doi.org/10.1016/j.ipl.2006.12.015

Ciocchetti C (2008) The future of privacy policies: A privacy nutrition label filled with fair information practices. John Marshall J of Comp & Info Law 26(1), URL `http://www.jcil.org/journal/articles/495.html`

Coroama V (2006) The smart tachograph - individual accounting of traffic costs and its implications. In: Fishkin KP, Schiele B, Nixon P, Quigley AJ (eds) Pervasive Computing: 4th International Conference, PERVASIVE 2006, Springer, Lecture Notes in Computer Science, vol 3968, pp 135–152

Coroama V, Langheinrich M (2006) Personalized vehicle insurance rates: A case for client-side personalization in ubiquitous computing. In: Kobsa A, Chellappa R, Spiekermann S (eds) Proceedings of PEP06, CHI 2006 Workshop on Privacy-Enhanced Personalization, Montreal, Canada, pp 56–59, http://www.isr.uci.edu/pep06/papers/PEP06_CoroamaLangheinrich.pdf

Culnan MJ, Bies RJ (2003) Consumer privacy: Balancing economic and justice considerations. Journal of Social Issues 59(2):323–342, DOI 10.1111/1540-4560.00067

Dashofy E, Asuncion H, Hendrickson S, Suryanarayana G, Georgas J, Taylor R (2007) Archstudio 4: An architecture-based meta-modeling environment. In: ICSE 2007: International Conference on Software Engineering, IEEE Computer Society, pp 67–68

DE (2009) German Federal Data Protection Act, as of 14 Aug. 2009. http://www.bfdi.bund.de/EN/DataProtectionActs/Artikel/BDSG_idFv01092009.pdf

DE-TML (2007) German telemedia law. http://www.gesetze-im-internet.de/tmg/

Disney (2002) Personal communication, chief privacy officer, Disney Corporation

Dourish P, Anderson K (2006) Collective information practice: Exploring privacy and security as social and cultural phenomena. Human-Computer Interaction 21(3):319–342

Earp JB, Baumer D (2003) Innovative web use to learn about consumer behavior and online privacy. Commun ACM 46(4):81–83, DOI 10.1145/641205.641209

EU (1995) Directive 95/46/EC of the European parliament and of the Council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal of the European Communities (23 November 1995 No L. 281):31ff, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML

EU (2002) Directive 2002/58/EC of the European parliament and of the Council concerning the processing of personal data and the protection of privacy in the electronic communications sector. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002L0058:EN:HTML

Fink J, Kobsa A (2002) User modeling in personalized city tours. Artificial Intelligence Review 18(1):33–74, DOI 10.1023/A:1016383418977

Foner L (1997) Yenta: a multi-agent, referral-based matchmaking system. In: AGENTS '97: Proceedings of the first international conference on Autonomous agents, ACM Press, pp 307, 301, URL http://dx.doi.org/10.1145/267658.267732

Frankowski D, Cosley D, Sen S, Terveen L, Riedl J (2006) You are what you say: Privacy risks of public mentions. In: 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA, pp 565–572, DOI 10.1145/1148170.1148267

Fredrikson M, Livshits B (2011) RePriv: re-imagining content personalization and in-browser privacy. In: 2011 IEEE Symposium on Security and Privacy (SP), IEEE, pp 131–146, DOI 10.1109/SP.2011.37

FTC (2000) Privacy Online: Fair Information Practices in the Electronic Marketplace. A Report to Congress. Federal Trade Commission, www.ftc.gov/reports/privacy2000/privacy2000.pdf

Gabber E, Gibbons PB, Matias Y, Mayer A (1997) How to make personalized web browsing simple, secure, and anonymous. In: Financial Cryptography'97, Lecture Notes in Computer Science, vol 1318, Springer Verlag, Berlin - Heidelberg - New York, pp 17–31, DOI 10.1007/3-540-63594-7_64

Gupta D, Digiovanni M, Narita H, Goldberg K (1999) Jester 2.0 (poster abstract): evaluation of an new linear time collaborative filtering algorithm. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM, Berkeley, California, United States, pp 291–292, DOI 10.1145/312624.312718

Hine C, Eve J (1998) Privacy in the marketplace. The Information Society 14:253–262, DOI 10.1080/019722498128700

Hitchens M, Kay J, Kummerfeld B, Brar A (2005) Secure identity management for pseudo-anonymous service access. In: Hutter D, Ullmann M (eds) Security in Pervasive Computing: Second International Conference, SPC 2005, Boppard, Germany, April 6-8, 2005. Proceedings, Springer Verlag, Berlin - Heidelberg, pp 48–55, DOI 10.1007/b135497

van der Hoek A (2004) Design-time product line architectures for any-time variability. Sci Comp Prog, special issue on Softw Variability Mgmt 53(30):285–304

van der Hoek A, Mikic-Rakic M, Roshandel R, Medvidovic N (2001) Taming architectural evolution. In: 9th ACM Symp. on the Foundations of Softw. Eng., pp 1–10

Huck SW (2012) Reading Statistics and Research, 6th edn. Pearson Education, Boston

IBM (2003) Personal communication, chief privacy officer, IBM Zurich

Ishitani L, Almeida V, Wagner J Meira (2003) Masks: Bringing anonymity and personalization together. IEEE Security & Privacy Magazine 1(3):18–23, DOI 10.1109/MSECP.2003.1203218

Jensen C, Potts C, Jensen C (2005) Privacy practices of internet users: Self-reports versus observed behavior. International Journal of Human-Computer Studies 63:203–227, DOI 10.1016/j.ijhcs.2005.04.019

Kelley PG, Bresee J, Cranor LF, Reeder RW (2009) A “nutrition label” for privacy. In: Proceedings of the 5th Symposium on Usable Privacy and Security - SOUPS ’09, Mountain View, California, DOI 10.1145/1572532.1572538

Kelley PG, Cesca L, Bresee J, Cranor LF (2010) Standardizing privacy notices: an online study of the nutrition label approach. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, ACM Press, Atlanta, Georgia, pp 1573–1582, DOI 10.1145/1753326.1753561

Knijnenburg B, Kobsa A, Moritz S, Svensson MA (2011) Exploring the effects of feed-forward and feedback on information disclosure and user experience in a context-aware recommender system. In: UMAP 2011 Workshop on Decision Making and Recommendation Acceptance Issues in Recommender Systems, Girona, Spain, http://www.ics.uci.edu/ kobsa/papers/2011-DEMRA-kobsa.pdf

Kobsa A (2002) Personalized hypermedia and international privacy. Communications of the ACM 45(5):64–67, DOI 10.1145/506218.506249

Kobsa A (2003) A component architecture for dynamically managing privacy constraints in personalized web-based systems. In: Dingledine R (ed) Privacy Enhancing Technologies: 3rd Intl. Workshop, PET 2003, Springer Verlag, pp 177–188

Kobsa A (2007a) Generic user modeling systems. In: Brusilovsky P, Kobsa A, Nejdl W (eds) The Adaptive Web: Methods and Strategies of Web Personalization, Lecture Notes in Computer Science, vol 4321, Springer Verlag, Berlin Heidelberg New York, pp 136–154, DOI 10.1007/978-3-540-72079-9_4

Kobsa A (2007b) Privacy-enhanced web personalization. In: Brusilovsky P, Kobsa A, Nejdl W (eds) The Adaptive Web: Methods and Strategies of Web Personalization, Springer-Verlag, pp 628–670, DOI 10.1007/978-3-540-72079-9_21

Kobsa A, Fink J (2003) Performance evaluation of user modeling servers under real-world workload conditions. In: Brusilovsky P, Corbett AT, Rosis Fd (eds) User Modeling 2003: 9th International Conference, Springer Verlag, pp 143–153, DOI

10.1007/978-3-642-02247-0_10
Kobsa A, Fink J (2006) An LDAP-based user modeling server and its evaluation. User Modeling and User-Adapted Interaction 16(2):129–169, DOI 10.1007/s11257-006-9006-5
Kobsa A, Schreck J (2003) Privacy through pseudonymity in user-adaptive systems. ACM Transactions on Internet Technology 3(2):149–183, DOI 10.1145/767193.767196
Kobsa A, Teltzrow M (2005) Contextualized communication of privacy practices and personalization benefits: Impacts on users' data sharing and purchase behavior. In: Martin D, Serjantov A (eds) Privacy Enhancing Technologies: Fourth International Workshop, PET 2004, Toronto, Canada, vol LNCS 3424, Springer Verlag, Heidelberg, Germany, pp 329–343, DOI 10.1007/11423409_21
Kobsa A, Koenemann J, Pohl W (2001) Personalized hypermedia presentation techniques for improving online customer relationships. The Knowledge Engineering Review 16:111–155, DOI 10.1017/S0269888901000108
Lwin M, Wirtz J, Williams JD (2007) Consumer online privacy concerns and responses: a power–responsibility equilibrium perspective. Journal of the Academy of Marketing Science 35(4):572–585
Malin B, Sweeney L, Newton E (2003) Trail re-identification: Learning who you are from where you have been. Technical Report LIDAP-WP12, Carnegie Mellon University, Laboratory for International Data Privacy, http://privacy.cs.cmu.edu/people/sweeney/trails1.pdf
McJones P (1997) Eachmovie collaborative filtering data set. URL `http://research.compaq.com/SRC/eachmovie/`
Metzger MJ (2004) Privacy, trust, and disclosure: Exploring barriers to electronic commerce. Journal of Computer-Mediated Communication 9(4), http://jcmc.indiana.edu/vol9/issue4/metzger.html
Miller BN, Konstan JA, Riedl J (2004) PocketLens: toward a personal recommender system. ACM Trans Inf Syst 22(3):437–476, DOI 10.1145/1010614.1010618
Morenoff E, McLean JB (1967) Application of level changing to a multilevel storage organization. Commun ACM 10:149–154, DOI 10.1145/363162.363183
Movielens (1997) Movielens - movie recommendations. http://www.movielens.org/, URL `http://www.movielens.org/`
Mulligan D, Schwartz A (2000) Your place or mine? privacy concerns and solutions for server and client-side storage of personal information. In: Proceedings of the tenth conference on Computers, freedom and privacy: challenging the assumptions, ACM, Toronto, Ontario, Canada, pp 81–84, DOI 10.1145/332186.332255
Nakashima E (2006) AOL search queries open window onto users' worlds. URL `http://www.washingtonpost.com/wp-dyn/content/article/2006/08/16/AR2006081601751.html`
Narayanan A, Shmatikov V (2008) Robust de-anonymization of large sparse datasets. In: SP '08: Proceedings of the 2008 IEEE Symposium on Security and Privacy, IEEE Computer Society, Washington, DC, USA, pp 111–125, DOI http://dx.doi.org/10.1109/SP.2008.33
Narayanan A, Shmatikov V (2010) Myths and fallacies of "personally identifiable information". Commun ACM 53(6):24–26, DOI 10.1145/1743546.1743558

Nissenbaum HF (2010) Privacy in context: technology, policy, and the integrity of social life. Stanford Law Books, Stanford, CA

Ohm P (2010) Broken promises of privacy: Responding to the surprising failure of anonymization. UCLA Law Review 57(6):1701–1777, URL http://uclalawreview.org/pdf/57-6-3.pdf

Opler A (1965) Dynamic flow of programs and data through hierarchical storage. In: Kalenich WA (ed) Information Processing 1965, Proceedings of IFIP Congress, New York, NY, vol 1, pp 273–276

Palen L, Dourish P (2002) Unpacking "privacy" for a networked world. In: CHI-02, Fort Lauderdale, FL, pp 129–136

Pew (2008) Privacy implications of fast, mobile internet access. http://www.pewinternet.org/Reports/2008/Privacy-Implications-of-Fast-Mobile-Internet-Access.aspx

Polat H, Du W (2003) Privacy-Preserving collaborative filtering using randomized perturbation techniques. In: IEEE International Conference on Data Mining, IEEE Computer Society, Los Alamitos, CA, USA, pp 625–628, DOI 10.1109/ICDM.2003.1250993

Polat H, Du W (2005) SVD-based collaborative filtering with privacy. In: Proceedings of the 2005 ACM Symposium on Applied Computing, ACM, Santa Fe, New Mexico, pp 791–795, DOI 10.1145/1066677.1066860

Rao JR, Rohatgi P (2000) Can pseudonymity really guarantee privacy? In: 9th USENIX Security Symposium, Denver, CO, pp 85–96, www.usenix.org/events/sec00/full_papers/rao/rao.pdf

Schafer J, Frankowski D, Herlocker J, Sen S (2007) Collaborative filtering recommender systems. In: Brusilovsky P, Kobsa A, Nejdl W (eds) The Adaptive Web, Lecture Notes in Computer Science, Springer Verlag, Berlin Heidelberg New York, pp 291–324, DOI 10.1007/978-3-540-72079-9_9

Schoenbachler DD, Gordon GL (2002) Trust and customer willingness to provide information in database-driven relationship marketing. Journal of Interactive Marketing 16(3):2–16, DOI 10.1002/dir.10033

Shankland S (2009) Google uncloaks once-secret server. http://news.cnet.com/8301-1001_3-10209580-92.html

Smith HJ, Milberg SJ, Burke SJ (1996) Information privacy: Measuring individuals' concerns about organizational practices. MIS Quarterly 20(2):167–196, URL http://www.jstor.org/stable/249477

Spiekermann S, Grossklags J, Berendt B (2001) E-privacy in 2nd generation e-commerce: Privacy preferences versus actual behavior. In: EC'01: Third ACM Conference on Electronic Commerce, Tampa, FL, pp 38–47

Sweeney L (2000) Uniqueness of simple demographics in the U.S. population. Technical report LIDAPWP4, Carnegie Mellon University, Laboratory for International Data Privacy

Teltzrow M, Kobsa A (2004) Impacts of user privacy preferences on personalized systems: a comparative study. In: Karat CM, Blom J, Karat J (eds) Designing Personalized User Experiences for eCommerce, Kluwer Academic Publishers, Dordrecht, Netherlands, pp 315–332, DOI 10.1007/1-4020-2148-8_17

TRUSTe (2010) Web privacy seal program requirements. Http://www.truste.com/pdf/Web_Privacy_Seal_Program_Requirements_Website.pdf

Turow J, King J, Hoofnagle CJ, Bleakley A, Hennessy M (2009) Americans reject tailored advertising and three activities that enable it. SSRN eLibrary http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1478214

Wang Y, Kobsa A (2006) Impacts of privacy laws and regulations on personalized systems. In: Kobsa A, Chellappa R, Spiekermann S (eds) Proceedings of PEP06, CHI 2006 Workshop on Privacy-Enhanced Personalization, ACM, pp 44–46, URL http://www.ics.uci.edu/~kobsa/papers/2006-PEP-wang-kobsa.pdf

Wang Y, Kobsa A (2007) Respecting users' individual privacy constraints in web personalization. In: Conati C, McCoy KF, Paliouras G (eds) User Modeling 2007: 11th Intl. Conf., Springer, pp 157–166, DOI 10.1007/978-3-540-73078-1_19

Wang Y, Kobsa A (2009a) Performance evaluation of a Privacy-Enhancing framework for personalized websites. In: Houben G, McCalla G, Pianesi F, Zancanaro M (eds) User Modeling, Adaptation, and Personalization: 17th International Conference, UMAP 2009, vol 5535, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 78–89, DOI 10.1007/978-3-642-02247-0_10

Wang Y, Kobsa A (2009b) Privacy-enhancing technologies. In: Gupta M, Sharman R (eds) Social and Organizational Liabilities in Information Security, IGI Global, pp 203–227, DOI 10.4018/978-1-60566-132-2.ch013

Wang Y, Kobsa A, van der Hoek A, White J (2006) PLA-based runtime dynamism in support of privacy-enhanced web personalization. In: SPLC'06: Proceedings of the 10th International Software Product Line Conference, IEEE Press, pp 151–162, DOI 10.1109/SPLC.2006.30

Xie E, Teo H, Wan W (2006) Volunteering personal information on the internet: Effects of reputation, privacy notices, and rewards on online consumer behavior. Marketing Letters 17(1):61–74, DOI 10.1007/s11002-006-4147-1

Young J (1978) Introduction: A look at privacy. In: Young J (ed) Privacy, John Wiley and Sons, New York

Zadorozhny V, Yudelson M, Brusilovsky P (2008) A framework for performance evaluation of user modeling servers for web applications. Web Intelli and Agent Sys 6(2):175–191, DOI 10.3233/WIA-2008-0136

Author Biographies

(1) DR. YANG WANG

Carnegie Mellon University, School of Computer Science, 4720 Forbes Ave., Pittsburgh, PA 15213, USA

Dr. Yang Wang is a Research Scientist in CyLab at Carnegie Mellon University, working in the area of usable privacy and security, personalization, and social computing. He received his Ph.D. in Information and Computer Science from University of California, Irvine.

(2) DR. ALFRED KOBSA

University of California, Irvine, School of Information and Computer Sciences, Irvine, CA 92697, USA

Dr. Alfred Kobsa is a Professor Alfred Kobsa in the Donald Bren School of Information and Computer Sciences of the University of California, Irvine. His research lies in the areas of user modeling and personalized systems, privacy, and information visualization. He is the editor of User Modeling and User-Adapted Interaction: The Journal of Personalization Research, and editorial board member of the Springer Lecture Notes in Computer Science and of three scientific journals. He edited several books and authored numerous publications in the areas of user-adaptive systems, human-computer interaction and knowledge representation.