UNIVERSITY OF CALIFORNIA,
IRVINE

# A Framework for Privacy-Enhanced Personalization

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Information and Computer Science

by

Yang Wang

Dissertation Committee:
Professor Alfred Kobsa, Chair
Professor André van der Hoek
Professor Gene Tsudik

2010

# DEDICATION

To my mother Zhengying Zhang
To my father Linbang Wang
To my wife Yun Huang
To my son Hengrui Wang

# TABLE OF CONTENTS

# LIST OF FIGURES

Page

# LIST OF TABLES

Page

# ACKNOWLEDGMENTS

This is for Alfred Kobsa, who ushered me to the wonderland of scientific research. Alfred took me when I was just a fresh graduate student with little research experience and patiently trained me over the years into a researcher. Getting a Ph.D. is not easy, but Alfred made it easier and enjoyable. I have always been inspired by his wisdom and rigor of research and his generosity and kindness. Alfred is not only a great advisor for research, but also a life-long mentor for just about anything. I am so fortunate that I have found the perfect advisor.

This is for André van der Hoek. If you want an easy committee member, do not ask André. But if you seek someone who can inspire you and provide insightful and critical comments, ask André. Yes, the rumor is true - no matter how busy he is, he will read your survey paper, topic proposal and dissertation. Be prepared, he will hit the bull's eye. André encourages and pushes people to reach their potentials. I am so glad and grateful that he did that to me.

This is for Gene Tsudik, a crypto guru. His intriguing class on security sparked my new enthusiasm in cryptography and security. Gene always brings a different and invaluable perspective to my research. His congenial and witty personality makes it such a pleasure to work with him. I am also thankful to work with him on the secure device paring project which broadens my horizon on security research.

This is for David Redmiles and Sanjeev Dewan who kindly served on my candidacy committee. Their thought-provoking questions and comments have shaped my research.

This is for Gloria Mark who taught me quantitative research and showed me doing research can have so much fun. I am also indebted to her generosity for paying me to attend my first academic conference out of her own pocket! I was just an co-author of a paper (my first academic publication!) based on her class project.

This is for Paul Dourish who taught me qualitative research and ethnography. I was not the best student, but his ethnography class was one of my most favorite classes at UCI. His class and writings have equipped me with new lenses and perspectives to look at and think about technologies. I am also very grateful for Paul's prompt and thoughtful advices and warm-hearted help in my ethnographic research and my job hunting and interviewing.

This is for Gary Olson and Judy Olson. Gary decided to take me to work with him on the digital photo frame project. Gary's knowledge and friendliness make everyone want to work with him. I also really appreciate that Gary spent a lot of time writing customized recommendation letters for me, especially given that he is not in my committee and has only worked with me for a relatively short period of time. Judy is as fabulous as Gary. They are both great role models for me and other young researchers - being research celebrities, they are so modest and amiable.

This is for David Kay who taught me how to be a good teacher. I am so appreciative of David's materials and advices that helped me ace my debut as a course instructor.

This is for Cristina Lopes, Bonnie Nardi, Richard Taylor, Don Patterson, Gillian Hayes, Yunan Chen, Debra Richardson, Hadar Ziv, Jason Robbins, Alex Thornton, Dan Frost, and the rest of faculty in Informatics. They have either collaborated with me or provided me with help and advices in advancing my academic career.

This is for Robert Newcomb, David van Dyk, Hal Stern, Daniel Gillen and the rest of faculty in Statistics. Most of Bob's stats classes started at 8am. I found myself wake up early (at least by my standard) with energy and excitement to go to Bob's classes. They instilled my passion into number crunching and hidden pattern hunting. David is more than a good teacher - he became my friend and long-term stats consultant. Hal's and Daniel's classes were exemplary - they really made me understand a lot of mathematical things under the hood of formulas and statistical tests and equipped me with skills to tackle practical data analysis problems.

This is for Bill Maurer, Tom Boellstorff, George Marcus and many other people in Anthropology. They helped me develop an anthropological sense.

This is for Suzanne Schaffer, my old office mate, life-long mentor and friend. I was so fortunate to sit in the same office with Suzanne for about 2 years in the ICS2 building. Her encouragement, kindness, wisdom, humor and support have always been there throughout my graduate school. I only wish I could learn more from her.

This is for Norman Su, David Nyugen, Bo Xing and Tiandan Chen. They are my buddies in graduate school. They made my Ph.D. days so much more enjoyable and memorable. They have always been reliable sources of support over the years. Norman takes pain to proofread and comment on my writings, and to listen to my countless practice talks. He also generously provided some of his Ruby code which I used in the system implementation for my user experiment. His music performance has joyfully entertained and enlightened my family. David is such a nice person and gentleman. No wonder why everybody likes him. David demonstrates to me how to be an earnest friend. David has always been a great colleague to discuss my research. His insights live in this dissertation. Bo and Tiandan are my conspirators. I always cherish our friendship.

This is for Sameer Patil and Xinru Page, my fellow group members. I really appreciate their support and kindness and look forward to collaborating with them in the future.

This is for Scott Hendrickson, Eric Dashofy, Ping Chen, Chris van der Westhuizen and the rest of the ArchStudio team. ArchStudio is such a great tool and I am proud of being a user.

This is for Jeff White, Sulagna Basu, and Deepika Gandhi. Their hard work has greatly helped my research.

This is for Xianghua Ding, Lihua Xu, Steve Abrams, Chris Jensen, Anita Sarma, Charlotte Lee, Patrick Shih, Bryan Semaan, Eric Baumer, Jennifer Rode, Amanda Williams, Johanna Brewer, Judy Chen, Lilly Irani, Eric Kabisch, Silvia Lindtner, Marisa Cohn, Yong Ming Kow, Jahmeilah Roberson, Ruy Cervantes, Nithya Sambasivan, Yongjie Zheng, Kristina

Winbladh, Jose Romero-Mariona, Alegria Baquero, Wenhui Zhang, Xin Zheng, Jun Lu, Di Ma, Mohamad Monibi, Miya Sylvester, Ben Koehne, Erik Trainer, Justin Chung, Ellie Harmon, Jed Brubaker, Sen Hirano, and many other colleagues and friends at ICS.

This is for John Chen, Chimin Chen, Esther Lin, Paul Lin, Joanna Ho, Jeffery Chiou, Lilly Wu, Yang Zhang, Mary Yan, Wei Zhang, and many others who inspire and guide my life.

This is for Yi Pan, Xiangrong Li, Liang Zhang, Yunhong Liu, Ziwei Xie, Pei Zhang, Zhen Bian, Jerry Peng, Liang Wu, Bill Gao, David Yang, Hongkai Zhao, Jack Shen, Joe Zhou, Qi Song, Muqing Lin, Kai Wang, Xiangwei Zhu, Mingbo Wang, Jie Kang, Walter Yuan, Wesley Sha and many other soccer teammates and friends. This is for the UCI Men's soccer club team. It was such a memorable experience to be part of the team. Without soccer and them, I cannot possibly finish my Ph.D.

This is for Song He, Lihua Li, Jun Wen, Jin Zuo, Jin Chen, Xinyu Chen, Yin Wang, and many other UCI badminton friends.

This is for Scott Mainwaring, my great mentor and close friend. Scott believed in my research enthusiasm in virtual currency and took me as an intern in the summer of 2007. I had such a great time doing field work with Scott in China. I am so fortunate to get to know him and work with him. His knowledge, wisdom and humor are always inspiring. I also want to thank Wendy March, Ken Anderson, Linda Illig, Maria Bezaitis and the rest of PAPR at Intel Research for their support and kindness.

This is for Daniel Billsus and Dave Hilbert, my great mentors at FXPAL. Daniel and Dave provided me with freedom and support to allow me to explore the internship research project. I only wish I could work and learn more from these two bright and musical UCI/ICS alumni. I also want to thank Jeremy Pickens, Laurent Denoue, Kandha Sankarapandian, Thea Turner, Gene Golovchinsky, Monica Gorkani, Qiong Liu, Tony Dunnigan, John Adcock, Bee Liew, Lynn Wilcox, my fellow intern friends and the rest of FXPAL.

This is for Rohit Khare, my great mentor at CommerceNet. Rohit is such a brilliant technologist. I only wish I could work and learn more from this bright and entrepreneurial UCI/ICS alumnus. I also would like to thank Kragen Sitaker, Chris Hibbert, Allan Schiffman, Jay (Marty) Tenenbaum, and the rest of CommerceNet.

This is for Kiana Fallah, Debra Brodbeck, Kris Bolcer, Gina Anzivino, Wendy Wehofer, Carol Rapp, Milena Wypchlak, Marty Beach, Nancy Myers, Sabina Jun, Donna Hedman, Carolyn Simpson, Susan Knight, and many other previous/current ICS staff for their wholehearted support.

This is for Lynn Hammeras, Lisa Martinez, teacher Armando, Holly, Kelly, Vi, Lisa, James, Jennifer, Christina, and many other teachers and student assistants at the UCI Child Care. They provided my son Hengrui Wang a peaceful and enjoyable environment to grow up and learn so that I could spend more time to concentrate on my research.

This is for all the people who participated in my experiment. I really appreciate their time and efforts in helping me evaluate my research.

This is for Kaiyong Huang, Zhaorong Wang, Yan Huang, Le Quan, Bo Min, and my other extended family members.

This is for Michael Poots, Esther Poots, Sean Huang, Kan Dai, Xuelong Wang, Hongyu Liu, Hongyu Zhou, Yu Tian, Manle Lin, Yunfeng Li, Dong Wang, Shen Ye, Bing Hu, Yin Wu, Jia Fan, Li An, and many other close classmates and friends.

This is for Wei Li, Tim Newman, Mary Weisskopf, Minghua Xiao, Shen Li, Runtian Luo, Maozhi Wang, Jinliang Jie, Zhongli Zhou, Yin Rong, Shiqian Guo, Darong Dai, and many other teachers in my life.

I would like to acknowledge several funding sources for supporting my work: the NSF grants IIS-0308277, CNS-0831526 and IIS-0808783, a Google Research Grant, the UCI Graduate Dean's Dissertation Fellowship, the Phi Beta Kappa International Scholarship awarded by the Alpha Association, Phi Beta Kappa Alumni of Southern California. Many people such as Alfred Kobsa, Gene Tsudik, Kiana Fallah, Amelia Regan, Kris Bolcer, Jonathan Lew, Long Hoang Nguyen, Cecilia Conrad, Alexei Maradudin, and Peggy Maradudin have done a lot of hard work to make these grants possible.

**Preliminary results of this dissertation were published as the following:**

Y. Wang, A. Kobsa (2009): Performance Evaluation of a Dynamic Privacy-Enhancing Framework for Personalized Websites. Intl. Conf. User Modeling, Adaptation, and Personalization (UMAP09), pp. 78-89.

S. A. Hendrickson, Y. Wang, A. van der Hoek, R. N. Taylor, A. Kobsa (2009): Modeling PLA Variation of Privacy-Enhancing Personalized Systems. Intl. Software Product Line Conference (SPLC09), pp. 71-80.

Y. Wang, A. Kobsa (2007): Respecting User's Individual Privacy Constraints in Web Personalization. Intl. Conf. User Modeling (UM07), pp. 157-166.

Y. Wang, A. Kobsa, A. van der Hoek, J. White (2006): PLA-based Runtime Dynamism in Support of Privacy-Enhanced Web Personalization. IEEE Conf. Software Product Line (SPLC06), pp. 151-162.

Y. Wang, A. Kobsa (2009): Privacy Enhancing Technologies. In M. Gupta and R. Sharman, eds.: Handbook of Research on Social and Organizational Liabilities in Information Security. Hershey, PA: IGI Global.

Y. Wang, A. Kobsa (2008): Technical Solutions for Privacy-Enhanced Personalization. In Constantinos Mourlas and Panagiotis Germanakos, eds.: Intelligent User Interfaces: Adaptation and Personalization Systems and Technologies. Hershey, PA: IGI Global.

Y. Wang (2006). Impacts of Privacy Laws and Regulations on Personalized Web-based Systems. ACM CHI06 Workshop on Privacy-Enhanced Personalization.

# CURRICULUM VITAE

## Yang Wang

### EDUCATION

**Doctor of Philosophy in Information and Computer Sciences** **2010**
University of California, Irvine *Irvine, California*

**Master of Science in Information and Computer Sciences** **2005**
University of California, Irvine *Irvine, California*

**Bachelor of Science in Computer Science** **2001**
Chengdu University of Technology *Chengdu, China*

### RESEARCH INTERESTS

Human-Computer Interaction
Computer-Supported Collaborative Work
Privacy and Security
Social Computing
User Modeling, Adaptation, and Personalization
Software Engineering

### RESEARCH EXPERIENCE

**Graduate Student Researcher** **9/2004–3/2010**
Department of Informatics, UC Irvine *Irvine, California*
Supervisor: Professor Alfred Kobsa

**Research Internship** **6/2007–9/2007**
People and Practices Research, Intel Labs *Beaverton, OR*
Supervisor: Dr. Scott Mainwaring

**Research Internship** **6/2006–9/2006**
Corporate Memory group, Fuji Xerox Palo Alto Laboratory (FXPAL) *Palo Alto, CA*
Supervisor: Dr. Daniel Billsus and Dr. David Hilbert

**Research Internship** **6/2005–9/2005**
zLab, CommerceNet *Palo Alto, CA*
Supervisor: Dr. Rohit Khare

## TEACHING EXPERIENCE

**Lead Instructor** **8/2009–9/2009**
University of California, Irvine *Irvine, California*
INF131: Intro to HCI

**Teaching Assistant** **2004–2009**
University of California, Irvine *Irvine, California*
Many undergraduate classes in Software Engineering and HCI

## PROFESSIONAL EXPERIENCE

**Mobile System Engineer** **11/2001–5/2002**
Chengdu Bell *Chengdu, China*

## REFEREED CONFERENCE PUBLICATIONS

Y. Wang, Scott Mainwaring (2010): Incentives in the Wild: Leveraging Virtual Currency to Sustain Online Community. To appear in iConference 2010.

Y. Wang, A. Kobsa (2010): Privacy in Cross-System Personalization. To appear in AAAI Privacy 2010 Symposium (Privacy2010).

Y. Wang, A. Kobsa (2009): Performance Evaluation of a Dynamic Privacy-Enhancing Framework for Personalized Websites. Intl. Conf. User Modeling, Adaptation, and Personalization (UMAP09), pp. 78-89.

S. A. Hendrickson, Y. Wang, A. van der Hoek, R. N. Taylor, A. Kobsa (2009): Modeling PLA Variation of Privacy-Enhancing Personalized Systems. Intl. Software Product Line Conference (SPLC09), pp. 71-80.

S. Lindtner, S. Mainwaring, P. Dourish, Y. Wang (2009): Situating Productive Play: Online Gaming Practices and Guanxi in China. IFIP Conf. on Human-Computer Interaction (INTERACT09), pp. 328-341.

A. Kobsa, R. Sonawalla, G. Tsudik, E. Uzun and Y. Wang: Serial Hook-Ups: A Comparative Usability Study of Secure Device Pairing Methods. Symp. on Usable Privacy and Security (SOUPS09), pp. 1-12.

S. Lindtner, B. Nardi, Y. Wang, S. Mainwaring, J. He, and W. Liang (2008): A Hybrid Cultural Ecology: World of Warcraft in China. ACM Conf. Computer Supported Cooperative Work (CSCW08), pp. 371-382.

Y. Wang, S. Mainwaring (2008): Human-Currency Interaction: Learning from Virtual Currency Use in China. ACM Conf. Human Factors in Computer Systems (CHI08), pp. 25-28.

Y. Wang, A. Kobsa (2007): Respecting User's Individual Privacy Constraints in Web Personalization. Intl. Conf. User Modeling (UM07), pp. 157-166.

Y. Wang, A. Kobsa, A. van der Hoek, J. White (2006): PLA-based Runtime Dynamism in Support of Privacy-Enhanced Web Personalization. IEEE Conf. Software Product Line (SPLC06), pp. 151-162.

N. M. Su, Y. Wang, G. Mark, T. Aieylokun, T. Nakano (2005): A Bosom Buddy Afar Brings a Distant Land Near: Are Bloggers a Global Community? Intl. Conf. Communities and Technology (C&T05), pp. 171-190.

## REFEREED BOOK CHAPTERS

Y. Wang, A. Kobsa (2009): Privacy Enhancing Technologies. In M. Gupta and R. Sharman, eds.: Handbook of Research on Social and Organizational Liabilities in Information Security. Hershey, PA: IGI Global.

Y. Wang, A. Kobsa (2008): Technical Solutions for Privacy-Enhanced Personalization. In Constantinos Mourlas and Panagiotis Germanakos, eds.: Intelligent User Interfaces: Adaptation and Personalization Systems and Technologies. Hershey, PA: IGI Global.

## REFEREED WORKSHOP AND DOCTORAL CONSORTIUM PAPERS

Y. Wang, A. Kobsa (2009). Privacy in Online Social Networking at Workplace. IEEE SocialCom09 Workshop on Security and Privacy in Social Networking.

Y. Wang, S. Mainwaring (2008). Ethnography at Play: An Exploratory Case Study of Chinese Users' Experience in Online Games. ACM CHI08 Workshop on Evaluating User Experiences in Games.

Y. Huang, N. Venkatasubramanian and Y. Wang (2007). MAPGrid: A New Architecture for Empowering Mobile Data Placement in Grid Environments. IEEE CCGrid07 Workshop on Context-Awareness and Mobility in Grid Computing.

Y. Wang (2006). Impacts of Privacy Laws and Regulations on Personalized Web-based Systems. ACM CHI06 Workshop on Privacy-Enhanced Personalization.

Y. Wang, A. Kobsa (2005). A Software Product Line Approach for Handling Privacy Constraints in Web Personalization. UM05 Workshop on Privacy-Enhanced Personalization.

N. M. Su, Y. Wang, G. Mark (2005). Politics as Usual in the Blogosphere. Intl. Workshop on Social Intelligence Design (SID05), Palo Alto, USA.

Y. Wang (2005). Constraint-Sensitive Privacy Management for Personalized Web-based Systems. Doctoral Consortium of the 10th Intl. Conference on User Modeling (UM05).

## MEDIA COVERAGE

Olga Kharif, “Virtual Currencies Gain in Popularity”, **BusinessWeek**, May 6, 2009.

Luca Chittaro, “The Color of The Penny”, **Il Sole 24 Ore** (major italian financial newspaper, circulation: 375’000 copies), April 3, 2008.

## AWARDS

Phi Beta Kappa International Scholarship, 2009

UCI Graduate Dean’s Dissertation Fellowship, 2009

NSF travel awards UM05, UM07, UMAP09

Graduate Student Fellowship, UCI, 2003

First Prize, National English Contest for College Students, China, 2000

Sichuan Province “All-round College Student Certificate”, Class A, China, 2000

Honorable Mention Award, National Mathematical Modeling Contest, China, 1999

CDUT Outstanding Student of Year 1998 and 2000, China

First Prize, CDUT English Speech Competition, China, 1999

# ABSTRACT OF THE DISSERTATION

A Framework for Privacy-Enhanced Personalization

By

Yang Wang

Doctor of Philosophy in Information and Computer Science

University of California, Irvine, 2010

Professor Alfred Kobsa, Chair

Web personalization has demonstrated to be advantageous for both online customers and vendors. However, its benefits may be severely counteracted by privacy constraints. Personalized systems need to address these privacy constraints, including users' personal privacy preferences as well as privacy laws and industry self-regulations that may be in effect.

This research aims to reconcile privacy and web personalization. In particular, this research proposes user-tailored privacy enforcement in personalization, i.e., catering privacy to the situation of each individual user. This research develops a framework for privacy-enhanced personalization based on product line architecture (PLA). This framework provides a systematic and flexible way to model, manage and enforce applicable privacy constraints in web personalization. Our systematic analysis of over 40 privacy laws shows that these legal requirements can not only affect the collection, storage and transfer of personal data, but also the methods that may be used to process the data in personalized systems. This framework dynamically selects and de-selects personalization methods during runtime in observance of each user's applicable privacy constraints.

A controlled experiment shows that users value the user-tailored privacy enforcement in personalization: they have higher regard for the privacy practices of a personalized system, disclose more information about themselves to the system, and do not perceive less personalization benefit. Besides, a performance evaluation of our framework shows that it is technically feasible for internationally-operating personalized sites even with the highest traffic today.

# Chapter 1

# Introduction

## 1.1 Background and Motivation

Personalized (or user-adaptive) systems cater their interaction to each user based on users' individual characteristics (Kobsa, 2003). To do that, personalized systems collect a large amount of user data (usually in an unobtrusive way) and "lay them in stock" for future adaptation. Generally, the more information becomes collected about users, the better will be the quality of personalization. Personalized systems employ various personalization methods (Kobsa et al., 2001), for instance, machine learning methods (Webb et al., 2001), to derive additional assumptions about users. The collected user data and derived assumptions about users are often used to build user profiles (a.k.a., user models), based on which services are personalized. As such, the process of personalization usually consists of user modeling and adaptation. User modeling is concerned with building user models, while adaptation creates personalized services based on the user models.

Personalization research has a root in Artificial Intelligence (AI) and early work has been traditionally applied in areas such as plan recognition (Allen, 1979), dialog systems (Kobsa and Wahlster, 1989) and tutoring systems (Kass, 1989). More recently, personalized systems have flourished on the World Wide Web. We use the term *web personalization* to denote personalized services on the web. Nowadays, web personalization can be found in various types of applications such as entertainment (Movielens, 1997), e-commerce (Goy et al., 2007), education (Brusilovsky and Millán, 2007), news (Billsus and Pazzani, 2007), healthcare (Cawsey et al., 2007), search (Micarelli et al., 2007) and collaboration (Soller, 2007). Concrete examples of personalized services include customized content (e.g., personalized finance pages or news collections), customized recommendations of movies and music, customized advertisements based on past purchase behavior, customized (preferred) pricing, customized form of representation (e.g., audio, text), tailored email alerts, and express transactions.

Industry reports show evidence that providing personalization brings websites many benefits such as attracting more users and improving brand loyalty (Hof et al., 1998; Cooperstein et al., 1999; Hagen et al., 1999). Consumer studies also indicate that users value personalized services on the Web (Personalization Consortium, 2001; Tam and Ho, 2003; ChoiceStream, 2005).

However, this win-win situation is undermined by privacy concerns (IBM, 1999; Forrester Research, 1999; DePallo, 2000; Teltzrow and Kobsa, 2004; Turow et al., 2009). For instance, a user who enjoys the personalized book recommendations provided by Amazon is likely to feel uneasy when her profile information is shared with others. Web users are not only concerned about disclosing their data online, but have also acted against websites' data collection practices (e.g., by leaving websites that require registration information or by entering fake registration information). Since personalized systems collect personal data, they are also subject to privacy laws and regulations if the respective individuals are in principle identifiable. Our analysis of more than 40 privacy laws (Wang et al., 2006a) shows that if privacy laws apply to a personalized website, they often not only affect the data that are collected by the website, the way in which the data is transferred and to which party it is transferred, but also the methods that may be used for processing them. For instance, the German Telemedia Law of 2007 (DE-TML, 2007) mandates personal data to be erased immediately after each session except for very limited purposes. This provision would affect the use of those machine learning methods where the learning takes place over several sessions.

To reconcile privacy and web personalization, two main types of privacy constraints need to be taken into consideration in designing and implementing web-based personalized systems: regulatory privacy requirements set out by various privacy laws and regulations, and users' personal privacy preferences/needs. Although regulatory privacy requirements are subject to amendments and revisions, the changes usually take a relatively long period of time (at least months, typically years). In contrast, users' privacy preferences/needs are highly situated, flexible and contingent (Palen and Dourish, 2002). This research focuses on understanding how both kinds of privacy constraints might affect the ways in which web-based personalized systems operate, and on developing technical solutions that aim to handle these privacy constraints without unduly compromising web personalization.

## 1.2 Research Overview

The overarching goal of this research is to reconcile privacy and web personalization. More specifically, this research aims at respecting various privacy constraints stemming from privacy laws and regulations as well as users' personal privacy preferences, while at the same time providing high-quality web personalization. The scope of this research and its research hypotheses will be discussed in this section.

### 1.2.1 Scope

Figure 1.1 shows a generic (application-independent) user modeling architecture proposed by Kobsa and Fink (2006). It consists of two main functional components: a user modeling system and a set of user-adaptive applications. The former is responsible for building and maintaining user models that represent and store different information/assumptions about users such as their beliefs, goals, plans and preferences. It also has reasoning capabilities (i.e., user modeling methods/components) for deriving additional assumptions based on existing ones. We define this as the *user data reasoning* aspect of personalized systems. The latter generates and delivers user-adaptive/personalized services (e.g., personalized book recommendations) to Internet users based on their information retrieved from the user modeling system.

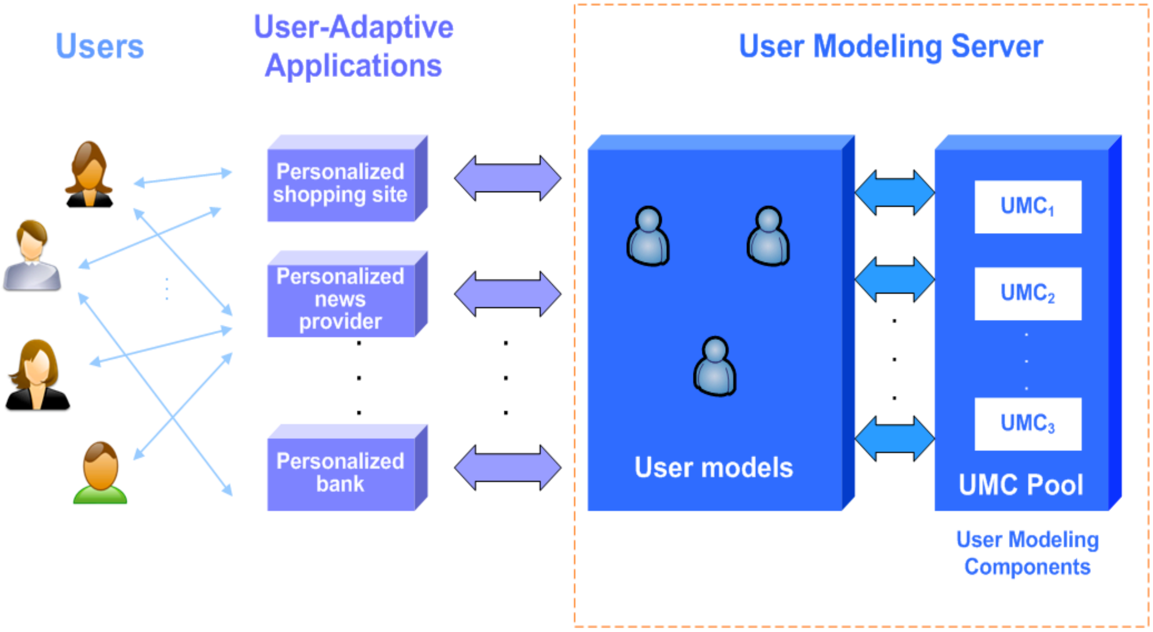


Figure 1.1: A Generic User Modeling Architecture and The Focus of This Research

A user modeling system can be implemented as part of a user-adaptive application. In this case, the user modeling system is usually called as a user modeling shell (Kobsa, 1990). Alternatively, the user modeling facility can be implemented as an independent user modeling server (UMS) (Kobsa, 2001) that can serve different user-adaptive applications at the same time. Conceptually, a UMS, the central venue of storage of users' information, provides a number of advantages such as user information acquired by one application can be used by other applications and vice versa (see (Kobsa, 2001) for more detailed discussions about the advantages). Physically, however, a UMS can be "centralized" (e.g., on a single machine or platform), or "distributed" across several machines or platforms to improve its availability and performance. Note that the distribution of a UMS is different from the notion of client-side personalization. In client-side personalization, user information is usually stored, and all user modeling and personalization carried out, at the user's machine so that the user has control over the personal data and their usage. In contrast, in physically distributed UMS, data is usually stored, and all user modeling and personalization performed, in machines controlled by the user modeling system rather than the user.

A large number of UMS have been developed and are widely employed for both academic and commercial purposes. Because of their heavy usage and critical role played in today's web personalization, this research mainly focuses on developing intelligent and effective privacy enhancement mechanisms for them. That is to say, from an architectural point of view (see Figure 1.1), this research does not focus on privacy enhancement in the interactions between end users and user-adaptive applications (e.g., privacy-enhancing user interfaces), nor on privacy enhancement in the interactions between user-adaptive applications and the UMS (e.g., access control of what user information can be acquired by a particular application). Instead, this research concentrates on how privacy constraints might affect the internal operations of a UMS (highlighted by a color box in Figure 1.1).

When privacy laws are applicable to a personalized website, they often not only affect the data that are collected by the website, but also the methods that may be used for processing the users' data (Kobsa, 2003; Wang et al., 2006a). This research is primarily concerned with the perspective of reasoning about users' data. More specifically, this research focuses on how privacy constraints might affect the usage of user modeling methods in a user modeling server and on mechanisms to responsively respect and handle applicable privacy constraints (e.g., changes of users' privacy preferences/decisions) in using permissible user modeling methods.

### 1.2.2 Hypotheses

In order to achieve our goal of reconciling privacy and personalization, we seek to develop a solution that (1) brings privacy and personalization values to end users, and (2) is technically feasible so that contemporary personalized sites even ones with heavy traffic can adopt our solution. To test our approach, this research formulates and examines the following hypotheses. We thereby define privacy constraints as legal and regulatory privacy requirements as well as the user's personal privacy preferences. We also use the term *personalized privacy-aware user modeling* to denote the practice of respecting applicable privacy constraints for each user when employing user modeling methods in web-based personalized systems.

- **Hypothesis 1**: *If a web-based personalized system respects applicable privacy constraints for each user in its usage of personalization methods, and informs its users about this privacy-aware practice, then users will have higher regard for the privacy practices of the system.*

  This hypothesis is important because users' privacy concerns are impediments of web personalization. Chellappa and Sin (2005) found that users' stated intention to use personalized services is negatively influenced by their privacy concerns. If users have higher regard for the privacy practices of a personalized system, they are more likely to embrace and engage with web personalization.

- **Hypothesis 2**: *If a web-based personalized system respects applicable privacy con-*

*straints for each user in its usage of personalization methods, and informs its users about this privacy-aware practice, then users will disclose more information about themselves to the system.*

This hypothesis puts forward since users' information is integral in generating web personalization (see Section 1.1). Generally, the more information users disclose about themselves, the more personalized systems know about the users, and the better will be the quality of web personalization.

- **Hypothesis 3**: *If a web-based personalized system respects applicable privacy constraints for each user in its usage of personalization methods, and informs its users about this privacy-aware practice, then users will be more likely to exhibit other privacy-sensitive behaviors, e.g., make online purchases.*

  Besides self-disclosure of their data to the personalized sites, users typically exhibit additional privacy-related behavior such as leaving personalized sites requiring registration, providing fake information for registration and making purchases on personalized e-commerce sites. In online purchases, users need to provide their financial information (e.g., credit card number or online bank account) which are highly sensitive personal information that users are reluctant to disclose (Ackerman et al., 1999; IBM, 1999). The examination of this hypothesis will shed light on how privacy-aware user modeling would affect users' disclosure of sensitive information about themselves. Particularly interesting is the fact that the information that is needed for an online purchase is difficult to fake, in the sense that if a user fakes it (e.g., the credit card number), the purchase or transaction may be invalidated. In comparison, users can easily fake other sensitive personal information such as their income.

- **Hypothesis 4**: *If a web-based personalized system respects applicable privacy constraints for each user in its usage of personalization methods, and informs its users about this privacy-aware practice, then this will not compromise users' perceived* *personalization benefits.*

Hypothesis 4 is important because the goal of this research is to reconcile privacy and personalization. If the privacy-aware user modeling does compromise users' perceived benefits of web personalization, then it fails to strike a good balance between privacy and personalization.

- **Hypothesis 5**: *Respecting applicable privacy constraints for each user in the usage of personalization methods is technically possible with reasonable computing resources even for contemporary personalized sites with heavy traffic.*

  The last hypothesis is concerned with the scalability of personalized privacy-aware user modeling, i.e., whether it can be carried out with contemporary personalized sites, and even those with heavy traffic.

In order to achieve a better understanding of the background of these hypotheses, this research asks the following questions:

- Q1: What are the privacy constraints in the domain of web personalization?
- Q2: How do these privacy constraints affect the internal operations of web-based personalized systems, particularly the usage of user modeling methods?
- Q3: Which privacy constraints or what aspects of them are under-addressed in current web-based personalized systems?
- Q4: In what ways can these aspects of privacy constraints be handled better?

Section 2.2 will address Q1 and Q2, and Section 3 will explore Q3 and Q4.

## 1.3 Summary of Contributions

The main contributions of this research lies in (1) analyzing, modeling, managing and enforcing privacy constraints in web personalization, and (2) conceptualizing, prototyping and evaluating a novel privacy-enhanced personalization framework.

### 1.3.1 Contributions from Analysis of the Impacts of Privacy Laws on Personalized Systems

- Our detailed analysis of privacy laws recognizes their impacts on personalized systems and reveals similarities, differences and trends in various privacy laws.
- We found that privacy laws not only affect data collection and storage, but also data processing/reasoning involved in web personalization.
- The practical implication of the previous finding is that legal and regulatory privacy requirements may allow or prohibit the usage of certain user modeling methods in personalized systems.

### 1.3.2 Contributions from PLA-based Framework for Privacy-Enhanced Personalization

- Our approach that is based on the conception of product line architecture (PLA) modularizes privacy constraints and personalization components. The result is a flexible approach that not only helps address the complexity of building personalized systems, but also strongly supports their evolution: as new privacy and personalization concerns arise, they can be modularly added to the PLA.

- Our framework respects and addresses the traditionally neglected impacts of privacy constraints on personalization methods (data processing/reasoning).

- Our approach does not only allow one to specify privacy requirements, but it also enforces their consequences on user modeling methods.

- Because of the explicit representation of legal requirements and their consequences in personalized systems, our approach helps make internal/external audit and system compliance with privacy constraints easier.

- Our approach acknowledges that individual users may have different privacy constraints. The privacy that the system affords is personalized to cater to each user.

- To our best knowledge, this is the first time that individual users' privacy constraints are treated as first-class design requirements and become part of system design specifications for personalized systems.

- Our approach allows privacy constraints to be addressed dynamically when users change their privacy preferences or when privacy laws change, even during runtime.

### 1.3.3 Contributions from User Evaluation of the Framework

- With our privacy enhancement mechanism, users have higher regard for the privacy practices of the personalized site.

- With our privacy enhancement mechanism, users disclose more information about themselves to the personalized site.

- With our privacy enhancement mechanism, users make more book purchases on the personalized site.

- Our privacy enhancement mechanism does not unduly compromise users' perceived personalization benefits.

### 1.3.4 Contributions from Performance Evaluation of the Framework

- The light-weight PLA representation, request distribution and multi-level caching mechanisms significantly improve the system performance.
- Contemporary personalized websites, even the ones with heaviest Internet traffic, e.g., Yahoo, can adopt our approach to provide privacy-enhanced web personalization to its users with a reasonable number of extra servers.

### 1.3.5 Recap of Contributions

- Our analysis of privacy laws reveal that they affect the usage of user modeling methods used to make inferences about users in personalized systems.
- Our PLA-based framework provides a flexible, extensible, and enforceable approach in modeling and addressing each user's privacy constraints in web personalization in a responsive manner.
- Our evaluations of the framework show that users value user-tailored privacy enforcement in personalization, and that it is technically feasible even for personalized sites with heavy traffic.

# Chapter 2

# Privacy Requirements and Their Impacts on Personalized Systems

In this chapter we describe various privacy constraints in the domain of personalized systems. Note that the privacy constraints discussed here are predominately about user/consumer privacy with regard to companies. People's privacy with regard to government, and interpersonal privacy are not the focus of this research.

## 2.1 Privacy Constraints

Privacy has been recognized as a fundamental human right at least since the seminal work of Warren and Brandeis (Warren and Brandeis, 1890). In recent decades, various privacy issues have been raised and have attracted substantive attention in society, due to the proliferation and advancement of innovative information technologies such as computers, the Internet, and recently mobile and ubiquitous computing applications. Despite its importance, the concept of privacy is difficult to grasp. Privacy is a truly multi-dimensional notion. It involves, but is not limited to, cultural, social, legal, psychological, political, economic and technical aspects.

Privacy has been studied for decades, and many different definitions of privacy have been proposed. For instance, prominent 19th-centry American justice Thomas Cooley defined privacy as “the right to be left alone” (Cooley, 1888). Warren and Brandeis (Warren and Brandeis, 1890) defined privacy as “the right of determining, ordinarily, to what extent his thoughts, sentiments, and emotions shall be communicated to others”. Westin (Westin, 1967) described privacy as “the ability to determine for ourselves when, how, and to what extent information about us is communicated to others”. Irwin Altman conceptualized privacy management as a dynamic and dialect boundary regulation process that involves “a selective control of access to the self or to one’s group” (Altman, 1975).

A number of scholars extend these theoretical notions of privacy to meet the challenges in an age of information technologies. Palen and Dourish (2002) drew from Irwin Altman’s privacy regulation theory (Altman, 1975) and argued that privacy is not static but highly nuanced, situated and contingent, something that is being constantly acted out. Dourish and Anderson (2006) suggested that privacy should not be studied and/or treated in isolation but rather in relation with the ecology of human information practices. Helen Nissenbaum conceptualized privacy as “contextual integrity”, which “ties adequate protection for privacy to norms of specific contexts, demanding that information gathering and dissemination be appropriate to that context and obey the governing norms of distribution within it” (Nissenbaum, 2004). Others take a more pragmatic approach towards conceptualizing privacy. For instance, Daniel Solove (Solove, 2006) commented that “privacy is in disarray and nobody can articulate what it means” and then proposed a concrete taxonomy of privacy in terms of information collection, processing, dissemination, and invasion.

Despite the existence of various conceptualizations of privacy, there is no unanimously agreed-upon definition of privacy. This is largely due to the fact that privacy is "an overwhelmingly large and nebulous concept" (Boyle and Greenberg, 2005). Young (Young, 1978) wittedly commented that "privacy, like an elephant, is...more readily recognized than described". Nevertheless, we take the position that privacy is personal, dynamic, and situated (context-dependent).

If privacy considerations are taken into account in the design of computer systems, they constrain the possible design space for such systems. Solutions that violate privacy constraints cannot be considered any more. Privacy constraints for computer systems stem primarily from two sources, namely from privacy laws and regulations and from personal privacy expectations of the computer users. Figure 2.1 shows the hierarchy of these constraints with a focus on privacy laws and regulations (Wang and Kobsa, 2009).

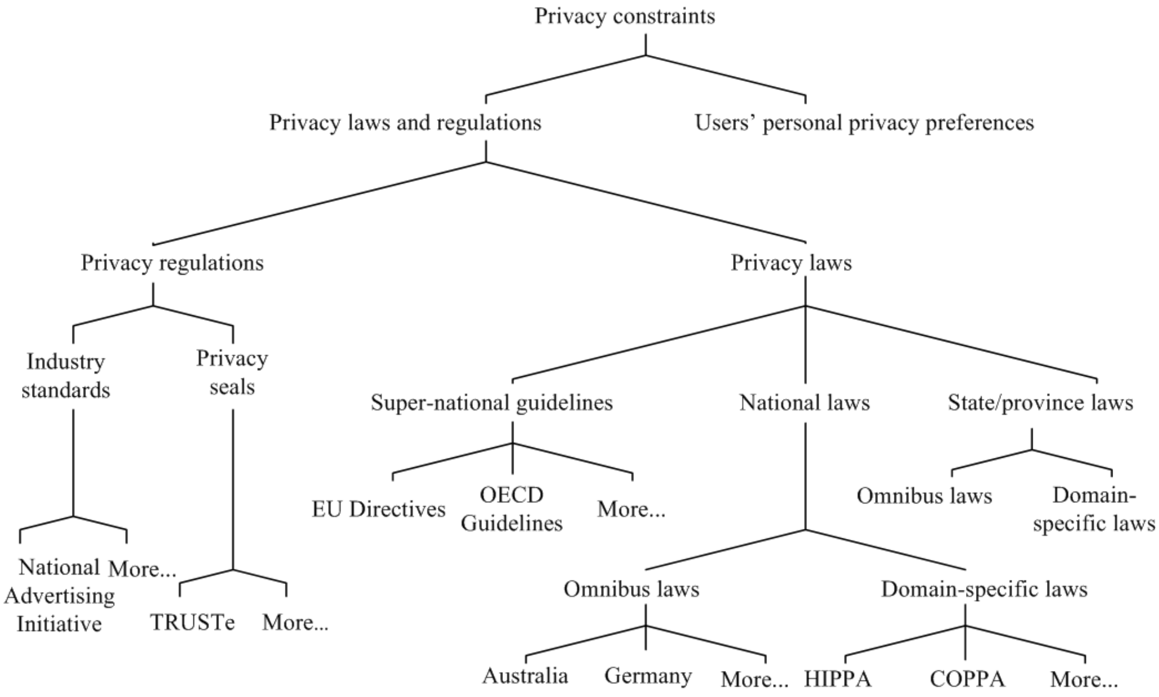


Figure 2.1: The Hierarchy of Potential Privacy Constraints

## 2.2 Impacts of Privacy Laws and Regulations

### 2.2.1 Privacy Laws

There are two schools of thoughts regarding privacy legislation (Xu, 2009). One school of thoughts view privacy as a fundamental human right, e.g., Young (1978) suggested that "the right to privacy is inherent in the right of liberty". The other school of thoughts view privacy as "a commodity", e.g., Waldo et al. suggest that privacy has instrumental value which "sustains, promotes, and protects other things that we value" (Waldo et al., 2007). In consequence, "human right" societies such as Australia, Canada, and European Union have legislated "omnibus" privacy laws that govern all instances of data practices in all sectors of their economies. In contrast, "commodity" societies did not legislate "omnibus" privacy laws. Some countries in this category, the U.S., for instance, has several sector-specific privacy laws and industry self-regulations (Smith, 2001, 2004).

We have witnessed a proliferation of privacy laws and regulations. There are currently more than 40 countries that have their own national privacy laws. Besides, various types of privacy regulations, industry seal programs, and company self-governing policies are launched like bamboo shoots after spring rain. When users are in principle identifiable (i.e., a user can be identified with a reasonable amount of effort), privacy laws and regulations often apply.

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 acquisition, notification about the purpose of use, permissible data transfer (e.g., to third parties and/or across national borders) and permissible data processing (e.g., organization, modification and destruction). Other requirements prescribe user opt-ins (e.g., asking for their consent before collecting their data), opt-out (e.g., of data collection and/or data processing) and user inquiries (e.g., regarding what personal information was collected and how it was processed and used). Others mandate the establishment of adequate security mechanisms (e.g., access control for personal data), the supervision and the audit of personal data processing.

Westin and van Gelder (2003) provide a historical account of privacy and data protection legislation, which Europe and the U.S. initiated parallel endeavors. The U.S. inaugurated the Fair Credit Reporting Act of 1970 (US, 1970) and the Privacy Act of 1974 (US, 1974) for the protection of citizen and consumer information databases. The Fair Information Practice Principles (FTC, 1973) that were first formulated by the U.S. Department of Health, Education and Welfare in 1973 became the basis for many privacy laws and regulations worldwide. A number of Western European countries, such as France, Germany and Sweden followed the trend in the late 1970s and early 1980s .

In 1980, the Organisation for Economic Co-operation and Development (OECD) drafted Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (OECD, 1980) These guidelines are however not binding for its currently 30 member countries, which include the U.S.. The European Union issued two privacy-related directives (EU, 1995, 2002) that set out the minimum standards for its member states to implement in their respective national privacy laws. The Asia-Pacific Economic Cooperation (APEC) recently also drafted a privacy framework (APEC, 2005), serving as recommendations for its currently 21 member countries including the U.S.

In the U.S., several sector-specific laws have come into effect such as the Health Insurance Portability and Accountability Act of 1996 (HIPAA) (HHS, 1996) for medical privacy, the Children's Online Privacy Protection Act of 1999 (COPPA) (FTC, 1999) for protecting children under the age of 13, the Gramm-Leach-Bliley Act of 1999 (US, 1999) for financial privacy, and the Sarbanes-Oxley Act of 2002 (SEC, 2002) for accounting and financial reporting. In 2000, the Federal Trade Commission (FTC) published a widely known report to Congress on Fair Information Practice Principles (FTC, 2000b). In the same year, the FTC also issued the so-called Safe Harbor Principles (FTC, 2000c) to meet the adequacy standard imposed by the EU. In 2006, the Association for Computing Machinery (ACM), the world's largest computer science association, announced recommendations of privacy principles drafted by its U.S. Public Policy Committee (USACM, 2006).

Our analysis (Wang et al., 2006a) of over 40 privacy laws found that if such laws apply to a personalized website, they often not only affect the data that is collected by the website and the way in which data is transferred, but also the personalization methods that may be used for processing them. The following are some example codes (Wang and Kobsa, 2006):

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 would receive. The common practice is to collect as much data about the user as possible, to lay them in stock, and then to apply those personalization methods that "fire" based on the existing 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 when asked, i.e. even during the current service. A case can probably be made that

[1] The traffic data is of communication networks (such as cell phone network and the Internet).

users should not only be able to make all-or-none decisions, but also decisions on 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 user information from, and supply the data to, different personalized applications. A UMS must not supply data to personalized applications if they intend to use those data for different purposes 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, 2006). *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 assign scores to users that significantly affect them.

We found that the privacy laws that impact personalized systems 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 Act. The reason is that these laws are particularly geared towards electronic communications while other privacy laws and regulations have a much broader scope. More countries are currently drafting such specific privacy laws to regulate telecommunication, teleservices, e-commerce, and even the usage of RFID tags.

### 2.2.2 Company and Industry Regulations

Many companies have internal guidelines 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 increasingly the TRUSTe privacy seal program).

The TRUSTe is a good and interesting representative in this space. A website that seeks seal certification is required to provide a detailed self-assessment of its privacy policy and practices and an certification application. The seal program issues a seal certificate based on the results of independent auditing on the website's compliance with regard to its claimed privacy policy. Once the certificate is granted, the seal program will keep monitoring the website's privacy practices and will handle complaints from customers by probing the website's practices and requesting the website to enact actions for resolution. If the website fail to resolve the issues, the associated certificate will be revoked. There are cases where TRUSTe revoked some sites' seals. However, since these seal programs do not impose any sort of minimum requirements on websites' privacy practices, websites can establish their own policies which may differ significantly in their qualities. Therefore, the bottom line is that a privacy seal does not mean good privacy practices and quite ironically studies such as (LaRose and Rifon, 2006) have found that on average websites that have seals are more privacy-invasive. Perhaps, the reason is that those privacy-invasive websites are aware of their questionable practices and then attempt to get privacy seals to disguise their privacy-invasive nature.

## 2.3 Impacts of Users' Online Privacy Concerns

Numerous opinion polls and empirical studies have revealed that Internet users have 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 (with a few conducted in 2008 and 2009), mostly in the United States. The following is a summary of a number of important findings (the percentage figures indicate the ratio of respondents from multiple studies who endorsed the respective view). See (Teltzrow and Kobsa, 2004; Kobsa, 2007b) for more details.

### 2.3.1 Personal Data.

1. Internet users who are concerned about the privacy or security of their personal information online: 70% - 89.5%;

2. People who have refused to give personal information to a web site at one time or another: 82% - 95%;

3. A 2008 report (Pew, 2008) by The Pew Internet & American Life Project noted that "59% of adults have refused to provide information to a business or company because they thought it was not really necessary or was too personal."

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 shares their data for a different than the original purpose: 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 (UMSs) (Kobsa, 2001) that collect data from, and share them with, different user-adaptive applications.

### 2.3.2 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%.

Besides traditional means of tracking users online such as cookies, new Internet tracking technologies are being developed and deployed. However, these newer tracking techniques such as “flash cookies” are much less known to ordinary users. “Flash cookies” basically provide the standard web cookies’ functionalities on the Macromedia Flash platform. This technology makes websites easier to track web usage on the Flash platform (which Macromedia claimed that the predominate majority of Internet users have installed). However, “flash cookies” cannot be managed through the standard web browser settings (BetterPrivacy, an optional Firefox add-on can achieve this though). Instead, one has to either go to

Macromedia's website to manage them or locate the "shared object files" on the local hard drive. Again, the biggest problem is that most users probably are not aware of the existence of the "flash cookie" practices.

According to a 2009 study on tailored advertising (Turow et al., 2009), if given a choice, 68% of Americans "definitely would not" and 19% "probably would not" allow advertisers to track them online even if their online activities would remain anonymous. 63% of Americans feel that laws should require advertisers to delete information about their Internet activity immediately. 69% of Americans would like to see a law giving them the right to access all of the information a Web site has collected about them. 62% of Americans erroneously believe that "if a website has a privacy policy, it means that the site cannot share information about you with other companies, unless you give the website your permission". 86% of young adults reject advertisements that are tailored based on their activities across multiple Web sites. 90% of young adults 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 2007 study (May Lwin, 2007) shows that strong business policy is effective in reducing the concerns of collecting low sensitive data from users, but ineffective for highly sensitive data, 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 really explain why highly sensitive data is collected for their concrete business contexts.

### 2.3.3 Other factors

Kobsa (Kobsa, 2007b) suggested that developers of personalized system should not feel discouraged by the stated privacy concerns and their potential negative impact on personalized systems. Rather, they should incorporate a number of mitigating factors into their designs that have been shown to encourage users' disclosure of personal data. Such factors include perceived value of personalization, previous positive experience with the site, the presence of a privacy seal, catering to individuals' privacy concern, etc.

## 2.4 Summary

Privacy and web personalization is in conflict. Users' privacy concerns/preferences tend to offset personalization benefits. Privacy laws and regulations exacerbate the conflict by setting requirements that further affect the ways in which web-based personalized systems operate. More specifically, when these legislation and regulations are in effect, they often not only affect the data that are collected by the website, the way in which the data is transferred and to which party it is transferred, but also the methods that may be used for processing them. By highlighting these significant impacts, we advocate more recognition of the importance of privacy in designing and implementing web-based personalized systems.

# Chapter 3

# Related Work

Since privacy and personalization is at odds, various solutions have been proposed to make tradeoffs between privacy and personalization. In this chapter, we first review generic, domain-independent privacy-enhancing technologies which can be adopted for the domain of web personalization. We then review technical privacy-enhanced solutions that are specifically designed for personalization.

## 3.1 An Analytical Framework for Evaluating Privacy-Enhancing Technologies

In order to systematically evaluate the effectiveness of these technologies, we propose an evaluation framework that draws on three analytical aspects regarding the solution being examined:

1. *What high-level principles the privacy solution follows*

   We identify a set of fundamental privacy principles from various privacy laws and regulations (e.g., notice and awareness) and treat them as high-level guidelines for enhancing privacy. Personalization quality principle aims to assess the quality of the personalized services supported by a particular solution. Other principles that are desirable for privacy enhancement (e.g., usability) are also recognized.

2. *What privacy concerns the privacy solution addresses*

   While privacy principles are high-level guidelines to enhance privacy, privacy concerns are more concrete and mundane. Ideally one would need user studies to examine how effectively solutions address users' changing and contingent privacy needs and preferences. Since running such studies for every evaluated solution is barely realistic, we instead chose to investigate privacy concerns that are somewhere in between high-level privacy principles and low-level contingent privacy needs of users.

   Furthermore, in order to better assess the privacy protections that privacy-enabling technologies afford, we propose to group them into the following three categories:

   - **Protection of identity**: this type of privacy protection aims to prevent users' true identities from being revealed (i.e., who they are).
   - **Seclusion**: this type of privacy protection attempts to prevent users from being bothered by unwanted contact or solicitation (e.g., spam emails).
   - **Control over data**: this type of privacy protection allows users to have control over their data, e.g. regarding what data can be collected or disclosed for what purpose, how the data will be used, and with whom the data may be shared or to whom it may be transferred.

3. *What basic privacy-enhancing techniques the solution employs*

   We look at the technical characteristics of privacy solutions, to critically analyze their effectiveness in safeguarding privacy and supporting personalization. The discussion of this evaluative aspect will be postpone to Section 3.3 where we review privacy-enhancing personalization solutions.

### 3.1.1 Principles

Privacy legislation and regulations are usually instantiations of more fundamental privacy principles. We select a core set of privacy principles that are frequently addressed in privacy laws and regulations, and add other principles/properties that are also desirable for privacy enhancement. This list of principles is by no means exhaustive, but meant to initiate a discussion on what principles are desirable for enhancing privacy effectively. The principles are grouped by their origin in the listing below.

**Privacy principles from privacy laws and regulations**

1. Notice/Awareness
   - Privacy policy: *Make [...] privacy policy statements clear, concise, and conspicuous to those responsible for deciding whether and how to provide the data* (USACM, 2006);
   - Notice upon collection: *Whenever any personal information is collected, explicitly state:*
     - *the precise purpose for the collection,*
     - *all the ways in which the information might be used,*
     - *all the potential recipients of the personal data,*
     - *how long the data will be stored and used* (USACM, 2006);
2. Data minimization

   *Before deployment of new activities and technologies that might impact personal privacy, carefully evaluate them for their necessity, effectiveness, and proportionality: the least privacy-invasive alternatives should always be sought* (USACM, 2006).
3. Purpose specification

   *The purposes for which personal data are collected should be specified not later* *than at the time of data collection and the subsequent use limited to the fulfillment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose* (OECD, 1980).

4. Collection limitation

   *There should be limits to the collection of personal data and any such data should be obtained by lawful and fair means [...]* (OECD, 1980).

5. Use limitation

   *Personal data should not be disclosed, made available or otherwise used for purposes other than those specified* (OECD, 1980).

6. Onward transfer

   *Personal data should not be transferred to a third country/party if it does not ensure an adequate level of protection* (EU, 1995; FTC, 2000c).

7. Choice/Consent

   *Where appropriate, individuals should be provided with clear, prominent, easily understandable, accessible and affordable mechanisms to exercise choice in relation to the collection, use and disclosure of their personal information* (APEC, 2005). The two widely adopted mechanisms are (FTC, 2000a):

   - Opt-in: *requires affirmative steps by the consumer to allow the collection and/or use of information*;

   - Opt-out: *requires affirmative steps to prevent the collection and/or use of such information.*

8. Access/Participation

   An individual should have right to:

   - *know whether a data controller has data relating to her* (OECD, 1980),

- *inspect and make corrections to her stored data* (USACM, 2006).

9. Integrity/accuracy

   *A data controller should ensure the collected personal data is sufficiently accurate and up-to-date for the intended purposes and all corrections are propagated in a timely manner to all parties that have received or supplied the inaccurate data* (USACM, 2006).

10. Security

    *Personal data should be protected by reasonable security safeguards against such risks as loss or unauthorized access, destruction, use, modification or disclosure of data* (OECD, 1980).

11. Enforcement/Redress

    Effective privacy protection must include mechanisms for enforcing the core privacy principles. At the minimum, the mechanisms must include (FTC, 2000c):

    - Recourse mechanisms for customers: *readily available and affordable independent recourse mechanisms by which an individual's complaints and disputes can be investigated and resolved and damages awarded where the applicable law or private sector initiatives so provide*;

    - Verification mechanisms for data controllers: *follow-up procedures for verifying that the attestations and assertions businesses make about their privacy practices are true and that privacy practices have been implemented as presented*;

    - Remedy mechanisms: *obligations to remedy problems arising out of failure to comply with these principles by organizations announcing their adherence to them and consequences for such organizations.*

**Anonymity-related principles**

12. Anonymity

    Anonymity means that users cannot be identified nor be tracked online.

13. Pseudonymity

    Pseudonymity also means that users cannot be identified, but they can still be tracked using a so-called alias or persona. The German Telemedia Law (DE-TML, 2007) mandates that profiling necessitates the use of pseudonyms or the prior consent of the user.

14. Unobservability

    A data controller cannot recognize that a system/website is being used or visited by a given user.

15. Unlinkability

    A data controller cannot link two interaction steps of the same user.

16. Deniability

    Deniability means that users are able to deny some of their characteristics or actions (e.g., having visited a particular website), and that others cannot verify the veracity of this claim.

**Other desirable principles for privacy enhancement**

17. User preference

    Different users can potentially have different privacy preferences. A data controller should tailor its privacy practices to each individual user's preferences.

18. Negotiation

    This principle calls for the support of negotiation between a user and a website, during which they can reach an agreement on privacy practices that the website may employ for the respective user.

19. Seclusion

    Seclusion means that users have the right to be left alone. Violations of this principle in the electronic world are popup ads and junk emails.

20. Ease of adoption

    Oftentimes privacy protection mechanisms rely on the presence of other infrastructures or technologies, and this fact may pose significant barriers for adoption. This principle relates to the readiness of *organizations* to adopt the examined privacy protection (e.g., whether the solution relies on special protocols or technologies that are proprietary or not readily available).

21. Ease of compliance

    This principle is concerned with the ease of fulfilling legal requirements by adopting a specific privacy protection solution.

22. Usability

    The privacy protection solution should be easy for *users* to adopt (e.g., the efforts required from users to utilize the solution should be reasonable).

23. Responsiveness

    The privacy protection solution should respond promptly to changes in users' privacy decisions.

| **Specification**<br>**Principle** | OECD Guidelines (OECD, 1980) | EU Directive on Data Protection (EU, 1995) | German Telemedia Law (DE-TML, 2007) | APEC Privacy Framework (APEC-FIP, 2004) | FTC Safe Harbor Principles (FTC, 2000c) | FTC Fair Info Practice (FTC, 2000b) | ACM Principles (USACM, 2006) |
|---|---|---|---|---|---|---|---|
| Notice/Awareness | X | X | X | X | X | X | X |
| Minimization | | | | | | | X |
| Purpose specification | X | X | X | X | X | | X |
| Collection limitation | | X | X | X | X | | |
| Use limitation | X | X | X | X | X | | X |
| Onward transfer | | X | X | | X | | |
| Choice/Consent | X | X | X | X | X | X | X |
| Access/Participation | X | X | X | X | X | X | X |
| Integrity/accuracy | X | X | X | X | X | X | X |
| Security | X | X | X | X | X | X | X |
| Enforcement/Redress | | X | X | | X | X | |
| Anonymity-related principles | | | | | | | |
| Anonymity | | | | | | | |
| Pseudonymity | | X | X | | | | |
| Unobservability | | | | | | | |
| Unlinkability | | | | | | | |
| Deniability | | | | | | | |
| Other desirable principles for privacy enhancement | | | | | | | |
| User preference | | | | | | | |
| Negotiation | | | | | | | |
| Seclusion | | | | | | | |
| Ease of adoption | | | | | | | |
| Ease of compliance | | | | | | | |
| Usability | | | | | | | |
| Responsiveness | | | | | | | |

Figure 3.1: Privacy guidelines/frameworks and privacy principles.

Privacy laws and regulations typically only include subsets of the above principles. For a comparison, Figure 3.1 shows a group of representative privacy laws and regulations in its columns, and the privacy principles discussed above in its rows. An "X" in a cell means that the framework includes the respective principle.

### Applying our privacy protection taxonomy to the principles

We now categorize the 23 identified principles based on the type of privacy protection they relate to. Note that the general category contains principles that pertain to all three types of privacy protection we introduced in Section 3.1. Figure 3.2 represents which category each privacy principle falls into.

| **Protection**<br>**Principle** | General | Protection of Identity | Seclusion | Control over data |
|---|---|---|---|---|
| Notice/Openness | X | | | |
| Minimization | | | | X |
| Purpose specification | | | | X |
| Collection limitation | | | | X |
| Use limitation | | | | X |
| Onward transfer | | | | X |
| Choice/Consent | X | | | |
| Access/Participation | | | | X |
| Integrity/accuracy | | | | X |
| Security | | | | X |
| Anonymity | | X | | |
| Pseudonymity | | X | | |
| Unobservability | | X | | |
| Unlinkability | | X | | |
| Deniability | | X | | |
| Enforcement/Redress | X | | | |
| User preference | X | | | |
| Negotiation | X | | | |
| Seclusion | | | X | |
| Ease of adoption | X | | | |
| Ease of compliance | X | | | |
| Usability | X | | | |
| Responsiveness | X | | | |

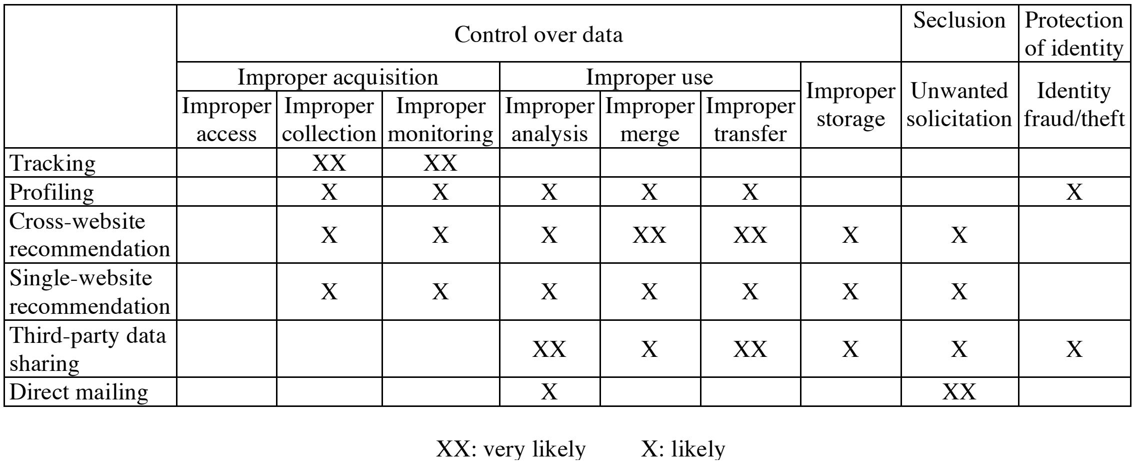
Figure 3.2: Categorization of principles based on the type of privacy protections

### 3.1.2 Privacy concerns

Whereas privacy principles are high-level guidelines for enhancing privacy, users' privacy concerns are more concrete and down to the earth. We discuss and analyze them here in order to also be able to evaluate the effectiveness of privacy enhancements from a subjective stance. Privacy concerns usually arise from characteristics of a specific application domain.

To illustrate this, we will focus on the potential privacy concerns that may arise in web personalization (Kobsa, 2007b), such as Amazon's personalized book recommendations.

Wang et al. (1998) presented a taxonomy of privacy concerns in Internet marketing that includes improper access, improper collection, improper monitoring, improper analysis, improper transfer, unwanted solicitation and improper storage. These high-level concerns as well as concerns about improper merging of data also apply in web personalization. Figure 3.3 shows what privacy concerns (columns) can arise from typical web personalization activities (rows).



| | Control over data | | | | | | | Seclusion | Protection of identity |
|---|---|---|---|---|---|---|---|---|---|
| | Improper acquisition | | | Improper use | | | Improper storage | Unwanted solicitation | Identity fraud/theft |
| | Improper access | Improper collection | Improper monitoring | Improper analysis | Improper merge | Improper transfer | | | |
| Tracking | | XX | XX | | | | | | |
| Profiling | | X | X | X | X | X | | | X |
| Cross-website recommendation | | X | X | X | XX | XX | X | X | |
| Single-website recommendation | | X | X | X | X | X | X | X | |
| Third-party data sharing | | | | XX | X | XX | X | X | X |
| Direct mailing | | | | X | | | | XX | |

XX: very likely X: likely

Figure 3.3: Potential privacy concerns in typical web personalization activities

## 3.2 Domain-Independent Privacy-Enhancing Technologies

In this section, we will review major privacy-enhancing technologies (PETs) that are not designed for any particular domain of applications. More specifically, we will review privacy policy languages, anonymity techniques, authentication and identity management, authorization and access control, usable security and privacy mechanisms. They provide technological building blocks for privacy-enhanced personalization solutions. We will evaluates these domain-independent PETs against the two analytical aspects of our framework (introduced in Section 3.1), namely what principles they follow, and what privacy concerns they address. This close examination of existing PETs will allow for a more comprehensive view of their pros and cons as well as their current gaps, and thus point out future research avenues.

### 3.2.1 Privacy policy languages

The U.S. Federal Trade Commission (FTC) defines a privacy policy as a comprehensive description of a company's information practices, accessible by clicking at a hyperlink on the company's website (FTC, 1998). Its aim is to enhance users' awareness of the privacy practices of the website. Privacy policies thus are directed at human readers. Privacy policy languages, in contrast, are intended to be machine-readable. They can be roughly divided into two types: external policy languages to describe websites' public privacy policies or users' privacy preferences, and internal ones to specify companies' or websites' internal rules for privacy practices. In general, external privacy policy languages are declarative without enforcement mechanism, while internal privacy policy languages are normative with support for enforcement.

#### External privacy policy language

*P3P: The Platform for Privacy Preferences*

Developed by the World Wide Web Consortium (W3C), the Platform for Privacy Preference (P3P 1.1) (Cranor et al., 2006) aims at increasing the transparency of websites' privacy practices in such a way that users can easily decide whether or not these websites meet their privacy expectations. Technically, P3P consists of two parts: (1) a standard machine-readable (XML) language/syntax that allows websites to describe their privacy practices regarding the collection, use, and distribution of personal information, and (2) a "handshake" protocol built on top of the HTTP protocol that enables P3P-enabled user agents (e.g., web browsers) to retrieve websites' P3P privacy policies automatically (Garfinkel and Cranor, 2002). Agents can also be configured to inform users about the sites' privacy policies, to notify them when those change, to warn them when those deviate from their pre-specified privacy preferences (expressed in languages like APPEL (Cranor et al., 2002) or XPref (Agrawal et al., 2003)), and to semi-automate or automate the decision whether or not to disclose the requested information on users' behalf.

A P3P policy file can be applied to a whole website or certain parts of it such as web pages, images, cookies, forms and even a single form field[1]. Every P3P policy contains a description of the legal entity responsible for the privacy policy, whether the site allows users to have access to the information collected about them, (optional) information regarding dispute resolution and remedy, and at least one statement. Each statement describes the data being collected (physical contact information, online contact information, purchase information, click stream data, etc.), the purpose(s) for collection (web site administration, research and development, profiling, etc.), whether the site supports user opt-in or opt-out for those purposes, what organizations will have access to the collected data (primary service provider only, delivery services, unrelated third parties, etc.), the retention of the collected data (single session, stated purpose, indefinitely, etc). Personalization can be considered as one type of purposes dubbed as "individual decision"[2] and similarly anonymous personalization as "pseudo decision".

P3P was designed as part of a broader privacy protection framework (including privacy legislation and enforcement) and is applicable to any web-based systems. P3P implementations include:

[1] The P3P 1.1 specification provides a new mechanism that binds a P3P policy to an XML element that does not have to be associated with a URI.

[2] Information may be used to determine the habits, interests, or other characteristics of individuals and combine it with identified data to make a decision that directly affects that individual.

- P3P user agents (such as Internet Explorer 6 (Microsoft, 2000) that supports cookie management as well as websites' privacy policies disclosure,
- AT&T Privacy Bird (Cranor and Reidenberg, 2002), an add-on to the Internet Explorer, that utilizes differently colored bird icons in the corner of the browser window to indicate whether or not a site's P3P policy matches the user's preferences,
- Privacy Bird Search Engine (Byers et al., 2004) that annotates regular search results with an indication to what extent the P3P policy of each site matches the user's requirements,
- P3P policy generators/editors/checkers (e.g., P3PEdit (P3PEdit, 2001)), and
- server-side P3P support (e.g., IBM Tivoli Privacy Manager For E-business (IBM, 2003a) that can enforce privacy policies internally in a system).

P3P's official website currently lists about 2900 websites worldwide that have adopted P3P 1.0[3]. The latest P3P adoption study conducted in the summer of 2005 (Egelman et al., 2006) estimates the overall P3P adoption rate at about 10% using a list of "typical" search terms taken from AOL users' queries, and the government adoption rate roughly at 36% (this is by far the largest sector to adopt P3P, which is probably due to the P3P adoption mandate of the E-Government Act[4] (USA, 2002)). The usage of P3P has also been proposed in the context of ubiquitous computing (Langheinrich, 2002).

Despite its relative popularity, P3P has a number of limitations. First, P3P does not include any technical mechanism for enforcing privacy policies. It is totally up to the websites to follow their stated privacy policies, and users cannot verify whether a site acts as promised.

Second, P3P (even the latest version 1.1) does not support different policies for different users, albeit offering users a choice of P3P policies is mentioned in P3P's future plan.

[3] http://www.w3.org/P3P/compliant_sites.php3

[4] The act mandates that government agencies post machine-readable privacy policies on their web sites.

Nevertheless, several proposals for individual negotiation of P3P policies have been made (Buffett et al., 2004; Preibusch, 2006).

Third, P3P might not be expressive enough to be able to fully encode the nuances of websites' privacy practices. For example, P3P cannot handle cases where privacy concerns crosscut more than one statement (e.g., that personal data that were obtained for different purposes may not be grouped (CZ, 2000)). Because of this lack in expressiveness and enforcement, it is difficult for websites to keep their P3P privacy policies, human-readable privacy policies and actual practices all consistent. P3P allows sites to further explain those nuances in the human-readable fields, which are not technically analyzable though.

Furthermore, P3P is not able to accurately capture the subtleties of privacy laws and regulations, nor does it develop a minimum set of privacy or security standards that web sites should follow. Therefore, websites cannot rely on P3P as a technical means to comply with relevant privacy provisions, and discrepancies might exist between the P3P policies and the applicable privacy laws which in turn can expose the websites to legal jeopardy (Cranor and Reidenberg, 2002).

Last but not least, P3P has been criticized for facilitating websites' data collection, rather than protecting users' privacy (Coyle, 1999). A convincing example is that the default value for data retention is "indefinitely" instead of "no retention" or "stated purpose". Besides, P3P makes it difficult for users to protect their privacy (for instance, users find changing defaults for cookie settings to be burdensome and confusing (EPIC and Junkbusters, 2000). P3P could also effectively excludes non-compliant sites (e.g., websites without P3P compact policies would be blocked by IE6) and P3P-compliant sites do not make themselves more trustable (a study has shown that among the top 500 companies, only nine out of the 65 sites that adopted P3P have P3P satisfactory to a pragmatic user who wants some privacy protection (Ashrafi and Kuilboer, 2005)).

In order to gauge P3P's expressiveness, we attempted to describe six representative privacy provisions that have the biggest impact on the internal operation of web personalization (Wang and Kobsa, 2006). The results show that P3P can express most of the provisions, with the following deficiencies: First, it needs more fine-grained expressiveness (e.g., retention time cannot be set in a continuous time scale; this can be easily solved by introducing an "expiry"-like sub-element for the retention element). Secondly, perhaps the biggest issue with the P3P language, it cannot express interactions across different statements; a potential solution is to introduce logic operators in the statement-group element so that different relationships between statements can be captured. Thirdly, the overall P3P framework is short of an interface with systems that enforce P3P privacy policies.

*APPEL: A P3P Preference Exchange Language*

APPEL (Cranor et al., 2002) was designed to complement P3P by allowing users to express their privacy preferences in terms of rules that specify certain conditions under which user information may be collected and used, so that P3P-enabled agents are able to check users' preferences against a website's P3P policy to make automated or semi-automated decisions whether or not users' data may be released to the website. A rule includes a behavior, an optional persona, optional explanation and prompt messages, and a number of expressions (Cranor et al., 2002). An expression is used to match a full XML element or a single attribute and its value in an XML element in the evaluated P3P policy.

APPEL only allows logical operations at nodes corresponding to P3P elements. The matching scheme of APPEL is problematic: a P3P policy can contain multiple statements, and a rule will fire and then stop being further evaluated against the policy if any of the statements satisfies the rule. Therefore, the remaining statements will be ignored. Because of this deficiency, APPEL only works correctly when rules express what is unacceptable rather than what is acceptable (Agrawal et al., 2003).

*XPref*

XPref (Agrawal et al., 2003) is another preference language for P3P and is based on the XPath language (Clark and DeRose, 1999), a W3C Recommendation for navigating and matching the hierarchical structure of an XML document. The biggest difference between XPref and APPEL is that APPEL uses the sub-elements of a rule to specify acceptable and unacceptable combinations of P3P elements, while XPref utilizes XPath expressions for the same purpose. XPref outweighs APPEL in that it can specify what is acceptable as well as what is unacceptable, and combinations of both. Agrawal et al (Agrawal et al., 2003) show that XPref subsumes APPEL, and that APPEL can be programmatically translated into XPref.

*Individual privacy policy negotiation*

Buffett et al. (2004) present a framework for the negotiation between users and websites about the disclosure of user information for compensation. It relies on utility theory to allow users to express the value associated with each piece and combination of personal data. The proposed PrivacyPact protocol enables the transmission of messages for negotiation. Preibusch (2006) identified relevant privacy dimensions (recipient, purpose, retention, and data) for negotiation and proposed a simple extension of P3P to allows for the expression and implementation of such negotiation processes.

### Internal privacy policy language

*EPAL: Enterprise Privacy Authorization Language*

EPAL (Ashley et al., 2003) is a formal language developed by IBM that allows enterprises to write their internal privacy policies, so that those can be enforced across IT applications and systems in an automated manner. To take advantage of EPAL, an enterprise first defines an EPAL vocabulary (i.e., concrete element types) that caters to its own needs, and then specifies its EPAL privacy policies. Applications that aim to share an EPAL policy obviously must agree on the vocabulary and interpret it in the same way.

An EPAL policy contains a set of privacy authorization rules that allow or deny requests. Rules have the following constituents (Ashley et al., 2003):

- the action for which authorization is requested (e.g., disclose or read)
- the data categories upon which these actions are going to be performed (e.g. medical records or contact info),
- the user categories that are affected (e.g., a department or a particular employee),
- the usage purposes (e.g., direct marketing or auditing),
- associated conditions (e.g., the purpose category must be billing purpose), and
- associated obligations (e.g., delete data after 7 days or obtain consent).

An authorization request contains a user category, an action, a data category, and a purpose. An authorization result is a statement that includes a ruling (allow or deny), a user category, an action, a data category, and a purpose. A rule may also contain conditions and obligations. Authorization results are used to determine if a request is allowed or denied. At first sight, EPAL policies seem quite similar to P3P policies since both describe privacy practices. However, they differ in the following ways:

- P3P is data-centric (i.e., a P3P statement covers different aspects of a specific type of data), while EPAL is access-centric (i.e., a EPAL rule refers to an instance of information access) (Stufflebeam et al., 2004);
- EPAL aims to describe the internal privacy practices of an enterprise, which are probably not to be shared with the public, while P3P is used to describe public privacy policies;

- unlike P3P that predefines the elements of a privacy policy, EPAL elements (such as actions or user categories) are abstract types which are mapped to actual instances of elements during the implementation;

- the design of EPAL takes privacy legislation and regulations into account, by including an obligation element;

- EPAL policies are machine enforceable: they are akin to access control policies in the security domain. An authorization engine parses the EPAL policies to generate a ruling given a request, and subsequently an enforcing environment/software will execute the ruling;

- Conflicting EPAL rules are allowed and solved by prioritizing rules, to allow for general rules and exceptions;

- P3P policies are always formulated in a positive manner (i.e., what is acceptable, not what is unacceptable), whereas EPAL can express both via the ruling element (allow or deny).

Although EPAL was designed to be expressive and flexible so as to capture evolving privacy legislation and customized privacy policy, we observe that it fails to express some privacy provisions, for example, personal data that were obtained for different purposes may not be grouped (CZ, 2000). Essentially, as with P3P, EPAL cannot express the interactions between different rules.

EPAL's current abstraction of actions is not sufficient for privacy authorization purposes. Actions need to be modeled with finer granularity (e.g., modeled as hierarchyType like data categories). For example, in the case of the action "personalization/inference", we believe that it should be further categorized into different specific personalization techniques (such as collaborative filtering and incremental machine learning). Our justification is based on the observation that if privacy laws apply to a personalized website, they may affect the methods (i.e., actions) that may be used for processing them (Kobsa, 2002). Or in other words, different specific actions may lead to opposite privacy authorization results. For instance, one-time machine learning methods that rely on a record of raw data from several user sessions are not permitted without the user's consent under the German Telemedia Law (DE-TML, 2007). In contrast, incremental machine learning methods that discard the raw data of each user session and only retain the learning results may be employed. However, in the current language both types of machine learning methods/actions are modeled as an abstract "inference" action and thus the critical distinction in authorization results is lost.

### *XACML: eXtensible Access Control Markup Language*

XACML[5] (OASIS, 2005) is a general-purpose access control policy language. XACML can be used to describe not only general access control policies/rules, but also access control decision requests and responses. The root element of an XACML policy document is a Policy or PolicySet element (a container for a set of policies). A Policy contains a set of access control Rules. A Rule includes a Condition (that can be nested) and the Rule's effect (Permit or Deny). If the Condition evaluates to true, the Rule's effect (Permit or Deny) is returned. If the Condition evaluates to false, it means the Rule is not applicable (NA is returned). If an error occurs during the evaluation, "Indeterminate" is returned.

When a user/subject wants to perform some action on a resource, she will make a request to the so-called "Policy Enforcement Point" (PEP) that protects the requested resource (e.g., a file system or web server). The PEP will generate a request that includes Subject, Resource, Action and optionally Environment attributes (values). The PEP will then submit the request to a Policy Decision Point (PDP), which will identify all policies that apply to the request by evaluating their Targets. A Target is a set of conditions associated with a PolicySet or Policy or Rule. When the Target evaluates to true, the corresponding Poli-

[5] At the time of writing, XACML 3.0 is still work in progress.

cySet/Policy/Rule applies to the request. The PDP will then evaluate the request against the applicable policies, yielding a response that consists of an access control decision regarding whether or not the request should be allowed, and optionally a list of obligations (i.e., actions that the PEP is obligated to perform before granting or denying access). Since each Rule and Policy may evaluate to yield different access control decisions, the XACML utilizes a collection of Combining Algorithms (e.g., Deny Overrides Algorithms) to derive a single final access control decision. Finally, the PDP returns the final decision back to the PEP, which can then enforce the decision (namely either allow or deny the access).

XACML also comes with an approved OASIS Standard profile for privacy policies, for modeling how personally identifiable information is collected and used. An attribute of "resource:purpose" defines the purpose for which the data resource was collected. Another attribute of "action:purpose" indicates the purpose for which access to the data resource is requested. A Rule element mandates that access shall be denied unless the purpose for which access is requested matches, by regular-expression match, the purpose for which the data resource was collected.

A thorough comparison of EPAL 1.2 and XACML 2.0 shows that XACML is a functional superset of EPAL and outweighs EPAL in expressing not only access control policies but also privacy policies. Specifically, XACML provides the following important features that EPAL lacks (Anderson, 2005):

- "the ability to combine results of multiple policies developed by different policy issuers;
- the ability to reference other policies in a given policy;
- the ability to specify conditions on multiple subjects that may be involved in making a request;

- the ability to return separate results for each node when access to a hierarchical resource is requested;

- support for subjects who must simultaneously be in multiple independent hierarchical roles or groups;

- policy-directed handling of error conditions and missing attributes;

- support for attribute values that are instances of XML schema elements; and

- support for additional primitive data types (including X.500 Distinguished Names, RFC822 names, and IP addresses)."

### Summary of privacy policy languages

Figure 3.4 shows a summary of various privacy policy languages and their characteristics. Two observations can be made in this table. First, negotiation and enforcement still call for wider support. Second, XACML seems to surpass all other existing privacy policy languages, although it lacks support for negotiation.

| | External privacy policy | Internal privacy policy | User preference | Expressiveness | Negotiation | Enforcement |
|---|---|---|---|---|---|---|
| P3P | + | | | + | | |
| APPEL | | | + | + | | |
| XPref | | | + | ++ | | |
| PrivacyPact | + | | + | | + | |
| EPAL | | + | | ++ | | + |
| XACML | + | + | + | +++ | | ++ |

+++: Very strong support ++: Strong support +: Support

Figure 3.4: Privacy policy languages and their characteristics

### 3.2.2 An integrated privacy management system based on privacy policy languages

The IBM Tivoli Privacy Manager is a comprehensive enterprise privacy management system that aims at supporting a variety of privacy enhancement functionalities (IBM, 2003a):

- “centralized authorship and management of an enterprise’s privacy rules,
- a natural language interface to author and manage privacy policies,
- translation of privacy policy from prose to P3P,
- enforcement of privacy policies across the enterprise’s IT infrastructure,
- monitoring access to personal information and generating detailed audit logs,
- notification and consent preferences for information sharing across the enterprise, and
- automatically generation of reports detailing compliance to corporate policies.”

This solution focuses on privacy protection in the sense of “control over data”. It marginally addresses “seclusion and barely touches the protection of identity. In other words, the Tivoli Privacy Manager cannot protect end users’ identities, which is understandable since it is geared towards enterprise privacy management.

### 3.2.3 Anonymity techniques

Anonymity of a user means that she cannot be identified nor tracked online. One way to improve anonymity is what Goldberg et al. Goldberg (1997) called “strip identifying headers and resend” approach. This approach has been used in anonymous email remailers

(Gülcü and Tsudik, 1996) and anonymous web browsing tools like Anonymizer (ConneXion, 1996), a web proxy that strips off identifying headers and source addresses from the web browser.

Another approach is “onion routing” which is built upon the notion of “mix network” (Chaum, 1982). A mix network is essentially a chain of proxy servers (called mixes). In onion routing, a message or packet is encrypted to each mix node using public key cryptography. The resulting encryption is like a layered “onion” with the original message in the innermost layer. As the message traverse over the network, each mix node strips off its own layer of encryption to reveal where to send the message next. Untraceability can be achieved unless all mix nodes are compromised. For example, Tor (Tor, 2004), a concrete onion routing system can provide anonymous communication such as web browsing, remote login sessions, instant messaging and other applications that rely on the TCP protocol.

The third major approach is centered on the concept of “k-anonymity” (Sweeney, 2002). It is concerned with a practical problem of releasing data about individuals without revealing identifying information about them. In a k-anonymized release, each individual’s record is indistinguishable from at least k-1 others’ records. A myriad of policies and techniques (e.g., clustering (Aggarwal et al., 2006)) have been proposed to achieve k-anonymity.

### 3.2.4 Authentication and identity management

Authentication seeks to ensure that a user is actually the person who she claims to be. This is usually achieved by employing a username in combination with a password, where the username is considered as a digital identity of the bearer and the password as her authentication. A more sophisticated and thus more secure scheme is the so-called two-factor authentication, which involves two independent ways for verifying identity. It may include a user having something (e.g., a bank ATM card or a time-dependent token card) and the user knowing something (e.g., a PIN).

One of the goals of the emerging identity management systems is to allow users to have more than one digital identity and be able to freely choose which identity to use. For example, Google allows its users to use different identities/accounts in its various applications (so that, for instance, one's interactions with Google Calendar will not be combined and used in Google's personalized search).

Another recent industry example is Microsoft's CardSpace (Chappell, 2006), an "identity metasystem" that allows users to create multiple virtual ID cards. Each virtual card created by the user would only contain the minimum amount of information (retrieved from an identity provider) that individuals will need to divulge to carry out the transaction to which the card applies. CardSpace thereby uses the metaphor of the various cards that we use to identify ourselves in the physical world, such as business cards, driver's licenses and credit cards. With these virtual cards, users no longer have to hassle with daunting passwords. CardSpace has been integrated into Microsoft's operating system Vista.

OpenID (OpenID, 2006) is an open specification of a truly distributed identity system. OpenID providers are essentially authentication brokers between users and OpenID-enabled websites. They allow users to log into an OpenID-suported website without registration, using a URI as a username that belongs to the user (e.g., the URL of her homepage or blog). Users' passwords and other credentials are safely stored by OpenID (which can be run by the user or by a third-party identity provider). Because of its open and distributed nature, ease of use, and easy adoption for websites (free libraries are available in most web programming languages), OpenID is gaining more and more momentum and emerges as the de-facto industry standard.

### 3.2.5 Authorization and access control

Authorization involves granting or denying specific access rights. In a classic access control model, an access matrix specifies what permissions each subject has on the resources the system retains. In a role-based access control model, permissions are assigned to roles instead of subjects directly (subjects can take on multiple roles, and multiple subjects can take on the same role). In a directory-based access control model, subjects are managed and organized in directories (e.g., in an LDAP server), and permissions are granted based on these different directories (Cannon, 2005).

Privacy policy languages such as P3P and XACML have an access control aspect since they prescribe who can access what information under what condition for what purpose.

### 3.2.6 Systems for empowering users in their privacy decisions

Security has long been primarily regarded as a technical and theoretical problem. It is well known though that many established security mechanisms are barely used in practice since they pose usability problems. A growing number of security researchers have therefore shifted towards so-called “usable security and privacy”, which studies the usability of security and privacy mechanisms. This emerging field aims at uncovering the reasons behind the mismatch between technical security mechanisms and their practical usage by end users, and on ways of bridging the gap to better meet users’ security needs.

Whitten and Tygar (1999) conducted a seminal usability analysis of PGP 5.0, a popular encryption tool, to find out why users failed to achieve their security goals (encrypting and decrypting email messages in this case). They found that this is largely due to interface design problems, causing a mismatch between users’ needs and the structure of the encryption technology. Bellotti and Sellen (Bellotti and Sellen, 1993) identified two primary sources for a number of potential security and privacy problems, from their experiences in ubiquitous computing: disembodiment (the actors are invisible in actions) and dissociation (actions are invisible to actors), both of which are visibility issues.

In the light of rendering the invisible visible (privacy threats in this case), Ackerman and Cranor (1999) proposed privacy critics that are semi-autonomous agents and can monitor users' online actions, warn users about potential privacy threats and suggest available countermeasures. Gideon et al. (2006) and Tsai et al. (2007) confirmed the effectiveness of such awareness mechanisms empirically.

de Paula et al. (2005) moved one step further. Instead of simply examining the usability of secure mechanisms, they framed security as an interaction problem (a practical, situated and contingent problem of decision making) and looked at a broader concern: "how security can manifest itself as part of people's interactions with and through information systems". In other words, security cannot be confined within components of a system specifically designed to attain security, but is an intrinsic and pervasive aspect of a broader context that includes end users, work practices and information systems. They argued that in practice the key issue is not how theoretically secure the underlying security mechanisms are, but rather to what extent end users can understand and make effective use of the secure mechanisms. They deliberately turned their "attention away from traditional considerations of expression and enforcement and towards explication and engagement". They designed Impromptu, a peer-to-peer file-sharing application based on supporting informed decision-making via two design principles: (1) the dynamic real-time visualization of system state, and (2) the integration of configuration and action. The former principle aims at helping users comprehend and assess the consequences of their actions when making privacy decisions. The later is based on the observation that "the separation of configuration and action may result in either overly rigid or ineffective control over security".

In short, these solutions underlie the strategy dubbed as "user empowerment" – helping users make informed privacy decisions.

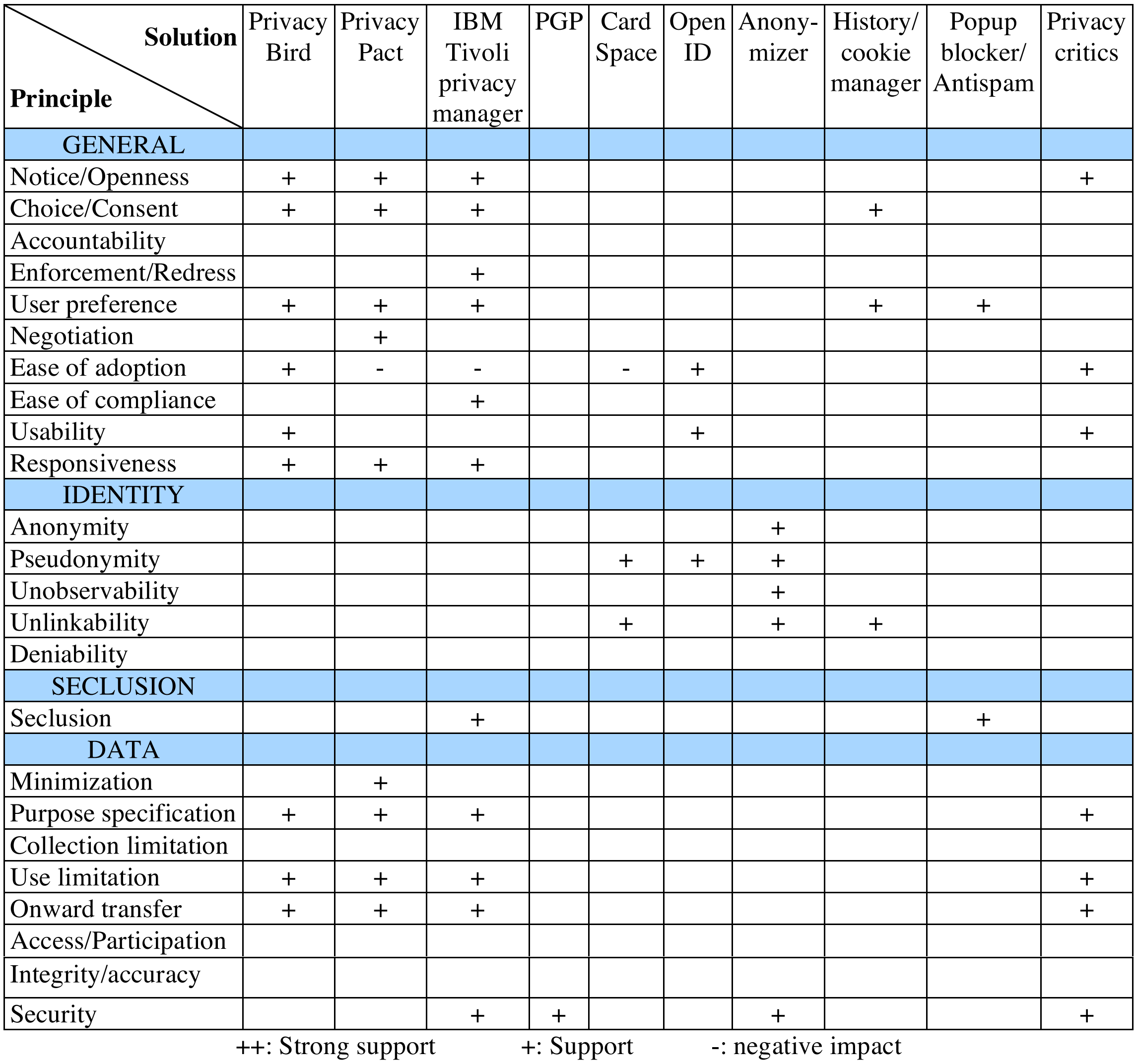
### 3.2.7 Discussion

Figure 3.5 below presents how a set of representative PETs address the privacy concerns, and Figure 3.6 shows in what ways these solutions follow the privacy principles. We now discuss some observations from these tables, and then propose implications for future research.

| | Control over data | | | | | | | Seclusion | Protection of identity |
|---|---|---|---|---|---|---|---|---|---|
| | Improper acquisition | | | Improper use | | | Improper storage | Unwanted solicitation | Identity fraud/theft |
| | Improper access | Improper collection | Improper monitoring | Improper analysis | Improper merge | Improper transfer | | | |
| Privacy Bird | | + | + | + | + | + | + | + | |
| Privacy Pact | | + | + | + | + | + | + | + | |
| IBM Tivoli privacy manager | | + | + | ++ | ++ | ++ | ++ | + | ++ |
| PGP | ++ | | | | | | | | |
| CardSpace | | | + | + | ++ | + | | | + |
| OpenID | | | + | + | ++ | + | | | + |
| Anony-mizer | | | ++ | + | + | + | | | + |
| History/ cookie manager | + | ++ | ++ | + | + | | + | | + |
| Popup blocker/ Antispam | | | | | | | | ++ | |
| Privacy critics | | + | + | + | + | + | + | + | + |

++: Effective +: Partially effective

Figure 3.5: How PETs address privacy concerns



| Solution<br>Principle | Privacy Bird | Privacy Pact | IBM Tivoli privacy manager | PGP | Card Space | Open ID | Anony-mizer | History/ cookie manager | Popup blocker/ Antispam | Privacy critics |
|---|---|---|---|---|---|---|---|---|---|---|
| GENERAL | | | | | | | | | | |
| Notice/Openness | + | + | + | | | | | | | + |
| Choice/Consent | + | + | + | | | | | + | | |
| Accountability | | | | | | | | | | |
| Enforcement/Redress | | | + | | | | | | | |
| User preference | + | + | + | | | | | + | + | |
| Negotiation | | + | | | | | | | | |
| Ease of adoption | + | - | - | | - | + | | | | + |
| Ease of compliance | | | + | | | | | | | |
| Usability | + | | | | | + | | | | + |
| Responsiveness | + | + | + | | | | | | | |
| IDENTITY | | | | | | | | | | |
| Anonymity | | | | | | | + | | | |
| Pseudonymity | | | | | + | + | + | | | |
| Unobservability | | | | | | | + | | | |
| Unlinkability | | | | | + | | + | + | | |
| Deniability | | | | | | | | | | |
| SECLUSION | | | | | | | | | | |
| Seclusion | | | + | | | | | | + | |
| DATA | | | | | | | | | | |
| Minimization | | + | | | | | | | | |
| Purpose specification | + | + | + | | | | | | | + |
| Collection limitation | | | | | | | | | | |
| Use limitation | + | + | + | | | | | | | + |
| Onward transfer | + | + | + | | | | | | | + |
| Access/Participation | | | | | | | | | | |
| Integrity/accuracy | | | | | | | | | | |
| Security | | | + | + | | | + | | | + |

++: Strong support +: Support -: negative impact

Figure 3.6: What privacy principles PETs follow

Two observations can be made in these two tables. First, privacy protection solutions form clusters. Solutions of the same type tend to address almost identical privacy concerns by following similar privacy principles, while different types of solutions address different but not necessarily disjoint concerns, and follow different but not necessarily disjoint principles. For example, P3P-enabled user agents such as Privacy Bird, PrivacyPact and Ackerman and Cranor's privacy critics address all listed privacy concerns except improper access. They do this by applying general principles (e.g., Notice/Openness) and data principles (e.g., purpose specification) but not identity principles. In contrast, identity management tools (such as CardSpace and OpenID) and anonymizers (e.g., Anonymizer) attend to concerns such as improper monitoring and improper use (e.g., improper merge) by observing identity principles (e.g., pseudonymity) but not data principles. This phenomenon indicates that research in data protection and identity management is still somewhat fragmented, albeit some overlap exists. Since both form integral parts of privacy enhancement, collaborations between the two research communities to integrate the two types of PETs would be desirable.

Second, no current PET effectively addresses all privacy concerns, nor follows all privacy principles. The IBM Tivoli privacy manager is the most comprehensive solution among those examined in this section. However, since it is designed and implemented as a server-side enterprise privacy management system, principles like usability (i.e., a PET solution should be easy for *end users* to adopt) are inevitably hard to achieve. Rather than mulling over whether one can develop a technical solution that effectively addresses all privacy concerns, our pragmatic strategy is to merely highlight directions that deserve more attention. For example, principles such as responsiveness, enforcement and ease of compliance are barely supported by the discussed solutions, except for the IBM Tivoli privacy manager.

Our investigation of existing privacy enhancing technologies yield the following insights:

- Privacy needs to be treated as a first-class requirement from the early onset in the design of an information system since, like for security and usability, it is extremely difficult if not impossible to “retrofit” a completed system to make it more privacy-friendly.

- While compliance has long been technically framed and treated as a server-side problem, we believe that the “user empowerment” strategy has a great potential for compliance since the “expression and enforcement” paradigm seems too rigid to accommodate users’ changing and context-dependent privacy desires.

- Since users' privacy needs and preferences are inherently dynamic and contingent, solutions need to cater to users' individual privacy needs. We start to see solutions like negotiable privacy policies that follow this promising direction.

## 3.3 Technical Solutions for Privacy-Enhancing Personalization

In this section we evaluate and discuss the major existing privacy-enhancing personalization solutions against the analytical framework introduced in Section 3.1, specifically, what basic privacy-enhancing techniques they employ, and how these solutions relate to the described principles and privacy concerns.

### 3.3.1 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 a user to have more than one pseudonym/account/role/persona, so that the user can keep apart different aspects of their online activities (e.g., work versus entertainment).

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, a password and an email address) for establishing an anonymous account at the website. Janus also supports anonymous email exchanges from a website to a user, and filters the potentially identifying information of the HTTP protocol to preserve user privacy.

Arlein et al. (2000) suggested an infrastructure that enables global user profiles to be maintained and accessed by different merchants. Users can control their data disclosure by grouping their information into profiles pertaining to different personae and can selectively authorize merchants to access these profiles. The infrastructure includes a persona server to assist users manage their personae. The persona server is separate from the profile database, so as to prevent linking different profiles of the same user. Besides, the infrastructure also has a tainting-based access control mechanism that allows merchants to designate which data about user interaction at their sites can be accessed by other merchants.

Ishitani et al. (2003) implemented a system called Masks (Managing Anonymity while Sharing Knowledge to Servers). The system consists of both server-side and client-side components, namely the Masks server and the privacy and security agents (PSAs). 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 enables the users to receive group-based personalization. The PSAs runs with users' web browsers and allows users to configure the masks as well as other functionalities such as blocking and filtering cookies and web bugs.

Kobsa and Schreck (2003) proposed 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.

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

Figure 3.7 presents an analysis of the aforementioned pseudonymous personalization systems along the following characteristics:

1. Alias-to-website cardinality

   The alias-to-website cardinality describes the relationship between the number of aliases pertaining to a user and the number of websites at which the alias(es) may be used. For example, a cardinality of 1:1 means that each user will have exactly one alias for every website, while 1:n means that a user has one global alias/profile for all websites, and m:n means that a user can have an arbitrary number of aliases for any number of websites.

2. User control

   User control denotes whether the system allows users to control the usage of their alias/profile at different websites.

3. Personalization

   This factor evaluates to what extent the websites can provide personalized services to users. For example, a site can provide personalized services using the user's interaction logs with this site, or it could use the logs from multiple sites.

4. Sender anonymity

   Sender anonymity indicates whether or not users are identified in the interactions.

5. Receiver anonymity

   Receiver anonymity indicates whether websites are identified in the interactions.

6. User Modeling Server (UMS) anonymity

   UMS anonymity indicates whether or not user modeling servers (or more general, the repositories that store the user models/profiles) are kept anonymous.

7. Content-based anonymity

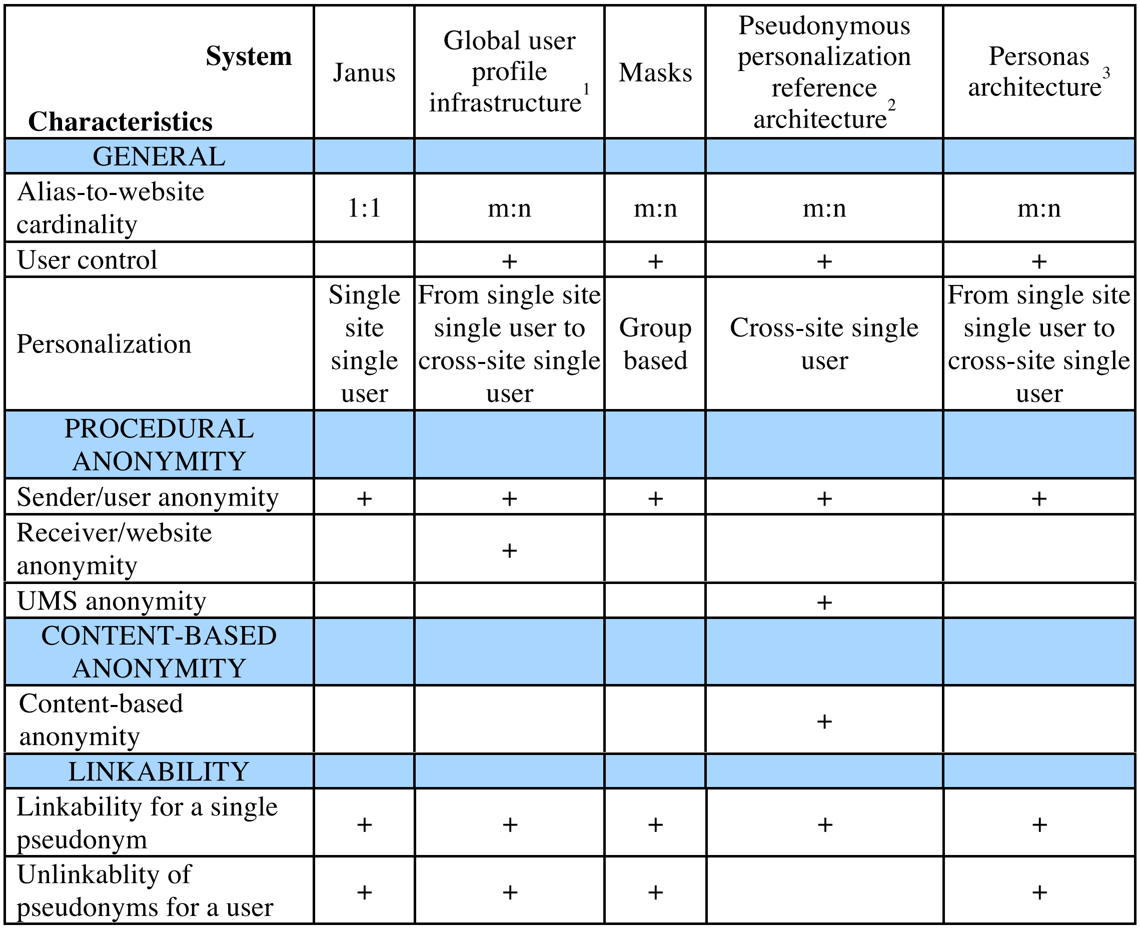
   Content-based anonymity prevails when no identification by means of the exchanged data is possible.

8. Linkability for a single pseudonym

   This characteristic indicates whether or not a user's interaction steps or sessions with one or multiple websites can be linked using one pseudonym of hers.

9. Unlinkability of pseudonyms for a user

   This characteristic indicates whether or not multiple pseudonyms pertaining to the same user can be linked.



| System<br>Characteristics | Janus | Global user profile infrastructure[1] | Masks | Pseudonymous personalization reference architecture[2] | Personas architecture[3] |
|---|---|---|---|---|---|
| GENERAL | | | | | |
| Alias-to-website cardinality | 1:1 | m:n | m:n | m:n | m:n |
| User control | | + | + | + | + |
| Personalization | Single site single user | From single site single user to cross-site single user | Group based | Cross-site single user | From single site single user to cross-site single user |
| PROCEDURAL ANONYMITY | | | | | |
| Sender/user anonymity | + | + | + | + | + |
| Receiver/website anonymity | | + | | | |
| UMS anonymity | | | | + | |
| CONTENT-BASED ANONYMITY | | | | | |
| Content-based anonymity | | | | + | |
| LINKABILITY | | | | | |
| Linkability for a single pseudonym | + | + | + | + | + |
| Unlinkablity of pseudonyms for a user | + | + | + | | + |

+: Support

Figure 3.7: Pseudonymous personalization systems and their characteristics

At first sight, pseudonymous personalization seems to be a panacea for all privacy problems because 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 customer by mail). Besides, users may still have additional privacy preferences such as not wanting to be profiled even when done pseudonymously only, to which personalized systems need to adjust. Moreover, Rao and Rohatgi (2000) pointed out that pseudonymity, or more broadly, hiding explicit identity information (e.g., name, email address) is not sufficient to guarantee privacy. They demonstrate using a technique from stylometry (a field of linguistics that uses syntactic and semantic information to ascribe identity or authorship to literary works), and principal component analysis of function words, to attack pseudonymity. Similar findings were made for database entries (Sweeney, 2002), web trails (Malin et al., 2003), and query terms (Nakashima, 2006).

### 3.3.2 Distributed personalization

Distributed personalization for safeguarding users' privacy has so far primarily been investigated in the domain of collaborative filtering (CF). Collaborative filtering is a popular technique for generating personalized recommendations using other users' preferences. The underlying assumption is that a user will prefer things that similar users like. In general, CF techniques use weighted combinations of nearest neighbor ratings to make predictions based on a user's preferences. A number of algorithms exist to determine proximity, including correlation between users, vector similarity methods, Bayesian clustering and Bayesian networks.

In recommender systems based on CF techniques, distribution may affect two aspects: the storage of personal profiles, and computation aspects (such as neighborhood formation and prediction generation). One argument why distribution leads to better privacy protection is that users may have better control over their own data if they are stored at the client side as compared to a central (user modeling) server. What is more important though is that CF computation is performed in a distributed and cooperative fashion rather than centrally. Personalization either takes places at the client side using merely the user's data, or is realized by specific privacy-preserving collaborative filtering schemes such as the ones described below.

**?** is a multi-agent distributed matchmaking system that learns about users by finding sets of keywords that characterize a user's interests. It matches users with similar interests by comparing their keywords without disclosing their identities. If a match is found, the Yenta clients can discretely negotiate to decide whether the matched users would like to reveal their identities to each other. Yenta utilizes anonymity/pseudonymity and encryption in protecting users' privacy.

Olsson (1998) describes a decentralized social filtering model that is built on interactions between collaborative software agents performing content-based filtering. This system is similar to Yenta but differs in its way of measuring similarity between different users via trust rather than interests as in Yenta.

Canny (2002a,b) outlined a peer-to-peer collaborative filtering model in which users' profiles are all stored at the client side so that users can fully control their data. The underlying multi-party computation scheme allows a community of users to compute an aggregate of their data (i.e., a singular value decomposition (SVD) model of the user-item matrix) based solely on vector addition so that individual data will not be disclosed. This non-disclosure property is achieved by using techniques including ElGamal encryption, homomorphic encryption and Zero Knowledge Proofs.

Miller et al. (2004) propose a peer-to-peer CF algorithm called PocketLens. For each individual user, PocketLens first searches for neighbors in the P2P network, then incrementally updates the user's individual item-item similarity model by incorporating one neighbor's ratings at a time (the neighbor's ratings will be discarded after updating the model), and finally generates recommendations based on the model. The paper also compares and discusses five implementation frameworks:

- a central server architecture where the key data is stored on a central server while the computations are performed at each individual node;

- a random discovery architecture that allows users to remain anonymous and uses Gnutella's ping/pong mechanism for finding neighbors;
- a transitive traversal architecture that allows clients to share their neighborhood lists by query flooding and thus enables neighborhood formation via a form of transitivity;
- a content-addressable architecture that adopts P2P file sharing networks, e.g., Chord, which places a deterministic overlay routing system over the network and provides a scalable and distributed lookup function (the II-Chord implementation described in the paper uses the network basically as a distributed storage mechanism to collaboratively build and maintain the item-item matrix); and
- a secure blackboard architecture that leverages the secure operations used in a secure online voting protocol and in Canny's work (Canny, 2002a,b), whereby each client writes encrypted partial results to a Write Once Read Many (WORM) blackboard and the final model is generated by incorporating those partial profiles.

Gilburd et al. (2004) introduce a k-TTP (trusted third party) model which suggests that privacy is preserved as long as no participant of a distributed (joint) computation learns statistics of a group with less than k members. This is less restrictive than an ordinary TTP model in the sense that it does not protect unauthorized access to statistics of individual users if less than k members participate in a joint computation, and is thus more flexible. The authors demonstrate that k-TTP enables more scalable distributed computation schemes. While the paper illustrates the idea of k-TTP by an association-rule mining algorithm, the same idea could be applied to personalization techniques such as collaborative filtering. Berkovsky et al.'s idea of super-peers echoes the same aggregation spirit (Berkovsky et al., 2006).

### 3.3.3 Privacy-preserving collaborative filtering

The aim of work in this area is to apply and extend privacy-preserving data mining techniques in the area of collaborative filtering. The common approach for achieving privacy preservation in data mining tasks is to replace each message exchange in an ordinary distributed data mining algorithm with a cryptographic primitive that provides the same information without disclosing the data of the individual participants. The research challenge here is to enable users to contribute their information for CF purposes without compromising their privacy (e.g., through exposure of their personal data). Here, privacy-preserving CF is treated as a secure multiparty computation problem where users and different websites jointly conduct CF computations based on their private data. These parties could be mutually untrusted, or even competitors. Typical ways of privacy preservation include encryption, aggregation, perturbation and obfuscation.

#### Encryption

In this type of work, CF computation is based on encrypted user data. An example is the abovementioned work of (Canny, 2002a), which describes a secure multi-party computation scheme that allows a community of users to compute an aggregate of their data without disclosing individual data by using homomorphic encryption and ElGamal encryption. More specifically, a combination of ElGamal encryption and homomorphic encryption allows vectors to be added by multiplying the encrypted addends, and the final result to be decrypted. Individual addends can be verified as valid data using zero knowledge proofs. The resultant aggregate SVD model can then be used to generate personalization.

### Randomized perturbation

Polat and Du (2003, 2005a,b) demonstrate the usage of randomized perturbation techniques (adding random numbers from a given range to the original data) in disguising the original user ratings before feeding them into CF algorithms based on correlation and singular value decomposition. The CF system thereby does not know the exact values of the original ratings, yet is still able to compute reasonably accurate recommendations. The underlying reason is that the CF algorithms often use aggregations like scalar products and sums, and that the perturbations tend to cancel themselves out.

### Aggregation

In this privacy-protecting approach (e.g., (Canny, 2002a)), users' personal data are aggregated in such a way that an individual's data cannot be identified.

### Community model

In this approach, CF computation (e.g., model generation) is carried out collaboratively by a community of clients. The difference to aggregation techniques is that a community model may not generate an aggregate model and may still reveal individual user's data, e.g., in the II-Chord implementation of PocketLens (Miller et al., 2004). Both aggregate and community model can also be considered as examples of distributed personalization, since they either store personal profiles or perform CF computation in a distributed manner.

**Obfuscation**

Another way of disguising users' personal data is via obfuscation. Berkovsky et al. (2005) describe a decentralized CF model in which user profiles are stored at the client side. In this approach, some of the personal data is replaced by some other data (which is either constant or drawn from some distribution). The authors demonstrate that relatively large parts of the user profile can be obfuscated while CF can still generate reasonably accurate recommendations. In their follow-up work (Berkovsky et al., 2006), they propose a decentralized recommendation generation scheme that is based on a hierarchical neighborhood topology. More specifically, users (peers) are organized into groups managed by super-peers. To enhance privacy, the super-peers choose only a random subset of their peers to form the neighborhood of similar users. To protect individual peers' privacy within a peer-group, the obfuscation techniques can be used and also only a subset of peers can be queried.

### 3.3.4 Scrutable personalization

Kay et al. (2003); Kay (2006) suggest putting scrutability into user modeling and personalized systems. By scrutability the authors mean that users can understand and control what goes into their user model, what information from their model is available to different services, and how the model is managed and maintained. Their user modeling system Personis applies three privacy-enhancing mechanisms to control the protection of each unit of personal information ("evidence") in the user model (Kay et al., 2003):

- expiration dates and purging of older evidence,
- compaction, for replacing a set of evidence from a single source with an aggregate, and

- morphing, which replaces an arbitrary collection of evidence.

For controlling the usage of evidences from the user model, Personis allows users to restrict the evidences that are available to applications, and the methods that may generate a user model and operate on it. Despite the desirability of scrutability from a privacy point of view, its implementation and control is currently very challenging, due to users' lack of understanding of these notions and of effective and efficient user interfaces to support them. Moreover, scrutability may reveal the personalization methods that a website uses, which may pose a problem in application areas in which those are considered to be competitive advantages and therefore confidential (e.g., in online retail websites).

### 3.3.5 Task-based personalization

Herlocker and Konstan (2001) propose a content-independent task-focused recommendation scheme. The scheme assumes that a traditional recommender system may already possess historical ratings data, and that recommendation is possible with data that pertain to the current session or specific task only (e.g., buying a martial arts DVD) rather than collecting a comprehensive profile of the user across multiple sessions. The system builds an item-item association model based on the legacy ratings, and uses the model to generate recommendations. The privacy improvement is that users do not need to disclose their historical ratings while still being able to receive task-focused recommendations. Cranor (2003) also supports task or session based personalization as a way to reduce privacy risks and make privacy compliance easier. However, the price is that the recommendations are not truly personalized, i.e., all users may receive the same recommendations for the same task.

### 3.3.6 Analysis of technical solutions for privacy-enhanced personalization

We have seen that different privacy enhancing solutions for personalized systems often implement several basic techniques. Figure 3.8 gives a summary of the techniques used in the discussed systems. Figure 3.9 shows how well a set of representative privacy protection solutions from the ones discussed above meet the privacy principles described earlier. Figure 3.10 presents how these solutions address the privacy concerns in web personalization described earlier. The following observations can be made:

First, several solutions aim for a balance between privacy and personalization. Examples include pseudonymous personalization, scrutable personalization and dynamic personalization. They all address a handful of privacy concerns and achieve at least reasonably good personalization.

Second, none of the solutions in Figure 3.8 uses all available privacy-enhancing techniques. We believe more comprehensive future solutions will need to incorporate a variety of basic privacy enhancing techniques.

Third, none of the solutions in Figure 3.10 addresses all privacy concerns, except Personis which relies on a “user empowerment” strategy. However, Personis does not address all the concerns effectively. For example, it does not provide comprehensible and effective user interfaces since most users do not usually possess mental models of the operation of user modeling systems.

Finally, we find that principles such as onward transfer, enforcement, user preference, negotiation, ease of compliance and responsiveness are currently insufficiently observed. Taking “onward transfer” as an example, no current privacy-enhancing solution in web personalization allows ”sticky” privacy policies that travel with data so that, e.g., user data cannot be copied and transferred by an entity that is only allowed to read the data. Techniques used in Digital Rights Management (DRM) (Rosenblatt et al., 2001) may be adapted for this purpose.

| **Technique**<br>**System** | A/P | En | SD | CD | Ag | CM | Pe | Ob | ScS | TP | DS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Yenta | X | X | X | X | | | | | | | |
| Trust-based Social Filtering (Olsson 1998) | | | X | X | | | | | | | |
| PocketLens Central Server | | | | X | | | | | | | |
| PocketLens Random Discovery | X | | X | X | | | | | | | |
| PocketLens Transitive Traversal | X | | X | X | | | | | | | |
| PocketLens II-Chord | | | X | X | | X | | | | | |
| PocketLens Secure Blackboard | | X | X | X | X | X | | | | | |
| k-TTP | | X | X | X | X | X | | | | | |
| Privacy Preserving CF (Canny 2002a) | | X | X | X | X | X | | | | | |
| Factor Analysis-CF (FA-CF) (Canny 2002b) | | X | X | X | X | X | | | | | |
| Random Perturbation-CF | | | | | | | X | | | | |
| Privacy Enhancing CF (Berkovsky et al. 2005) | | | X | X | | | | X | | | |
| Hierarchical Neighborhood Topology-CF (HNT-CF) (Berkovsky et al. 2006) | | | X | X | X | | | X | | | |
| Personis | | | X | X | X | | X | | X | | |
| Task-based Personalization | | | | | | | | | | X | |
| Privacy-Tailored Personalization | | | | | | | | | | | X |

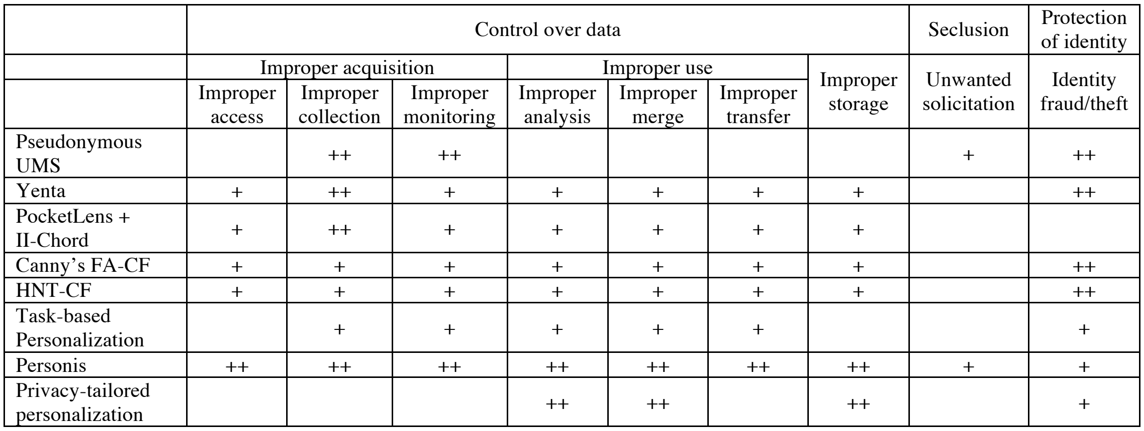
A/P: Anonymity/pseudonymity En: Encryption SD: Storage distribution
CD: Computation distribution Ag: Aggregation CM: Community model
Pe: Perturbation Ob: Obfuscation ScS: Scrutability support
TP: Task-based personalization DS: Dynamism support

Figure 3.8: Basic privacy protection techniques used in privacy-enhanced personalization solutions

| Solution<br>Principle | Pseudonymous UMS | Yenta | PocketLens + II-Chord | Canny's FA-CF | HNT-CF | Task-based CF | Personis | Privacy-tailored personaliztion |
|---|---|---|---|---|---|---|---|---|
| GENERAL | | | | | | | | |
| Notice/Awareness | | | | | | | ++ | |
| Choice/Consent | | + | + | + | + | | + | + |
| Enforcement/Redress | + | + | | + | | | + | ++ |
| User preference | | | | | | | + | ++ |
| Negotiation | | + | | | | | | |
| Ease of adoption | - | | | - | | | | + |
| Ease of compliance | | | | | | | | ++ |
| Usability | | | | | | | - | |
| Responsiveness | | | | | | | | ++ |
| Personalization quality | ++ | + | ++ | + | ++ | + | + | ++ |
| IDENTITY | | | | | | | | |
| Anonymity | ++ | | + | | | | | |
| Pseudonymity | ++ | + | | ++ | + | | + | |
| Unobservability | ++ | | + | + | + | | | |
| Unlinkability | | | + | ++ | | | | |
| Deniability | | | | | + | | | |
| SECLUSION | | | | | | | | |
| Seclusion | | | | | | | | |
| DATA | | | | | | | | |
| Minimization | | | | + | + | ++ | | ++ |
| Purpose specification | | | | | | | + | + |
| Collection limitation | | | | | | | + | |
| Use limitation | + | | | | | | + | ++ |
| Onward transfer | | | | | | | | |
| Access/Participation | | | | | | | ++ | |
| Integrity/accuracy | | | | | | | + | |
| Security | + | + | | + | + | | | |

++: Strong support +: Support -: Negative impact

Figure 3.9: An analysis of privacy protection solutions in web personalization



| | Control over data | | | | | | | Seclusion | Protection of identity |
|---|---|---|---|---|---|---|---|---|---|
| | Improper acquisition | | | Improper use | | | Improper storage | Unwanted solicitation | Identity fraud/theft |
| | Improper access | Improper collection | Improper monitoring | Improper analysis | Improper merge | Improper transfer | | | |
| Pseudonymous UMS | | ++ | ++ | | | | | + | ++ |
| Yenta | + | ++ | + | + | + | + | + | | ++ |
| PocketLens + II-Chord | + | ++ | + | + | + | + | + | | |
| Canny’s FA-CF | + | + | + | + | + | + | + | | ++ |
| HNT-CF | + | + | + | + | + | + | + | | ++ |
| Task-based Personalization | | + | + | + | + | + | | | + |
| Personis | ++ | ++ | ++ | ++ | ++ | ++ | ++ | + | + |
| Privacy-tailored personalization | | | | ++ | ++ | | ++ | | + |

++: Effective     +: Partially effective

Figure 3.10: How existing solutions address privacy concerns in web personalization

### 3.3.7 Discussion

We now discuss the major findings of our analysis of existing technical PEP solutions from two points of views, namely the one of users and of websites.

#### Users

User would like to enjoy personalized services of websites while at the same time have their individual privacy needs respected (Kobsa, 2007b). The traditional strategy for addressing users’ privacy needs is through expression and enforcement – users specify their privacy needs which are then translated into formal expressions and finally enforced in technical solutions.

There are several problems with this strategy. First, privacy decisions (e.g., whether to disclose one’s telephone number in a particular situation) are inherently contingent and situated. As DiGioia and Dourish (2005); Dourish and Anderson (2006) point out, the artificial separation of configuration and action may be overly rigid or ineffective. Second, it is a known fact that users' actual behaviors may diverge from their stated privacy attitudes or preferences (Spiekermann et al., 2001b). Third, we observe that currently available technical privacy languages fall short of expressing users' highly flexible and nuanced privacy needs. This may well be an inevitable "social-technical gap"(Ackerman, 2000) between human activities/decisions and what we can support technically. Forth, even if users' privacy decisions could be accurately translated into enforceable specifications, we notice that the majority of existing solutions lack enforcement mechanisms that respond to users' unpredictable changes of privacy decisions in an effective manner.

We see two emerging ways of alleviating or solving these problems:

1. by empowering users to make informed decisions (e.g., by giving them insights into the consequences of their actions through visualizations of system states and events, by enabling them to carry out their privacy decisions rather than merely expressing them through integration of configuration and action (de Paula et al., 2005), or by providing scrutability support in user models (Kay, 2006));
2. by supporting the negotiation between users and websites to reach a consensus on the privacy practices of websites (e.g., (Buffett et al., 2004; Preibusch, 2006)).

### Websites

One of the pressing challenges that websites face today is the need to provide competitive value-added personalized services to its users while complying with a growing number of regulatory privacy requirements. From our survey, we recognize deficiencies in the area of compliance (see Figure 3.10). More specifically, we witness that compliance-related principles such as enforcement and ease of compliance are mostly not addressed, with the exception of a few solutions based on the abovementioned "expression and enforcement" strategy such as in the IBM Tivoli privacy manager (IBM, 2003a). From the previous section we can infer though that this approach may run into problems when users become involved.

In the light of this, we coarsely categorize regulatory privacy requirements into two types. The first type consists of requirements that can be met without user involvement (we call them "website-exclusive" requirements). An instance of this type is "*usage data must be erased immediately after each session*" *(*except for very limited purposes) (DE-TML, 2007). The second type consists of requirements that may include privacy decisions of the user (we call them "user-involving" requirements). Examples are "*users must be able to withdraw their consent to the processing of traffic and location data at any time* (EU, 2002) ", and "*value-added (e.g. personalized) services based on traffic or location data require the anonymization of such data or the user's consent* (EU, 2002) ".

Since "user-involving" requirements could be fulfilled by users‘ involvement (giving their consent), we believe that this type of privacy requirements might also be well addressed by using some of the alternatives to the expression and enforcement approach that were discussed in the previous section. We expect new solutions to emerge in the future that follow these alternate directions.

In contrast, the traditional strategy of expression and enforcement is by and large appropriate and effective for fulfilling the website-exclusive obligations. First, because of its website-exclusiveness, the user empowerment alternative is obviously irrelevant. Second, the separation of expression and enforcement is no longer a problem here, for three reasons: (1) website-exclusive requirements are usually unambiguous and rigid, and thus amenable to accurate formal expressions; (2) there are tools available that can automatically translate textual requirements into specifications in formal languages like P3P (e.g., IBM's Sparcle (Karat et al., 2005)); and (3) once put into effect, privacy laws and regulations are fairly stable, and changes are normally known a few months before they become effective.

While expression can become much easier with support through tools like Sparcle, enforcement is still quite challenging, for the following reasons.

- An effective enforcement mechanism needs to cover the whole lifecycle of user data from collection to usage to transfer, etc.
- In centralized user modeling systems (which collect and supply user information from and to different websites for usually different purposes), the complexities of defining different permissible purposes for collecting and using personal data must be addressed.
- For legacy systems it is likely that privacy had been disregarded during their design and implementation. As with usability, research has revealed though that privacy and security cannot be an afterthought in system design (Dourish et al., 2004; de Paula et al., 2005; Dourish and Anderson, 2006). The support of the enforcement of privacy in legacy systems is therefore likely to be very hard.

## 3.4 Escape Strategies for Privacy-Enhacing Personalization

Besides the technical solutions discussed above, websites also use simple escape strategies to cope with privacy constraints of international users.

### 3.4.1 Largest permissible common subset

Ideally, this approach means that only those personalization methods are used that satisfy all privacy laws and regulations. The Disney website, for instance, observes both the U.S.

Children's Online Privacy Protection Act (COPPA) as well as the European Union Directive (Disney, 2002). This solution is likely to run into problems if more than a very few jurisdictions are involved, since the largest common subset of permissible personalization methods may then become very small.

### 3.4.2 Different country/region versions

In this approach, personalized systems have different country versions, each of which uses only those personalization methods that are permitted in the respective country. If some countries have similar privacy laws, their versions can be combined using the above-described largest permissible common subset approach. For example, IBM's German-language pages comply with the privacy laws of Germany, Austria and Switzerland (IBM, 2003b), while IBM's U.S. site meets the legal constraints in U.S. As with the largest permissible common subset approach, this approach also has scaling problems as soon as the number of countries/ regions, and hence the number of different versions of the personalized system, increases.

### 3.4.3 Discussion

These simple escape strategies may be a reasonable first step for personalized websites that are only dealing with a small number of jurisdictions. However, they simply cannot be scaled up easily, if not impossible. In addition, they cannot address users' individual privacy preferences. They can be at best sensitive to different jurisdictions but cannot differentiate two users who are from the same country/jurisdiction but have disparate privacy preferences. In short, these strategies fail to provide a flexible, systematic and scalable solution for addressing privacy constraints that may differ among users.

## 3.5 Summary

Privacy and web personalization are in tension with each other. The more user data websites collect and utilize, the better are generally the personalized services they provide but the more potential privacy concerns may arise. With the enactment of privacy laws and regulations worldwide, the conflict is even more acute because personalized websites are obliged to comply with their provisions, which often have remarkable impacts on how personalization may be performed.

In analyzing the related work in the area of privacy-enhancing technologies in general, and technical solutions for privacy-enhancing personalization in particular, we propose and apply a multi-faceted approach, consisting of privacy guidelines, privacy concerns, and privacy-enhancing characteristics of these solutions. We relate these facets to each other and reveal trends and identify deficiencies.

### 3.5.1 Gaps in existing work

Our analysis identifies the following gaps in existing work:

- *Gap 1: overlooked impacts of privacy constraints on personalization methods*

  As discussed earlier, privacy constraints from either laws or users' personal preferences, when applicable, may have profound impacts on personalization methods (e.g., the cross-session machine learning example in the introduction). Current approaches tend to focus on privacy enhancement from a perspective of data *collection and storage* but overlook the perspective of data "*processing*", i.e., making inferences based on user data (e.g., profiling).

- *Gap 2: lack of system-level enforcement of privacy constraints*

System-level enforcement of privacy constraints is still challenging. This is partly due to the lack of system-wide enforcement engine in personalized systems. Besides, a full-fledged privacy constraint enforcement mechanism needs to cover the whole life cycle of user data in a personalized system from collection to processing and usage to transfer, among other things. General-purpose privacy enforcement solutions such as the IBM Tivoli Privacy Manager need to be adapted in web-based personalized systems.

- *Gap 3: insufficient catering to individual user's privacy preferences*

  Since users' privacy needs and preferences are inherently dynamic and contingent, solutions need to cater to users' individual privacy needs. Current personalized systems largely lack this support with few exceptions (e.g., negotiable P3P policy ).

- *Gap 4: inadequate responsiveness of privacy enhancement solution to users' changes of privacy decisions*

  Because privacy is highly personal, situated and contextual, users may change their privacy decisions anytime. An effective and efficient privacy enhancement solution should respond promptly to changes of users' privacy decisions. This "responsiveness" of privacy enhancement is mostly unexplored.

# Chapter 4

# Our Approach

## 4.1 Our Privacy-Enhanced User Modeling Framework

Our vision is to provide personalized privacy management where the personalization process is tailored to each individual user's privacy constraints. In this section, we describe how we approach the problem in the context of personalized system design, and present our personalization framework and its underlying privacy-enabling mechanism in details.

### 4.1.1 User modeling server

Most personalized systems employ a user modeling system, usually in a client-server fashion, which is then called a User Modeling Server (UMS). A UMS stores and represents user characteristics and behavior, integrates external user-related information, applies user modeling methods to derive additional assumptions about the user, and allows multiple external user-adaptive applications to retrieve user information from the server concurrently (Kobsa, 2007a). UMSs are widely used for supporting user-adaptive applications. Our solution enhances a regular UMS by a new dimension of personalization, namely adaptation to each user's potentially different privacy constraints.

For many personalization goals, more than one method can often be used that differ in their data and privacy requirements and their anticipated accuracy and reliability. For example, a personalized website could use incremental machine learning to provide personalization to visitors from Germany (where user logs must be discarded at the end of a session to comply with the German Telemedia Act (DE-TML, 2007), see Code 5 in Section 2.2.1), while it can use possibly better one-time machine learning with user data from several sessions to provide personalization to web visitors from the U.S. who are not subject to this constraint.

Since UMSs are the central repositories for personal information in personalized systems and the loci of personal data processing, our solution focuses on using a product line architecture for UMSs, with which we address privacy and personalization issues.

### 4.1.2 Product line architecture

*Software architectures* provide high-level abstractions for representing a system's structure, behavior, and key properties. These are generally expressed using an *architecture description language* (ADL) (Medvidovic and Taylor, 2000), which captures concepts such as the elements from which systems are built, interactions among those elements, patterns that guide their composition, and constraints on these patterns (Perry and Wolf, 1992).

Whereas "normal" software architectures define the architectural structure of a single software system, a *product line architecture* (PLA) simultaneously defines the architectural structure for a set of closely-related systems or products (Bosch, 2000). As such, it must provide a basis by which an architect may understand and manipulate the commonalities and variabilities existing among each product constituting the PLA and also must support the creation of each individual product architecture from the PLA, for instance to deploy an individual product to a client.

PLAs have been increasingly used in industrial software development and significant improvement in terms of reduced development cost and time have been reported (Bosch, 2000; Clements et al., 2001; Atkinson et al., 2002). The Software Product Line Hall of Fame (SEI, 2009) lists many corporations such as Boeing, Ericsson, HP, Lucent, Nokia, Philips, and Toshiba as well as government agencies such as the U.S. Naval Research that have successfully adopted PLA. From these positive experiences, it is reasonable to say that the PLA methodology is emerging as a "best practice" for coping with software variability.

A number of ADLs support the specification of product line architectures, typically distinguishing *core elements* from *variation points*. Variation points, architectural elements themselves, specify places in the PLA where differences exist among specific product architectures. For instance, Koala uses switches (van Ommering et al., 2000), xADL 2.0 (Dashofy et al., 2005) and Ménage (Garg et al., 2003) allow optional, variant, and optional variant elements, and COVAMOF utilizes optionals, alternatives, optional variants, variants, and values (Sinnema et al., 2004). Most express these differences in some form of configuration or constraint language and promote commonly-used rules to first-class language constructs. For instance, xADL 2.0 uses Boolean expressions and Koala has a language construct for switches that route connections to one of several alternative interfaces.

In this section, we will describe a Boolean guard approach of expressing PLA variability that we first used (Wang et al., 2006b). Basically, 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. A product instance can be selected out of a product line architecture by resolving the Boolean guards of each variation point (van der Hoek, 2004). In Section 4.3, we will present an alternative approach for modeling

PLA variability and a detailed comparison with the Boolean guard approach.

### 4.1.3 Framework overview

Our goal is to design a user modeling framework that respects users' potentially different privacy constraints in a flexible, systematic and scalable manner. Inspired by the idea of treating software as a product line to support software variability from design-time to invocation-time to run-time (van der Hoek, 2004) and several other works in the field of dynamic architecture and run-time architecture evolution (Magee and Kramer, 1996; Oreizy et al., 1998; Chen et al., 2003; Garg et al., 2003; Georgas et al., 2005), we propose a dynamic, privacy-enhanced personalization framework. This framework encapsulates different personalization methods in individual components and, at any point during runtime, ascertains that only those components can be operational that are in compliance with the currently prevailing privacy constraints. Moreover, this framework can also dynamically select the component with the optimal anticipated personalization effects among those that are currently permissible (Kobsa, 2003). We conceptualize and operationalize this framework as a PLA. Simplistically speaking, this framework gives every user their own UMS instance which incorporates those user modeling methods only that meet the user's current privacy constraints (Wang et al., 2006b). Doing so allows us to provide a framework that solves the problem of handling privacy constraints in web personalization in a generic fashion, to take advantage of commonalities among different needs for privacy and personalization, and to dynamically update different privacy and personalization strategies in a *modular* fashion, not requiring that the UMS be entirely rebuilt upon each change.

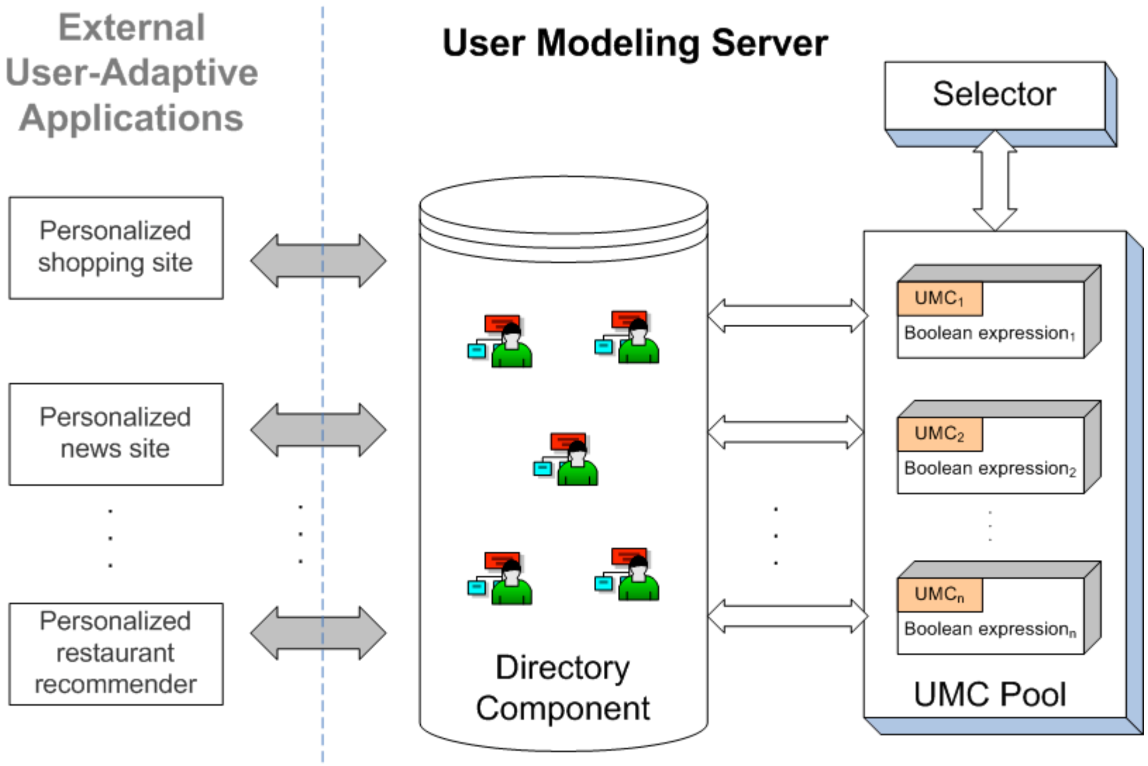


Figure 4.1: A privacy-enhanced user modeling framework

Figure 4.1 shows an overview of our PLA-based user modeling framework. This framework is largely inspired by and based on a user modeling server (UMS) architecture proposed by Kobsa and Fink (Kobsa and Fink, 2006). Compared with their server architecture, our framework adds the Selector and makes the UMCs optional components associated with Boolean guards in the PLA. More specifically, our framework consists of external user-adaptive applications, the Selector, and the LDAP-based UMS of Kobsa and Fink (Kobsa and Fink, 2006) which includes the Directory Component and a pool of user modeling components (UMCs). External personalized applications can query the UMS for existing user information, so as to provide personalized services to their end users, and can supply additional user information to the UMS. The Directory Component is essentially a repository of user models, each of which stores and represents not only users' characteristics, behavior and inferences, but also their potentially different individual privacy constraints. The UMC Pool contains a set of UMCs, each of which encapsulates one or more user modeling methods (e.g., collaborative filtering (Resnick et al., 1994)) that make inferences about users based on existing user data. Each UMC forms an *optional element* (van der Hoek et al., 2001) associated with a Boolean guard in the PLA.

A particular personalization architecture containing only those UMCs that are allowed to operate under a user's prevailing privacy constraints can be selected from the PLA by the Selector, and then instantiated to provide services to the external personalized applications as a UMS for the respective user. The novel privacy enhancement consists in every user having their own instance of the UMC Pool, each containing only those user modeling components that meet the privacy requirements for the respective user (users with identical privacy constraints share the same instance).

### 4.1.4 Modeling privacy constraints and their impacts on UMCs

A Boolean guard captures whether its associated UMC is allowed to operate under a set of identified privacy constraints. A Boolean guard is a logic combination of Boolean expressions, which are defined during a manual analysis of the impacts of potential privacy constraints on a UMC. Privacy constraints that apply to a user can be privacy laws and regulations that are in effect, as well as the user's own personal privacy preferences. Those privacy constraints are expressed in name-value pairs and used as bindings for the Boolean guards associated with each UMC. If the Boolean guard is resolved to be true, then the associated UMC will be selected in the resulting personalization architecture; otherwise, the UMC will not be included.

The syntax for constructing the Boolean expressions and Boolean guards follows the Backus-Naur Form (BNF) (Knuth, 1964), which is a set of context-free grammars to define a formal language. This syntax is defined as follows (Garg et al., 2003):

```
<BooleanGuard> ::= <BooleanExp>

<BooleanExp> ::= <And> | <Or> | <Not> |

    <GreaterThan>

<GreaterThanOrEquals> |

    <LessThan> | <LessThanOrEquals> | <Equals> |

    <NotEquals> | <InSet> | <InRange> |

    <Bool> | <Paren>

<And> ::= <BooleanExp> && <BooleanExp>

<Or> ::= <BooleanExp> || <BooleanExp>

<Not> ::= !<BooleanExp>

<GreaterThan> ::= <LeftOperand> > <RightOperand>

<GreaterThanOrEquals> ::= <LeftOperand> >= <RightOperand>

<LessThan> ::= <LeftOperand> < <RightOperand>

<LessThanOrEquals> ::= <LeftOperand> <= <RightOperand>

<Equals> ::= <LeftOperand> == <RightOperand>

<NotEquals> ::= <LeftOperand> != <RightOperand>

<InSet> ::= <LeftOperand> @{ <Set> }

<InRange> ::= <LeftOperand> @[ <RightOperand>, <RightOperand> ]

<Paren> ::= ( <BooleanExp> )

<Set> ::= <RightOperand> | <RightOperand>, <Set>

<LeftOperand> ::= Variable

<RightOperand> ::= Variable | Value

<Bool> ::= true | false
```

For our purposes, the variables, Boolean expressions and Boolean guards all have privacy semantics. For example, assume that we have a UMS that serves users from Germany and the United States of America (USA). This UMS contains a UMC that employs one-time machining learning combined with a clustering technique to generate personalized music recommendations for a user. This UMC analyzes both the user's browsing history over several sessions in a personalized online music store and her demographic data such as gender, address and occupation. Upon our analysis of potential privacy constraints, we recognize two conditions for German citizens according to the German Telemedia Law (DE-TML, 2007):

1. Keeping multiple session logs are prohibited, without consent.
2. Combining user profiles retrievable under pseudonyms with data relating to the bearer of the pseudonym.

We also realize that there is no applicable privacy laws in the USA for this scenario. Despite tracking users online is not explicitly prohibited by laws in this scenario, existing literature has shown that Internet users do concern about being tracked online. Therefore, we create a variable to capture this potential privacy constraint (users' personal privacy preference, in this case). To model the potential privacy constraints in this scenario, we define the following variables and their possible values:

```
Country = Germany | USA
SessionLogPref = single | multiple | unspecified
TrackingPref = allow | disallow | unspecified
CombineProfilePref = allow | disallow | unspecified
```

A user can explicitly specify their privacy preferences (possibly indicating consent) for a set of pre-defined options, or leave these preferences unspecified. If preferences are not explicitly stated, the system chooses features that provide the most information to the system and are legally allowed by the user's country/location. The resulting Boolean guard and its constitute Boolean expressions read as follows:

```
  (SessionLogPref == ‘‘multiple’’
    || (SessionLogPref == ‘‘unspecified’’
   && Country != ‘‘Germany’’))
&&
  (CombinePorfilePref == ‘‘allow’’
    || (CombinePorfilePref == ‘‘unspecified’’
     && Country != ‘‘Germany’’))
&&
  (TrackingPref != ‘‘disallow’’)
```

For this simple example, the above Boolean guard is reasonably manageable. But, what if now the system has to serve users from dozens of countries, how difficult would it be to update the Boolean guard? We will discuss issues with the Boolean guard approach, present an alternative approach and compare both approaches in Section 4.3.

Suppose we have a German user A who did not specify any personal privacy preferences. Nevertheless, if A is in principle identifiable, the German Telemedia Law (DE-TML, 2007) would apply. Therefore, the bindings for this users are:

```
Country = ‘‘Germany’’
SessionLogPref = ‘‘unspecified’’
TrackingPref = ‘‘unspecified’’
CombineProfilePref = ‘‘unspecified’’
```

Applying these bindings to evaluate the above Boolean guard results false, therefore the UMC will not be selected and used for this user.

Let us assume that we have two American users B and C. User B does not have any personal privacy preferences, while user C expresses that he or she does not want to be tracked online. The bindings for user B are:

```
Country = ‘‘USA’’
SessionLogPref = ‘‘unspecified’’
TrackingPref = ‘‘unspecified’’
CombineProfilePref = ‘‘unspecified’’
```

While the bindings for user C are:

```
Country = ‘‘USA’’
SessionLogPref = ‘‘unspecified’’
TrackingPref = ‘‘disallow’’
CombineProfilePref = ‘‘unspecified’’
```

Applying user B’s and user C’s bindings to evaluate the above Boolean guard results true and false, respectively. Therefore, the UMC will be selected and used for user B but not user C. This example illustrates that both users’ personal privacy preferences and their countries/locations matter in this process. Two users from the same country/location can get different UMC pool due to the differences in their individual privacy preferences.

### 4.1.5 Dynamic selection process

The Selector monitors the start and end of user sessions via bind and unbind operations onto the UMS by the external applications. When the Selector detects the start of a user session, it initiates a *Privacy Context Detection* process that will collect all the active privacy constraints and then generate corresponding bindings. A similar process will be carried out whenever during a user session the Selector learns about new or changed privacy requirements (which for all practical purposes will stem from user preferences since privacy laws and regulations are unlikely to change during a session).

The bindings are then fed into the Selector that will carry out a *PLA selection process* (Garg et al., 2003). Firstly, the Boolean guards of all UMCs are evaluated based on the given bindings, to determine whether or not these UMCs may be included in the personalization architecture for the current user session. Secondly, a binary Privacy Constraint Satisfaction (PCS) vector is constructed whose n$^{th}$ element represents whether or not the n$^{th}$ UMC may be used. The Selector checks whether a run-time system instance with such a PCS already exists. If so, the Selector will assign the user session to the existing run-time system instance that has the same PCS. If not, the Selector will perform *PLA Pruning* (Garg et al., 2003) that automatically removes any disallowed components from the architecture, and then the Selector instantiates a new run-time system instance for the user session. Figure 4.2 presents the pseudo-code of the above process.

```
The Selector monitors the start and end of user sessions:
    On bind (start):
        Privacy Context Detection:
            Collect active privacy constraints;
            Generate variable bindings;
        PLA selection, based on bindings:
            Evaluate Boolean guards for UMCs;
            Construct a new PCS vector V;
        IF there already exists an identical PCS THEN
            Assign the user session to the existing
            run-time system instance, say $instance_i$;
            $instance_i$ . numSessions ++;
        ELSE
            PLA Pruning:
                Prune out UMCs whose Boolean
                guards are resolved to FALSE;
            Instantiate a new run-time system instance
            for the user session, say $instance_{n+1}$);
            $instance_{n+1}$ . numSessions = 1;
    On unbind (end):
        $numSessions_{current}$ - -;

If new/changed user privacy preferences are detected, a
similar process starts as on bind.
```

Figure 4.2: Dynamic selection process

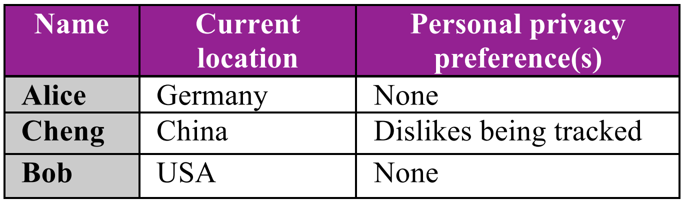
### 4.1.6 An illustrative example

We now describe a concrete example of our user modeling framework. We will use this example to illustrate the technical details of how our framework provides privacy-enhanced personalization services.

#### The example scenario

Let us assume that UniversalFriends.com is a website that is operated in the USA by a signatory of the U.S. Network Advertisers Initiative (NAI) (NAI, 2006). The goal of this website is to bridge physical distances between people and to foster world-wide friendships through information technology. It provides personalized services to help customers make friends worldwide. Upon registration, users will be asked to choose a pseudonymous user ID along with a password and to disclose some information about themselves (e.g., their hobbies). They will be given some space on the UniversalFriends web server to create their own homepages. The system will recommend a personalized list of likely friends based on a user's characteristics, and will automatically send invitations for pair-wise virtual meetings.

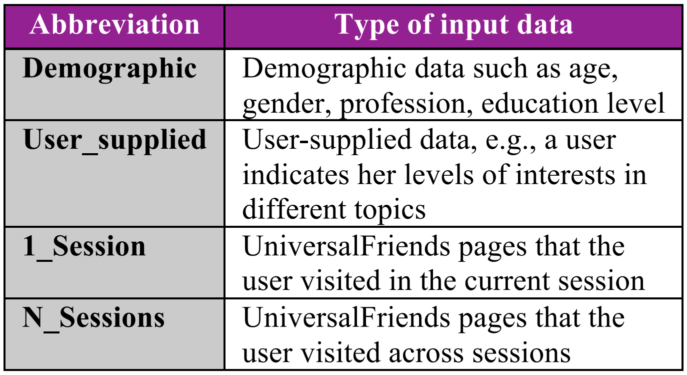
We have three hypothetical users, Alice, Cheng and Bob. Figure 4.3 describes their characteristics.



| Name | Current location | Personal privacy preference(s) |
|---|---|---|
| **Alice** | Germany | None |
| **Cheng** | China | Dislikes being tracked |
| **Bob** | USA | None |

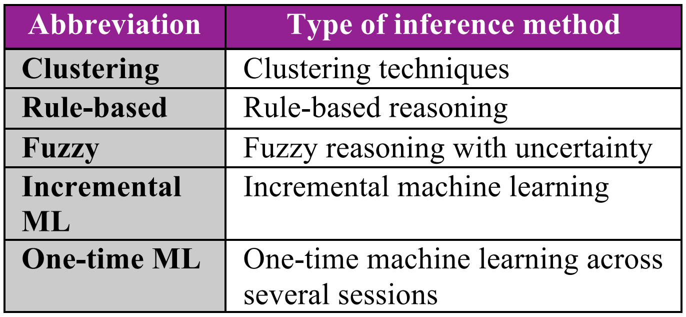
Figure 4.3: Our hypothetical users

The UniversalFriends web server relies on our privacy-enabling personalization infrastructure to infer information about users in order to recommend potential friends. Figure 4.4 and Figure 4.5 show the types of input data and the available inference methods, respectively. Figure 4.6 summarizes the usage of data and inference methods for each user modeling component.



| Abbreviation | Type of input data |
|---|---|
| **Demographic** | Demographic data such as age, gender, profession, education level |
| **User_supplied** | User-supplied data, e.g., a user indicates her levels of interests in different topics |
| **1_Session** | UniversalFriends pages that the user visited in the current session |
| **N_Sessions** | UniversalFriends pages that the user visited across sessions |

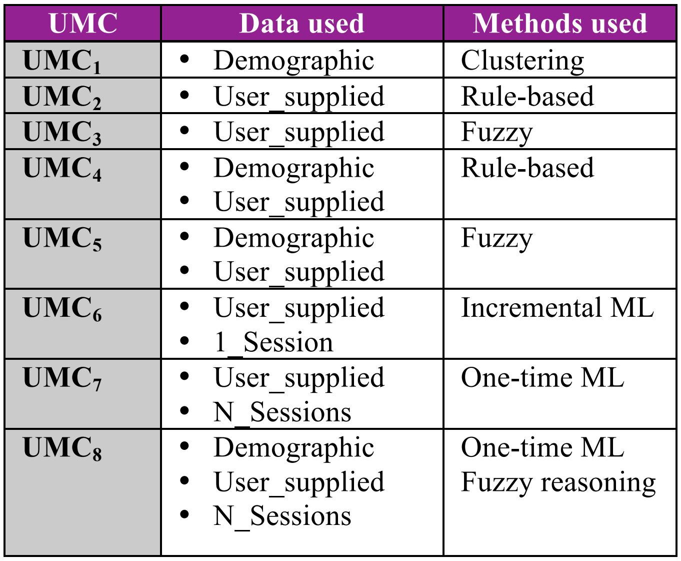
Figure 4.4: Types of input data



| Abbreviation | Type of inference method |
|---|---|
| **Clustering** | Clustering techniques |
| **Rule-based** | Rule-based reasoning |
| **Fuzzy** | Fuzzy reasoning with uncertainty |
| **Incremental ML** | Incremental machine learning |
| **One-time ML** | One-time machine learning across several sessions |

Figure 4.5: Types of inference methods

For example, $\text{UMC}_1$ can recommend people in the same profession cluster. If a user indicates a high interest in a specific topic, $\text{UMC}_2$ can infer that she would like to meet people with similar ratings for a topic; alternatively in this case, $\text{UMC}_3$ can infer with 95% confidence that she would like to meet people with similar ratings for the topic.



| UMC | Data used | Methods used |
|---|---|---|
| **$UMC_1$** | • Demographic | Clustering |
| **$UMC_2$** | • User_supplied | Rule-based |
| **$UMC_3$** | • User_supplied | Fuzzy |
| **$UMC_4$** | • Demographic<br>• User_supplied | Rule-based |
| **$UMC_5$** | • Demographic<br>• User_supplied | Fuzzy |
| **$UMC_6$** | • User_supplied<br>• 1_Session | Incremental ML |
| **$UMC_7$** | • User_supplied<br>• N_Sessions | One-time ML |
| **$UMC_8$** | • Demographic<br>• User_supplied<br>• N_Sessions | One-time ML<br>Fuzzy reasoning |

Figure 4.6: UMC pool

**Interaction with the personalized system**

Users can interact with the system as follows:

1. Users log into UniversalFriends.com using their registered user names and passwords.

2. The website gathers users' current privacy constraints including those imposed by privacy laws and regulations, and their privacy preferences. Users can specify their privacy preferences and change them anytime during the interaction with the personalized system. For instance, if they feel that a specific piece of privacy law or regulation is too strict to get otherwise much better personalization, they can give their consent to certain system actions that are otherwise legally prohibited (e.g., the storage of personal data across sessions).

3. For every user, a summary webpage shows:

   (a) their prevailing privacy constraints, and

(b) the selected UMCs used in producing the personalized service, and the excluded UMCs and the reasons for their exclusion (i.e., the specific privacy constraints).

**Privacy-enabling personalization process**

The privacy constraints that apply to each of the three individual users and their implications for the UMCs are discussed below (the corresponding variables, Boolean expressions and Boolean guards can be defined as described in Sections 4.1.4):

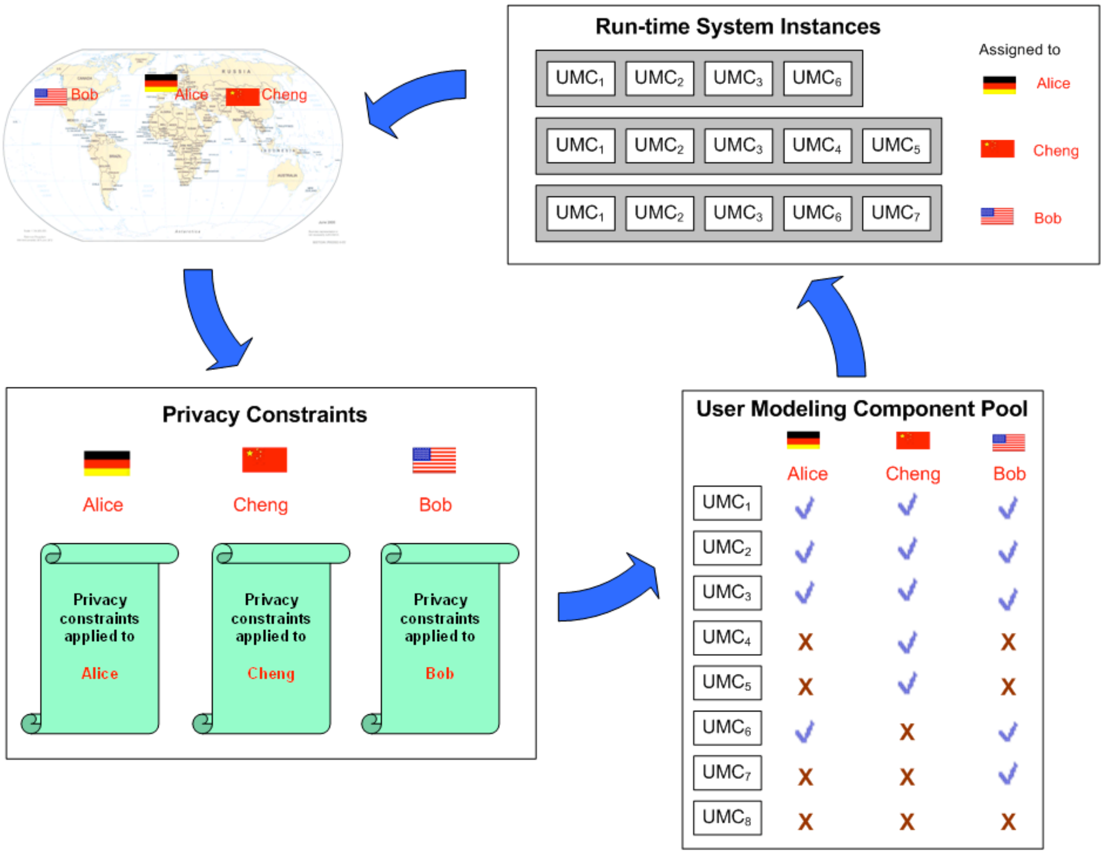
For Alice, the German Telemedia Act applies, with the following consequences:

- $\text{UMC}_4$, $\text{UMC}_5$ and $\text{UMC}_8$ are illegal because the law prohibits combining user profiles retrievable under pseudonyms with data relating to the bearer of the pseudonym.
- $\text{UMC}_7$ and $\text{UMC}_8$ are illegal because the law mandates personal data to be erased immediately after each session except for very limited purposes.

Therefore, $\text{UMC}_4$, $\text{UMC}_5$, $\text{UMC}_7$ and $\text{UMC}_8$ cannot be used for Alice without her explicit consent.

While no privacy law applies to Cheng, she has her own personal privacy preference, such as that she “dislikes being tracked”. Hence $\text{UMC}_{6,}$$\text{UMC}_7$ and $\text{UMC}_8$ cannot be used because the system may not keep track of the pages she visits on UniversalFriends.com.

For Bob from the United States, $\text{UMC}_4$, $\text{UMC}_5$ and $\text{UMC}_8$ cannot be used according to the NAI self-regulation if he does not consent to merging non-personally identifiable usage data with personally identifiable demographic data.


Run-time System Instances
Assigned to
UMC$_1$ UMC$_2$ UMC$_3$ UMC$_6$
Alice
UMC$_1$ UMC$_2$ UMC$_3$ UMC$_4$ UMC$_5$
Cheng
UMC$_1$ UMC$_2$ UMC$_3$ UMC$_6$ UMC$_7$
Bob
Bob
Alice
Cheng
Privacy Constraints
Alice
Cheng
Bob
Privacy constraints applied to Alice
Privacy constraints applied to Cheng
Privacy constraints applied to Bob
User Modeling Component Pool

|  | Alice | Cheng | Bob |
|---|---|---|---|
| UMC$_1$ | ✓ | ✓ | ✓ |
| UMC$_2$ | ✓ | ✓ | ✓ |
| UMC$_3$ | ✓ | ✓ | ✓ |
| UMC$_4$ | X | ✓ | X |
| UMC$_5$ | X | ✓ | X |
| UMC$_6$ | ✓ | X | ✓ |
| UMC$_7$ | X | X | ✓ |
| UMC$_8$ | X | X | X |

Figure 4.7: Privacy-enhanced personalization process

4.7 illustrates the process of selecting and instantiating personalization architectures for each user according to their individual privacy constraints (as we explained in Section 4.1.4). Note that, in this case, three different architectural instances are created since each user has different privacy constraints.

### Prototype implementation

The prototype system is composed of three basic components: a Context Detector, an Instance Manager, and a light version of ArchStudio (ArchStudio, 2005). To simplify matters, we did not yet include a Directory Component. Figure 4.8 gives a high-level overview of the system structure.

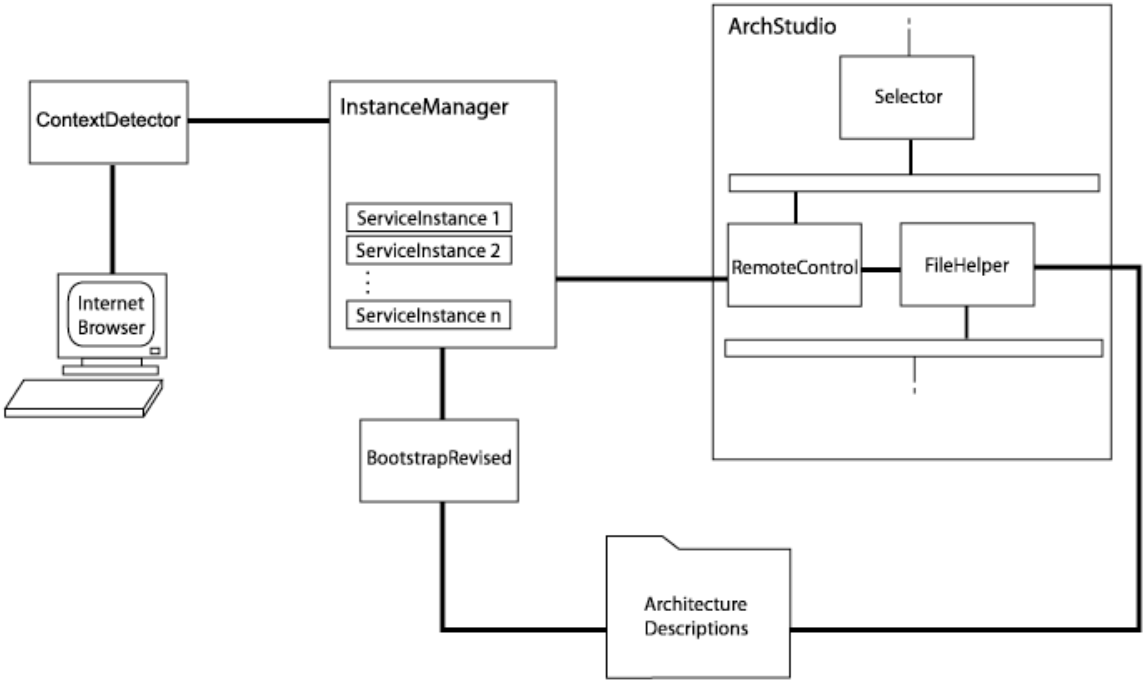


Figure 4.8: System architecture

The Context Detector is the component that interfaces with a user's web browser, collecting her privacy constraints and relaying them to the next component, the Instance Manager. The ArchStudio component is mainly used for its Selector, which generates the architecture descriptions (expressed in xADL 2.0 (Dashofy et al., 2005) and selected from an overall PLA description) for the personalization architectures, or "personalized system instances" tailored to each individual user based on their privacy constraints. The Instance Manager is the central core of the system. It responds to the requests of the Context Detector and uses ArchStudio to build the personalized system instances.

All three main components of the system are implemented in Java and communicate via the Java Remote Method Invocation (RMI) framework. Using this method, it is possible for the components to be distributed across more than one machine, but this is currently not the case. The Instance Manager and the RemoteControl subcomponent of ArchStudio extend the remote interface and sign their names to the RMI registry, allowing the Context Detector and ArchStudio to access the Instance Manager directly, as well as allowing the

Instance Manager to invoke ArchStudio's Selector functionality.

Minor miscellaneous components of the system include BootstrapRevised (a modified version of the Bootstrapper from the orginal ArchStudio), which the Instance Manager uses to initialize architecture descriptions into running instances. In Figure 4.8, the stored architecture descriptions produced by ArchStudio are simply represented as a file directory located on the server machine. Each UMC has its implementation stored as a JAR file in the server machine. The BootstrapRevised component will in turn evoke the JAR files for those UMCs that are included in the final architecture description resulted from PLA selection and pruning. The web pages produced by the Context Detector are served via Apache Tomcat servlets, which are also able to make requests of the Instance Manager directly once a user's system instance is produced.

When a user first interacts with the system using her web browser, she will be prompted by the Context Detector for her privacy constraints. When submitted, these constraints are transferred to the Instance Manager as a new user request. They are packaged by the Instance Manager and posted to the ArchStudio Component for selection processing. Then a customized architecture is selected and its description saved to a file. The Instance Manager, thereupon, receives a request to instantiate the newly completed architecture. It first analyzes the new architecture to construct a PCS Vector describing which UMCs are included in the description. This PCS Vector is compared with those of the currently running instances. If one of them matches, then no new instantiation takes place but rather the found instance is used. If no running service instance matches the new architecture description, BootstrapRevised is invoked to turn the architecture into a running service instance. This new service instance is assigned to the user, who may now access its functionality via requests to the Instance Manager. If the user's privacy constraint information changes later on, the process may be restarted to consider the new constraints.

### 4.1.7 Discussion

Our approach uses the concept of PLA to model the variability that exists in a UMS, and to dynamically select architectural instances of the UMS to cater to the specific needs of a particular user. The approach, thus, considers the privacy constraints that apply to an individual user at a given time and dynamically selects and instantiates a personalization architecture that provides personalized services to this specific user. The result is a flexible approach that not only helps address the complexity of building personalized systems, but also strongly supports their evolution: as new privacy and personalization concerns arise, they can be modularly added to the product line architecture.

However, there are still two major issues with the current incarnation of this approach. The first issue is the system performance of the PLA-based UMS since dynamic architectural reconfiguration during runtime is usually resource-intensive. In Section 4.2, we will discuss three performance-enhancing mechanisms: light-weight representation of PLA, computation distribution, and caching. Another issue emerges from our initial experience of using a Boolean guard language to express the privacy constraints and their impacts on the UMS. In particular, we find it can be difficult to update the Boolean guards when privacy constraints change. In Section 4.3, we will revisit this modeling issue, introduce an alternative approach, and compare these two approaches.

## 4.2 Distributed Framework

In order to cope with the challenge of resource-intensive dynamic run-time architectural reconfigurations caused by the PLA-based user modeling framework tailoring to each user's privacy constraints, we explore the ideas of computation distribution, caching and light-weight PLA representation.

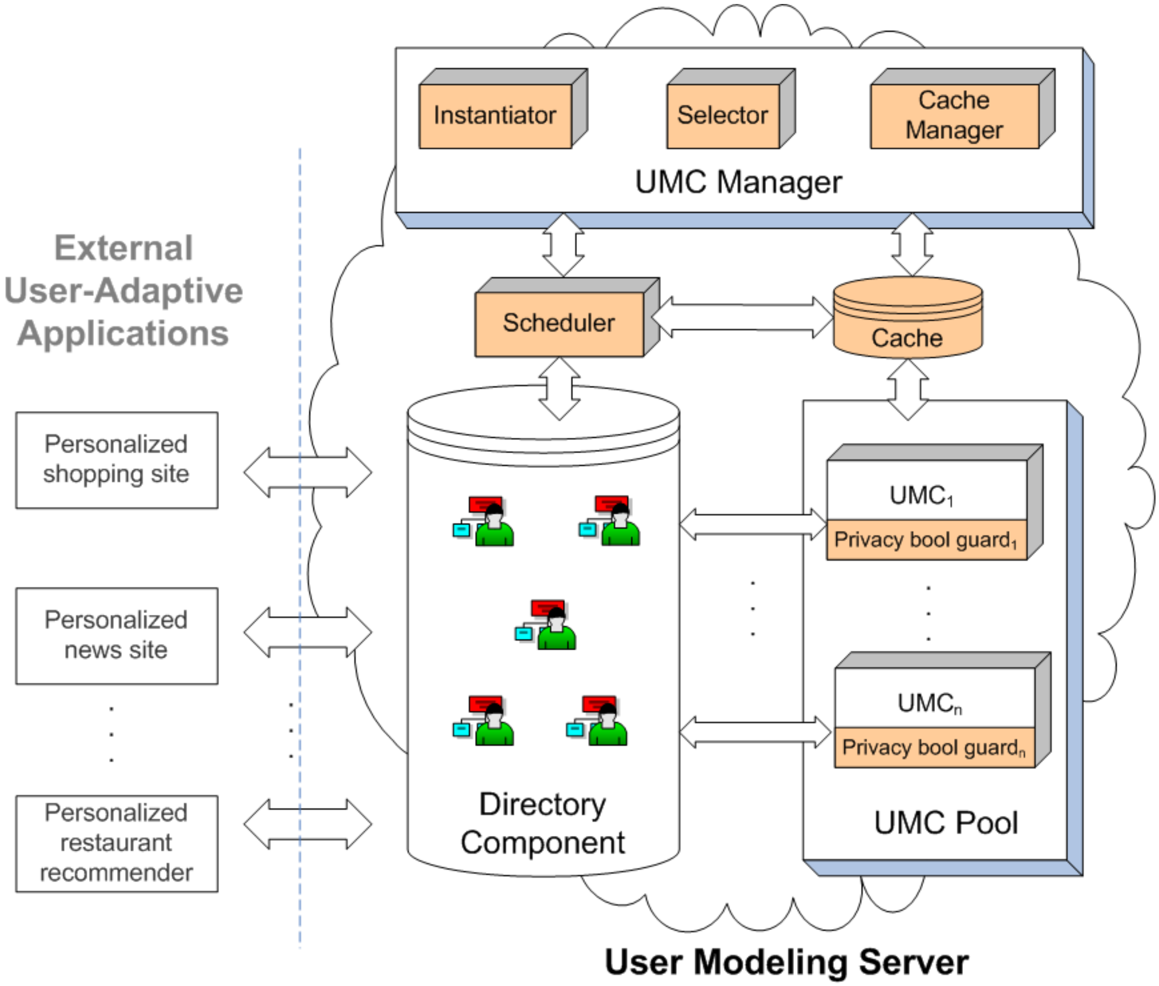


Figure 4.9: Distributed privacy-enhanced user modeling framework

### 4.2.1 Overview

Figure 4.9 shows an overview of our distributed framework[1]. Compared with our framework (refer to as the baseline framework thereon) introduced in Section 4.1, it now adds a Scheduler and a modified UMC manager.

To briefly recap how the baseline framework works (details in Section 4.1). External user-adaptive applications 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

[1] The shaded parts are our privacy-related additions to the user modeling server described in (Kobsa and Fink, 2006).

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., collaborative filtering). UMCs make inferences about users based on existing information in the user models and then add the derived user information to the user models (Wang and Kobsa, 2007).

To enable PLA operations (e.g., product architecture selection), 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 were core components, and UMCs were 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 (e.g., Combining_Profile == true). As such, we use these Boolean variables bearing privacy semantics to represent users' privacy preferences as well as applicable privacy regulations. In practice, the values of these Boolean variables can come from the evaluation of privacy conditions expressed in a privacy policy language (see 3.2.1 for a discussion of these languages).

In the following, we will describe the UMC Manager in more detail and then discuss distribution issues.

### 4.2.2 UMC Manager

The UMC Manager was implemented to support PLA selection and instantiation as well as our caching mechanism. It 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-value pairs for the Boolean guards in the PLA, e.g., Combining_Profile = false which would represent that the user or some privacy norm relating to the user session disallow the merging of profiles relating to the same user. The Selector selects a particular product architecture out of 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., 2006b) whose n$^{th}$ element represents whether or not the n$^{th}$ UMC may be included in the selected product architecture.

**Instantiator.** The Instantiator takes a PCS 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 may be created.

**Cache Manager.** We designed a multi-level caching strategy that is shown in Figure 4.10. The Cache Manager controls caches of both individual users' privacy bindings and their associate PCS vectors (i.e., the results of the PLA selection). More specifically, when a new user session starts, the Cache Manager checks the privacy binding cache whether the system has an existing user session with the same privacy bindings (i.e., a user with identical privacy norms and individual privacy preferences). If it finds one, the new session will be assigned to the same system instance as the existing session. If no such binding can be found, the Cache Manager will further check the PCS cache since a PCS may meet the constraints of more than one privacy binding. Only if no such PCS can be found either, the Instantiator will start a new instance for this user session.

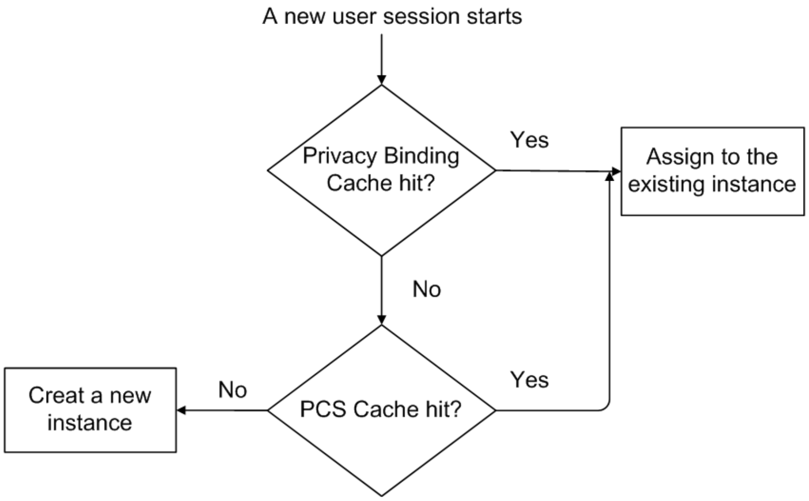

A new user session starts
Privacy Binding Cache hit?
Yes
Assign to the existing instance
No
PCS Cache hit?
Yes
No
Creat a new instance

Figure 4.10: Multi-level caching mechanism

### 4.2.3 Distribution

In order to cope with potentially millions of concurrent users, the enhanced UMS needs to be distributed. In Figure 4.9, the cloud denotes the distribution of processing over a network of machines. Distribution of the LDAP-based Directory Component and the UMC Pool have been addressed in (Kobsa and Fink, 2006). We 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 in the framework to assign incoming user sessions to various hosts, and a database to store the privacy binding cache and the PCS cache.

### 4.2.4 Implementation

In this section, we describe the implementations of major components and operations in our distributed framework (the first two were varied in the different conditions of our performance evaluation described in Section 5.1).

### PLA Representation, Selection and Instantiation

As explained above, our privacy-enhancing user modeling framework was designed as a PLA. Therefore, the core of the framework involves the following tasks: generation of a PLA for the system architecture, selection of UMCs based on the bindings of the privacy Boolean guards, and instantiation of the selected architecture for the user modeling system.

*ArchStudio-based Implementation.*

In our preliminary implementation (Wang et al., 2006b), we adapted functionalities from ArchStudio 3 (ArchStudio, 2005) to perform the above tasks. ArchStudio 3 is an architecture-centric development environment, built on the C2 architectural style (Taylor, 1996). It provides excellent support for PLA modeling and development. This system has been meanwhile upgraded to ArchStudio 4 (Dashofy et al., 2007), built on the Myx architectural style (ArchStudio, 2008). The Myx style provides better system performance because it allows unmediated synchronous procedure calls between components in the architecture. In the C2 style, component interactions are always asynchronous and mediated by connectors. We therefore chose ArchStudio 4 for our final test system and implemented it in the Myx style (we call it the Myx version).

*Our Customized Implementation.*

The standardization and extensibility of the XML-based PLA representation come at a price: XML processing can be expensive and thus affect the overall system performance. This is especially the case when the PLA has a large number of components. Therefore, we designed a light-weight alternative to the xADL 2.0 representation, called PLA Object Notation (PLAON)[2]. It contains an array of component objects. Each optional component object stores its privacy Boolean guard in an array, each element representing a privacy Boolean variable. Privacy bindings are in turn stored as a binary array, each element denot-

[2] This is inspired by the idea of JSON (www.json.org) as a light-weight alternative of XML for web programming.

ing the binding for a privacy Boolean variable. Our customized selector can then use the privacy binding array to resolve the Boolean guard array. Again the results of the selection will be a PCS vector, implemented as a binary array. Our customized instantiator reads from the PCS array to start components whose values in the PCS array are 1. Since our customized implementation represents the PLA semantics in a succinct object notation and omits any XML processing, we expect it to perform better than the original Myx-based implementation.

### Multi-Level Caching

Caching is the other factor that we vary in our performance evaluation (see Section 5.1 for details). As described earlier, if two users have the same privacy bindings, or the same PCS vectors after selection, then they can share the same user modeling system instance. This reuse would save the system from performing unnecessary architectural selections and instantiations in such cases.

### 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 would select the lightest-loaded host that can still handle this session with its resources. This resource-aware scheduling was used in all conditions of our experiment.

## 4.3 Revisit modeling privacy constraints and their impacts

In Section 4.1.4, we describe using Boolean guards to express privacy constraints and their impacts on the PLA-based personalized system. In this section, we will revisit this topic by discussing issues of the Boolean guard approach, presenting an alternative approach using change sets and relationships (Hendrickson and van der Hoek, 2007; Wang et al., 2009) and comparing the two approaches.

The analysis and discussion in this section rely on concepts from configuration management, which we describe next.

### 4.3.1 Configuration management

The discipline of *configuration management* (CM) has been primarily concerned with capturing the evolution of a software system at the source code level (Estublier et al., 2005). For this, it has extensive and detailed mechanisms and procedures for storing multiple versions of code and allowing multiple developers parallel access to that code (Conradi and Westfechtel, 1998). Automated conflict detection and merge routines help in reconciling overlapping changes that may arise as a result of parallel development (Mens, 2002).

Of interest to this paper are the concepts of extensional and intensional versioning (Conradi and Westfechtel, 1998). In *extensional versioning*, the configuration management system focuses on managing versions of artifacts that result after making changes. Typically, a version graph is used to relate different versions of an artifact; developers retrieve a particular version, modify it, and then add the new version to the graph when complete.

In contrast, *intensional versioning* makes changes a first class entity, inverting the relationship between versions and changes (Conradi and Westfechtel, 1998). Instead of ensuring that each version is uniquely stored and accessible, intensional versioning stores each change as a change set (a "delta") independently from the other changes. So, instead of requesting a version of an artifact, developers retrieve an artifact by requesting a series of change sets from which a "version" is constructed. Similarly, after modification of this "version," the delta between this new and the original version is stored as an individually-identifiable change set. This has the advantage that new incarnations of an artifact can be composed by mixing and matching different change sets.

### 4.3.2 Representing PLA variations

As discussed in Section 4.1.2, variations are at the core of PLAs. In our PLA-based user modeling framework, each UMC is associated with a Boolean guard which reflects the impact of privacy constraints on this UMC. These UMCs are where the variations lie in the PLA.

There are multiple ways of representing variations in PLAs. Currently, the approaches towards modeling PLAs are predominantly extensional, i.e., they model a single, monolithic architecture that simultaneously represents all possible products using variation points and guards of some form, e.g., Menage (Garg et al., 2003) uses Boolean guards. 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. A product instance can be selected out of a product line architecture by resolving the Boolean guards of each variation point at design-time, invocation-time or run-time (van der Hoek, 2004). "While extensional approaches adequately model PLA variation, they suffer from a sizable mismatch between conceptual variability (i.e., the features through which architects logically view and interpret product differences) and actual variability (i.e., the modeling constructs through which the logical differences must be expressed). As a result, the actual model exhibits a high degree of redundancy, scattering and tangling of the conceptual model it represents making it difficult to interpret and modify" (Hendrickson and van der Hoek, 2007).

Alternatively, *intensional* (Conradi and Westfechtel, 1998) approaches are gaining ground, e.g., (Bell Labs Lucent Technologies, 1997; Batory, 2005; Cottenier et al., 2006; Hendrickson and van der Hoek, 2007). With intensional approaches, product architectures are *composed* from different modeling constructs that represent features at some level. Hendrickson and van der Hoek (2007) presented an intensional approach where an architect composes product architectures from a collection of change sets and is guided by constraints expressed as relationships. Together, change sets and relationships form the basis for modeling features and feature models. They also found that conceptual variability is better expressed using the modeling constructs of change sets and relationships than the modeling constructs of an extensional approach.

### 4.3.3 A motivating example

We now introduce a motivating example to ground our analysis and discussion. Consider an online movie recommender system called MyMovie[3]. To make personalized recommendations, MyMovie can utilize three features: 1) *cross-site tracking*, to observe what other websites a user visits, 2) *single session logs*, which provide information about the pages a user has visited on our website during her current session, and 3) *multiple session logs*, which additionally provide information about the pages a user has visited on our website in the past, during previous sessions. The first feature is optional while the singe and multiple

[3] Inspired by the MovieLens system (www.movielens.org) and the Netflix system (www.netflix.com)

session log features are alternatives, meaning that only one may be selected. We refer to these features and the feature model that binds them, collectively as the *software model*.

Being a privacy-enhancing personalization (PEP) system, MyMovie also has a *privacy model*, which defines the various privacy constraints placed on the system accoding to the users stated preferences and country/location. We have three hypothetical users, Alice, Bob and Chris from Germany, the United Kingdom (UK) and the United States (US), respectively. The privacy model must ensure that the product obtained from the software model adheres to the individual laws of each country.

Initially, our system's privacy model recognizes two conditions for German citizens[4]:

- Cross-cite tracking is prohibited, without consent.
- Multiple session logs are prohibited, without consent.

Upon navigating to our website, a user can explicitly specify their privacy preferences (possibly indicating consent) for a set of pre-defined options, or leave these preferences unspecified. Our system uses these preferences along with the user's location to dynamically generate a personalized system adhering to the user's specific set of privacy constraints. If preferences are not explicitly stated, the system chooses features that provide the most information to the system and are legally allowed by the user's location. Thus, German citizens without explicit preferences are provided a system with the single session logs feature while citizens from the United Kingdom and the United States without explicit preferences are provided systems with the cross-site tracking and multiple session logs features. Preferences, of course, if stated, override these defaults.

[4] Based on the German Telemedia Law (DE-TML, 2007)

### 4.3.4 Extensional and intensional modeling

In this section, we model our example PLA both extensionally and intensionally, noting intricaces of both approaches. We conclude this section with a reflection on the two modeling approaches.

#### Modeling extensionally

Taking an extensional modeling approach, we use variation points to denote the places where variabilities occur in the PLA. The resulting *software model* for our system is shown in Figure 4.11.

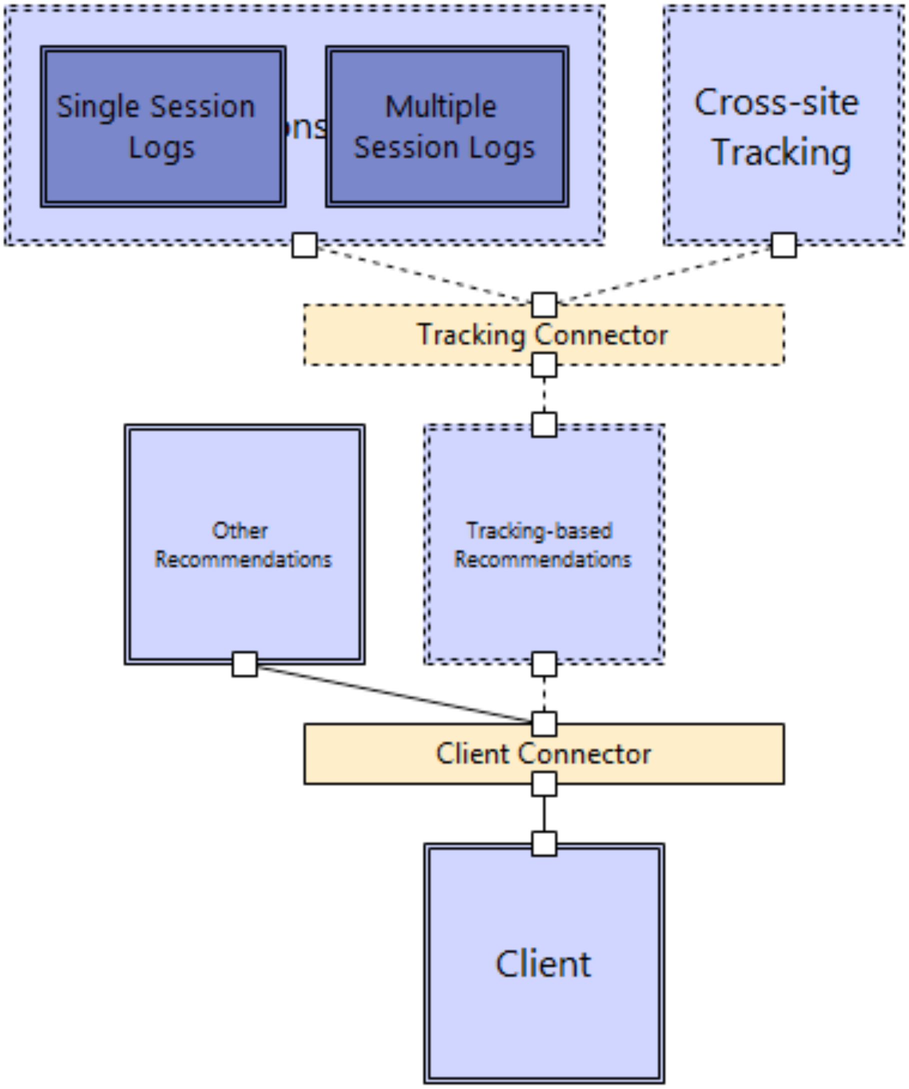


Figure 4.11: Extensional Model

In this figure, core elements belonging to all product architectures are shown as solid boxes or lines, optional elements as dashed boxes or lines, and variant elements as large boxes containing the variants. Additionally, but not shown, each variation point is annotated with a Boolean guard that indicates when that element is to be included in or excluded from a particular product architecture.

To model the privacy model of our system, we define the following variables and their possible values:

```
Country = Germany | UK | USA
SessionLogPref
  = single | multiple | unspecified
CrossSiteTrackingPref
  = allow | disallow | unspecified
```

These variables are used in the Boolean guard expressions of each variation point. For example, the Boolean guard for the "Cross-site Tracking" component (using a Java-like notation) is:

```
CrossSiteTrackingPref == "allow"
||(CrossSiteTrackingPref == "unspecified"
    && Country != "Germany")
```

The above guard states that the architectural element is to be included when the user has explicitly allowed it, or when the user has not stated a preference and the user is in a country other than Germany. A product is selected by evaluating each variation point's Boolean guard against values assigned to each variable, indicating whether the variation point is included in, or excluded from, the desired product.

As we can see, the software model and privacy models immediately become entangled as the privacy model is expressed in terms of the impact of a user's preferences and location on each and every variation point in the system. Furthermore, as features overlap in the extensional approach, guards become more complex. For instance, the "Tracking-based Recommendations" component should be included when any of the three features ("Single Session Logs", "Multiple Session Logs", and "Cross-Site Tracking") are enabled. The resulting guard is:

```
(SessionLogsPref == "single"
 ||(SessionLogsPref == "unspecified"
     && Country == "Germany"))
||(SessionLogsPref == "multiple"
 ||(SessionLogsPref == "unspecified"
     && Country != "Germany"))
||(CrossSiteTrackingPref == "allow"
 ||(CrossSiteTrackingPref == "unspecified"
     && Country != "Germany"))
```

From these example guards, three problems become apparent. First, information is *scattered.* The clause in the first example is the condition for selecting the cross-site tracking feature. However, the same clause is repeated at the end of the second example. For an architect to determine all elements affected by that feature, he must examine each and every guard throughout the architecture. Second, information is *tangled.* The second example contains three main clauses, each capturing the condition for selecting the "Single session logs", "Multiple session logs", and "Cross-site Tracking" features, respectively. To modify this guard, an architect must mentally extract these concepts, modify them, and then recombine them to update the guard's expression. Finally, information is *redundantly* expressed. The links and connectors surrounding the "Tracking-based Recommendations" component have the same guard, because they are included or excluded in unison with the component. These factors make interpreting and updating an extensional PLA a tedious and error prone task.

A deployed system using our extensonal PLA example would have a complete copy of the entire PLA as presented above. A new user visiting the website would be prompted for her privacy preferences. Once obtained, these values and the users country would be directly plugged in to the variables used in the PLA guards. These guards would be resolved to obtain a specific system configuration matching the user's privacy constraints, which would then be instantiated for the user. If the user changes her privacy preferences, a potentially different system will then be instantiated for the user.

Modeling the entire system extensionally requires a total of 3 variables: 2 for the users preferences regarding each feature and 1 for the user's country. The model has 10 variation points (the "Session Logs" component and each of its variants have their own Boolean guard) and there are a total of 5 unique Boolean guards in the model.

### Modeling intensionally

Instead of using variation points and guards, the intensional modeling approach uses change sets and relationships to represent the PLA. We chose to use four categories of change sets in our intensional model: feature change sets and relationships which are used to model the software model, preference change sets and country change sets which are used to model the privacy model, and mapping change sets which are used to connect the two models.

The *feature change sets* prefixed by "Feature" are used to model the structrual features of the software system, as shown in Figure 4.12. For each element in a feature change set, an annotation with a "+" means that the element is added by the change set and an "x" means that the element is removed. The presence of a clear, dashed, "ghost" element indicates that the change set references, but does not actually modify, that particular element. For instance, the "Feature: Single Session Logs" change set adds the "Single Session Logs" component, its interface, and a link to an interface on the "Tracking Connector". However, it does not modify the "Tracking Connector" itself.

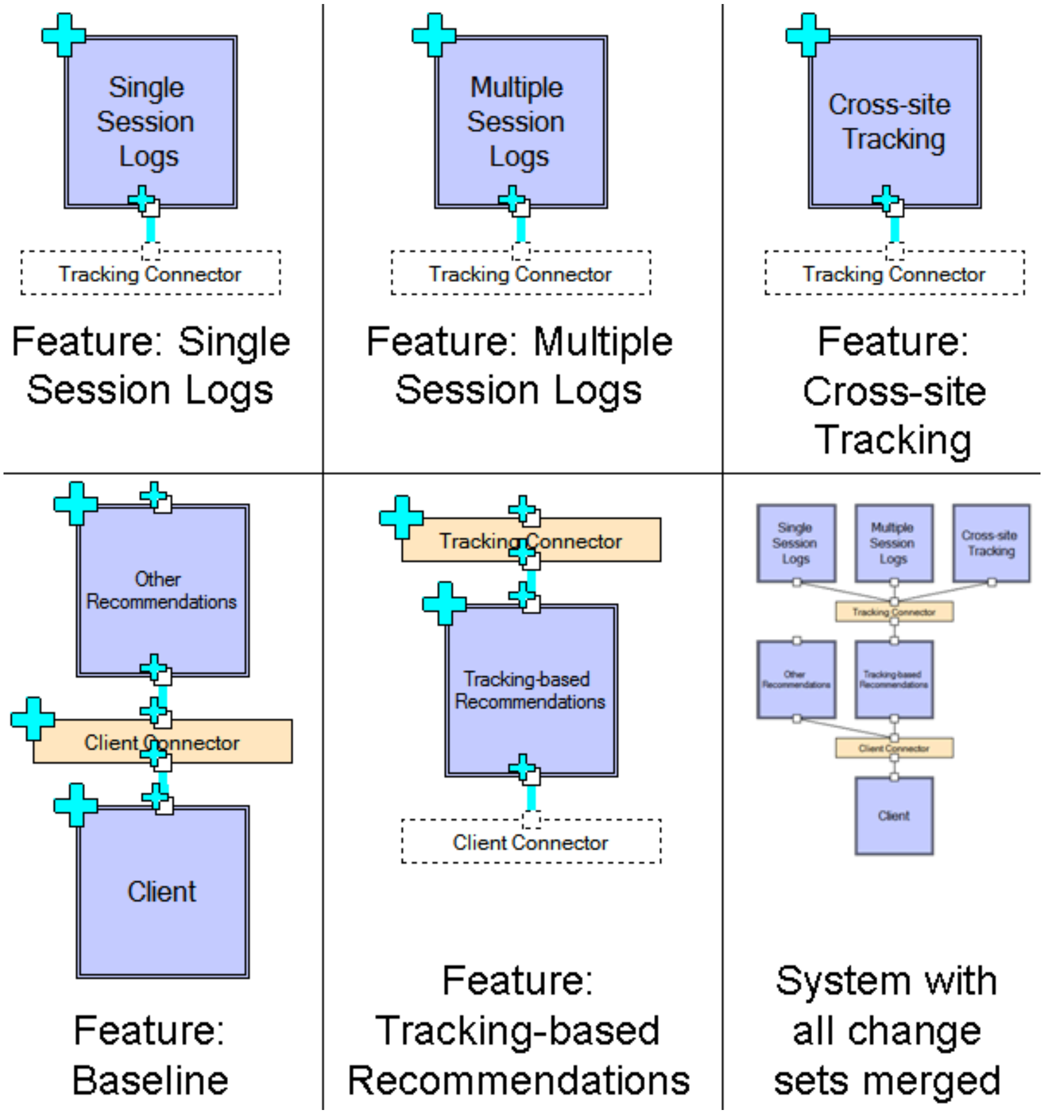
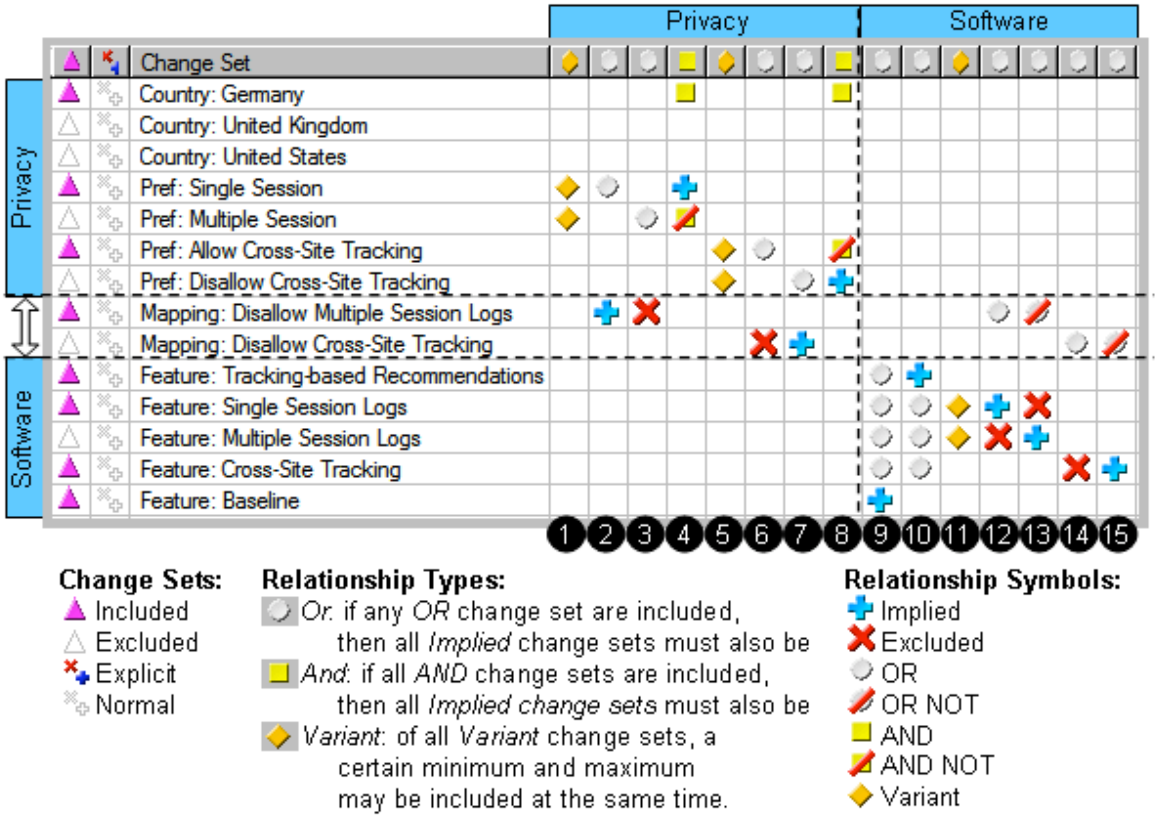


Figure 4.12: Feature Change Sets

These feature change sets are merged together to compose different product architectures. The “Feature: Baseline” change set adds elements common to all products while the three feature change sets on the top row of Figure 4.12 add elements that are unique to their respective features. In the intensional model, areas of feature overlap are expressed as separate change sets that add the elements involved in the feature overlap. The “Feature: Tracking-based Recommendations” change set is such a change set, adding, for instance, the “Tracking-based Recommendations” component needed by all three features. A relationship is used to ensure that this change set is included whenever any of the feature change sets are included (discussed later).

The entire collection of change sets and relationships is shown in Figure 4.13. Each row represents a change set and each column to the right of the “Change Set” column represents a relationship (numbered from 1 to 15). Note that change sets and relationships are grouped according to the two domain models from which they originated: change sets and relationships for the software model are in the lower-right quadrant of the spreadsheet and managed solely by software architects, as indicated by the “Software” labels. Change sets and relationships for the privacy model as managed by legal professionals are in the upper-left quadrant of the variability spreadsheet, as indicated by the “Privacy” labels.

Change sets added by legal professionals that are used in the privacy model are *empty*, only serving as “place holders” or “switches” for the privacy-related concepts they wish to model. These privacy-related change sets include the *preference change sets* prefixed by “Pref” that represent user preferences and the *country change sets* prefixed by “Country” that model users’ countries.



|  |  |  |  | Privacy |  |  |  |  |  |  |  | Software |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | ▲ | ✖✚ | Change Set | ◆ | ○ | ○ | 🟨 | ◆ | ○ | ○ | 🟨 | ○ | ○ | ◆ | ○ | ○ | ○ | ○ |
| Privacy | ▲ | ⊹ | Country: Germany |  |  |  | 🟨 |  |  |  | 🟨 |  |  |  |  |  |  |  |
|  | △ | ⊹ | Country: United Kingdom |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | △ | ⊹ | Country: United States |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | ▲ | ⊹ | Pref: Single Session | ◆ | ○ |  | ✚ |  |  |  |  |  |  |  |  |  |  |  |
|  | △ | ⊹ | Pref: Multiple Session | ◆ |  | ○ | ⧄ |  |  |  |  |  |  |  |  |  |  |  |
|  | ▲ | ⊹ | Pref: Allow Cross-Site Tracking |  |  |  |  | ◆ | ○ |  | ⧄ |  |  |  |  |  |  |  |
|  | △ | ⊹ | Pref: Disallow Cross-Site Tracking |  |  |  |  | ◆ |  | ○ | ✚ |  |  |  |  |  |  |  |
| ⇳ | ▲ | ⊹ | Mapping: Disallow Multiple Session Logs |  | ✚ | ✖ |  |  |  |  |  |  |  |  | ○ | ⊘ |  |  |
|  | △ | ⊹ | Mapping: Disallow Cross-Site Tracking |  |  |  |  |  | ✖ | ✚ |  |  |  |  |  |  | ○ | ⊘ |
| Software | ▲ | ⊹ | Feature: Tracking-based Recommendations |  |  |  |  |  |  |  |  | ○ | ✚ |  |  |  |  |  |
|  | ▲ | ⊹ | Feature: Single Session Logs |  |  |  |  |  |  |  |  | ○ | ○ | ◆ | ✚ | ✖ |  |  |
|  | △ | ⊹ | Feature: Multiple Session Logs |  |  |  |  |  |  |  |  | ○ | ○ | ◆ | ✖ | ✚ |  |  |
|  | ▲ | ⊹ | Feature: Cross-Site Tracking |  |  |  |  |  |  |  |  | ○ | ○ |  |  |  | ✖ | ✚ |
|  | ▲ | ⊹ | Feature: Baseline |  |  |  |  |  |  |  |  | ✚ |  |  |  |  |  |  |
|  |  |  |  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |

**Change Sets:**
▲ Included
△ Excluded
✖✚ Explicit
⊹ Normal

**Relationship Types:**
○ *Or*: if any *OR* change set are included,
then all *Implied* change sets must also be
🟨 *And*: if all *AND* change sets are included,
then all *Implied change sets* must also be
◆ *Variant*: of all *Variant* change sets, a
certain minimum and maximum
may be included at the same time.

**Relationship Symbols:**
✚ Implied
✖ Excluded
○ OR
⊘ OR NOT
🟨 AND
⧄ AND NOT
◆ Variant

Figure 4.13: All Change Sets and Relationships

The *mapping change sets* prefixed by “Mapping” server as the common terms of discourse through which legal professionals and their privacy model interact with software architects and their software model. Professionals from each domain must agree on the meaning of these mapping change sets. Legal professionals add relationships to ensure that a selection of a country and preferences imply the appropriate set of mapping change sets, shown in the upper-left quadrant of Figure 4.13. Software architects add relationships to ensure that a selection of mapping change sets imply the appropriate feature change sets, shown in the upper-left quadrant of Figure 4.13.

Several representative relationships in Figure 4.13 are read as follows:

Relationship 4 expresses an internal mapping within the privacy model. In this case, the legal professionals chose to have a country selection indicate a default set of preferences that adhere to each country's privacy concerns. These, of course may be overridden by the users actual preferences. An AND relationship type, this relationship reads: if the "Country: Germany" change set is selected, and the user has not selected the "Pref: Multiple Session" preference change set, then the "Pref: Single Session" preference change set should be selected.

Relationship 2 expresses a mapping from the privacy model to the mapping change sets that are common to both models. An OR relationship type, this relationship ensures that when the "Pref: Single Session" change set is selected, the "Mapping: Disallow Multiple Session Logs" change set is also selected.

Relationship 12 expresses a mapping from the mapping change sets to the software model. An OR relationship type, this relationship ensures that when the "Mapping: Disallow Multiple Session Logs" change set is selected, the "Feature: Single Session Logs" change set is also be selected and that the "Feature: Multiple Session Logs" change set is *not* selected.

Relationship 10 expresses an internal mapping within the software model. It captures the dependence of the three feature change sets of Figure 4.12 on the "Feature: Tracking-based Recommendations" change set that adds elements common to these three features, as discussed earlier. An OR relationship type, this relationship ensures that when any of the feature change sets are selected, the "Feature: Tracking-based Recommendations" change set is also selected.

Note that in Figure 4.13 the upper-right and lower-left quadrants are empty, indicating no cross-over occurs between these two models, other than through the mapping change sets. This is because the two domain models remain independent, interconnected only through the common mapping change sets. As such, professionals in each domain may freely modify the change sets and relationships used to capture the concepts relevant to their respective domains. Interaction is only needed when mapping change sets are added, removed, or their purpose is changed.

Our intensional PLA would be deployed in a similar fashion as the extensional PLA: the system would have a complete copy of the entire PLA as presented above and new users would be prompted for their privacy preferences. Once obtained, the user's preferences and country would be mapped to an initial selection of change sets. Thus, the use of change sets as opposed to variation points would be transparent to the end user – only the legal and software professionals need know its details. The system would then automate the process of selecting mapping change sets and feature change sets based on the initial selection of country and preference change sets, as dictated by the system's relationships. To do so, the system would continually scan through the relationships, selecting implied change sets and unselecting excluded change sets for any relationship that is violated until either one of two conditions occur. First, if a configuration does not violate any relationships, then a complete system matching the user's privacy constraints has been configured, which would then be instantiated for the user. Otherwise, the system will eventually hit a violated variant relationship or revisit a previous selection of change sets, indicating that a system cannot be composed because of an error in the relationships such as a contradiction.

The entire system, modeled intensionally, has a total of 14 change sets: 4 for user preferences, 3 for a user's country, 2 for mappings between the two domains, and 5 for system features. The model includes 15 relationships: 4 for interdependencies between privacy constraints (Relationship 1, 4, 5, 8), 3 for interdependencies between architectural features (Relationship 9, 10, 11), and the remaining 8 for mappings from the privacy domain to the software domain. We note that the initial work of building the intensional model is more involved as it models more domain concepts explicitly, but it enables a greater degree of separation between both domain models and their interactions, making them easier to interpret, evolve and maintain.

### 4.3.5 Evaluation of PLA evolution

In this section, we evaluate both models in terms of three likely scenarios involving the evolution of a privacy-enhancing personalized system: First, a new law is introduced that adds a new requirement not previously modeled in either domain. Second, a software architect chooses to modify their model by refactoring a feature. Third, a law is modified so as to include additional countries, but does not require or prohibit any new system features. Each scenario modifies the systems that result from the previous scenario, building on the previous scenarios' modifications.

**Scenario 1: modifying both models**

This scenario is based on a European Union (EU) Directive on Privacy and Electronic Communications (2002) (EU, 2002) mandating that tracking-based services require anonymization or user's consent. Anonymization of tracking affects both domain models because neither currently captures such a concept. We discuss updating the software model and the privacy model for each approach.

*Scenario 1: extensional model*

To update the extensional PLA, we first focused on modifying the software model. To do so, we changed the "Tracking-based Recommendations" component in to an optional-variant that contains an "Anonymous Tracking-based Recommendations" variant that anonymizes the data and an "Identifiable Tracking-based Recommendations" variant that does not anonymize the data. The new component is shown in Figure 4.14.

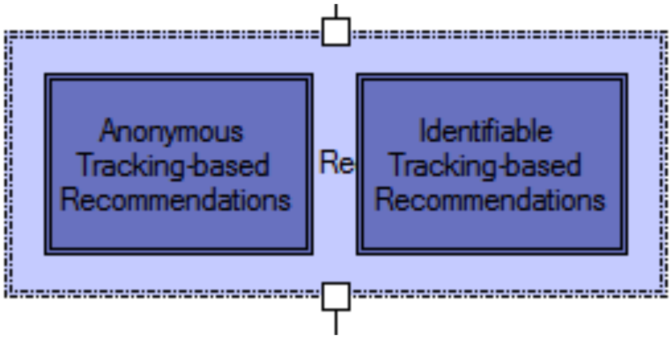

Anonymous Tracking-based Recommendations
Re
Identifiable Tracking-based Recommendations

Figure 4.14: New Optional-Variant

While the effort involved in the introduction of the new variants was minimal, we ran in to two limitations when trying to assign them guards. First, the extensional model disallows the calculation of intermediate values, such as whether or not the user's country is within the European Union. As such, we had to treat the new anonymization constraint as if it were placed on each country within the European Union individually, namely Germany and the UK. Second, we found that we could not assign guards to the new variants without considering the privacy constraints, which are part of the privacy model, since these are directly repersented in the guards. Instead, to assign Boolean guards the architect and legal professionals would have to work together to ensure that both the privacy and software models were correctly reflected in each guard.

The resulting extensional model contained one additional variable to express a user's preference regarding anonymization and two new, unique Boolean guards, one for each new variant. The resulting model has 4 variables, 12 variation points, and 7 unique Boolean guards.

*Scenario 1: intensional model*

To update the intensional PLA, we again focused on updating the software model first. We added a new change set which replaces the "Tracking-based Recommendations" component with a "Anonymous Tracking-based Recommendations" component. This change set is shown in Figure 4.15. We then updated two relationships in the software model to correctly integrate the new feature change set with the other feature change sets.

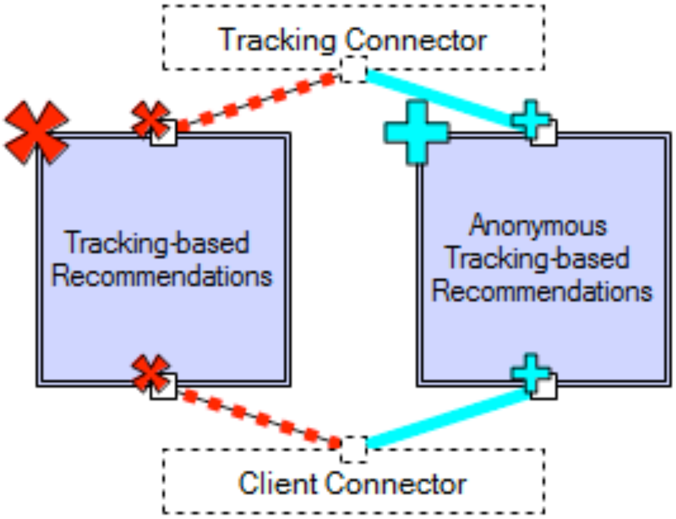

Tracking Connector
Tracking-based Recommendations
Anonymous Tracking-based Recommendations
Client Connector

Figure 4.15: New Change Set

To update the privacy model, we added 3 change sets and 5 relationships. Of significance was the fact that we could explicitly model the concept of being “In the European Union”. To do so, we created a “Country: Member of the EU” change set and added a relationship that implied this change set when the user indicated that they were in Germany or the UK.

To the intensional model, we added a total of 5 new change sets and 9 relationships, we also modified 2 relationships within the software model.

*Insights from scenario 1*

In this scenario, we found that we could express new concepts in the intensional model (i.e., whether a country is part of the EU) that we could not directly express in the extensional model. We also note that updating the intensional model involved more actual changes, but allowed us to focus on smaller parts of the model while making those changes, such as focusing on one domain model at a time.

### Scenario 2: modifying the software Model

In this scenario, the architect decides to refactor the anonymization feature introduced in the previous scenario. Instead of having two variants of the “Tracking-based Recommendations” component, one with anonymization and one without. The architect chooses to utilize a single “Tracking-based Recommendations” component and an optional “Anonymization” component. This change is conceptually confined to only the software model of the system.

*Scenario 2: extensional model*

To implement this architectural change, the architect reverts the optional-variant component back to an optional component and adds a new optional component, as shown in Figure 4.16.

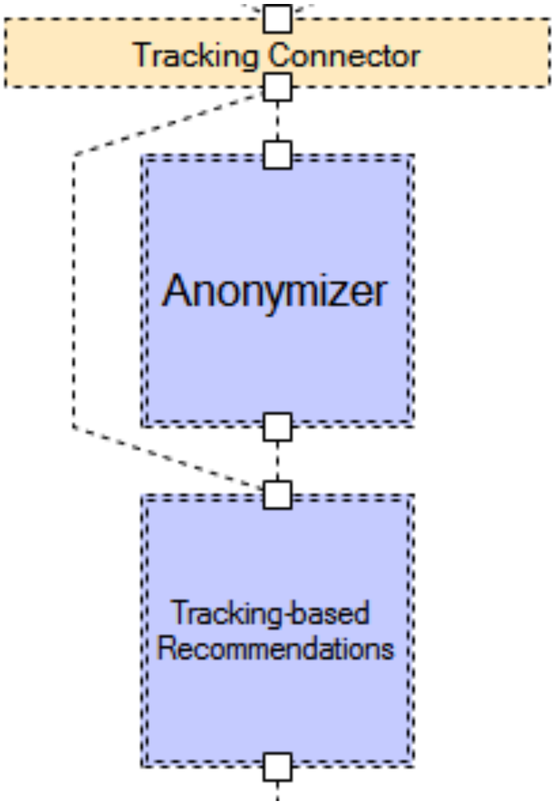


Figure 4.16: New Optional Component

Updating the model, required making the structural changes and then updating the guards. Again, we found that we cannot assign guards without considering both domain models because they are intertwined. Additionally, the guard for the "Anonymizer" component turned out to be more complex than that of the "Tracking-based Recommendations" component because it had to account for yet another preference variable. The guard for the new link in Figure 4.16 was equally complex.

The resulting extensional model contained two new, unique Boolean guards, however the two guards introduced in the previous scenario were removed. The resulting model has 4 variables, 13 variation points, and 7 unique Boolean guards.

*Scenario 2: intensional model*

In comparison, to update the intensional model, we modified the change set introduced in the previous scenario to that shown in Figure 4.17.

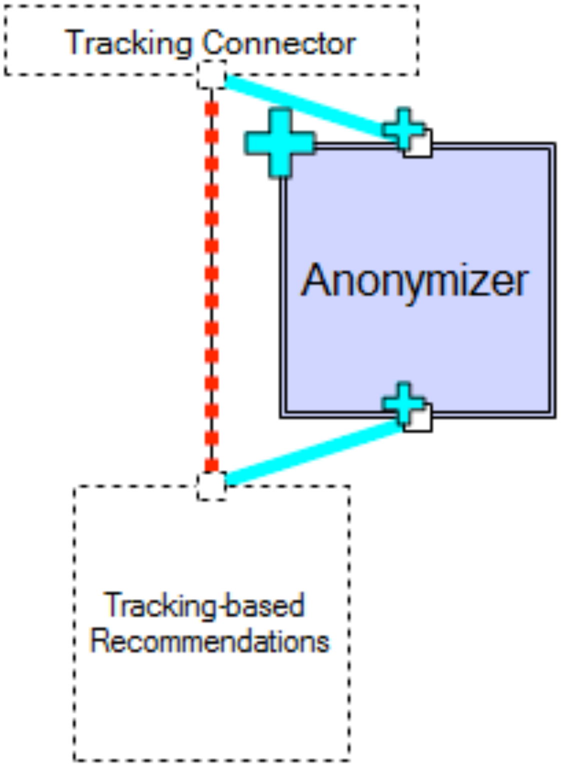


Figure 4.17: Revised Change Set

These modifications were contained within the single change set and so we did not need to consider any other concepts modeled in the system, even the relationships of the software model. Instead, we already knew that the particular change set was included or excluded from the model when appropriate.

*Insights from scenario 2*

We found that, for the extensional model, the amount of work involved was similar to that of the first scenario – each guard had to be interpreted, only a few were updated, and both domain models had to be considerd when updating the guards. In contrast, the intensional model proved far superior in this scenario. The feature was already represented using a change set, which was introduced in the previous scenario. Thus, we did not have to examin anything beyond the boundaries of that change set, instead we simply modified the change set's contents.

### Scenario 3: modifying the privacy-constraint model

For the final scenario, we consider a change that is confied to the privacy model. Here, we imagine that the German privacy law that prohibits cross-site tracking has been adopted by the European Union, making it applicable to all of its constituent countries.

*Scenario 3: extensional model*

No structural modifications were necessary for the extensional model, however a *significant* number of Boolean guards had to be modified. In fact all Boolean guards needed to be updated except for four. This was because so many other optional parts of the extensional model need to be present to utilize the "Cross-site Tracking" component, such as the "Tracking-based Recommendations" and "Anonymizer" components as well as the links and connectors between them. Specifically, the guards of 9 differnet variation points, consisting of 4 unique guards needed to be updated. Though many guards were updated, the overall number of variables, variation points and unique Boolean guards of the resulting model was the same.

*Scenario 3: intensional model*

To make this change in the intensional model, Relationship 4 from Figure 4.13 was duplicated and modified to reference the "Country: Member of the EU" change set added in the scenario 1 rather than the "Country: Germany" change set. While it would have been possible to simply modify Relationship 4 itself, we chose to duplicate it so that the privacy model accurate reflects the fact that both Germany *and* and the EU prohibit cross-site tracking without conscent, in case the laws change in the future.

*Scenario 3: insights*

We found in this scenario that the extensional model had to be significantly changed in order to reflect changes in the privacy model. This is because the privacy model is implicitly embedded through the extensional model at each point of variation. In the intensional model, we found that the graphical representation of relationships aided in identifiying the specific relationship that we wanted to copy, we simply looked at the relationships referencing the cross-site tracking preference change sets.

### 4.3.6 Discussion

We observe a number of tradeoffse between the two approaches from our initial modeling of the systems in Section 4.3.4 and the evolution scenarios in Section 4.3.5.

We found during the initial modeling and subsequent modification of the intensional and extensional models that introducing *new* concepts into the intensional model was generally more tedious, but easier than introducing these concepts in the extensional model. It was more *tedious* because, when compared to the Boolean guards in the extensional model, relationships in the intensional model tended to express small, individual dependencies. At the same time, the intensional model captured a more complete specifications (such as expressing whether a country is part of the EU, in Scenario 1) of both domain models. As a result more change sets and relationships were necessary to model the system. However, it was *easier* to model the concepts in the intensional model for these same reasons: smaller, isolated relationships were easier to express and the model was capable of modeling the additional concepts needed. This is as compared to the Boolean guards that were modeled in the extensional approach. For these, it was initially very difficult to determine how the relevant variabiles could be combined into single expressions and concepts such as the existence of the EU could not be modeled.

Even though the number of modeling constructs used in the intensional model was greater, interpreting and verifying the intensional model was significantly easier because we were able to focus on smaller, more isolated concepts. For instance, to verify that relationships captured the desried concepts correctly, we were generally able to talk through a logical argument. We did make a truth table for Relationship 4 in Figure 4.13, which is perhaps the most unintuitive relationship, but it was small. In contrast, to verify that the Boolean expressions of the extensional model captured the desired concepts, we had to resort to creating much larger truth tables.

Since the intensional approach models domain concepts explicitly and independently, it enables a better separation of the domain models. The privacy model explicitly represents various privacy constraints and their interdependencies. The software model explicitly represents architectural features and their interdependencies. For example, in order to implement the change introduced in Scenario 2, only the change sets and relationships of the software model needed to be examined and in turn only one change set needed to be updated. On the other hand, implementing the change in the extensional model involved both domain models as they were entangled in the Boolean guard expressions.

The clearer separation of domains also leads to a better division of work. Law professionals can primarily focus on the privacy model, while software architects can largely focus on the system model. Only the mappings from privacy constraints to architectural features demand knowledge of both domains and thus collaborations of lawyers and architects. For instance, in order to realize the change introduced in Scenario 3 in the intensional model, law professionals only needed to update a single relationship in the privacy model without any involvement of the software architects. In contrast, implementing the same change in the extensional model required both the law professionals and the software architects. In future work, it will be useful to add support for grouping and categorizing change sets and relationships according to their respective domain model.

## 4.4 Summary

In this section, we present a user modeling framework that leverages the concept of PLA to model the variability that exists in the privacy and personalization domain, and dynamically selects architectural instances to tailor the PLA to the specific needs of a particular user. The framework, thus, considers the privacy constraints that apply to an individual user and dynamically selects and instantiates a personalization architecture that provides personalized services to this specific user. The result is a flexible approach that not only helps address the complexity of building personalized systems, but also strongly supports their evolution: as new privacy and personalization concerns arise, they can be modularly added to the PLA.

We also describe how to use change sets and relationships to express the privacy model (privacy constraints and their dependencies), the system model (system components and their dependencies), and the relationships between the privacy model and the system model. We demonstrate that comparing with the extensional approach with Boolean guards, the intensional approach with change sets and relationships leads to clearer separation of the privacy models and the system models, and better supports their evolutions.

In a nutshell, we propose the PLA-based approach because of (1) its modular modeling capability, (2) its any-time architectural reconfiguration support, and (3) its excellent tool support, in particular ArchStudio for PLA representation, selection and instantiation, and (3) its increasing industry adoption (thus companies are easier to adopt this PLA-based approach).

How would this approach then help address or bridge the four gaps identified in existing work (see Section 3.5.1)? Our PLA-based approach not only models (Gap 1) but enforces (Gap 2) privacy constraints and their impacts on personalization methods in web-based personalized systems. Since our framework supports run-time dynamic architectural reconfiguration and instantiation, it can respond immediately when a user changes his or her privacy constraints. As such, it allows a web-based personalized system to adjust its privacy practices with regard to individual users (Gap 3) in a highly dynamic and responsive manner (Gap 4). Therefore, we believe our approach demonstrates a promising direction to address all the four gaps.

Despite its advantages, this PLA-based approach also poses a significant challenge in system performance since theoretically the resource-intensive runtime architectural reconfiguration and instantiation could be evoked for every single user session (if no two user sessions share the same privacy constraints). To mitigate this issue, we propose a combination of techniques including computation distribution, caching and light-weight representation of PLA. Whether these techniques can help improve the performance of this PLA-based approach to provide privacy-enhanced personalized services to a large number of users will be explored in Section 5.1. Besides, what if end users are enabled to control their privacy constraints and to see the effects of these constraints on the personalized system? How would this affect end users' perceptions about the privacy practices of the system and perhaps more importantly users' behaviors on the system? We will investigate this issue in Section 5.2.

# Chapter 5

# Evaluations

In order to assess the impacts of the PLA-based dynamic architectural reconfiguration and instantiation on the performance of personalized systems, we conducted a simulation-based system performance study. Besides, we conducted a controlled user experiment to assess the effects of this privacy enhancement from users' standpoint.

## 5.1 Performance Evaluation

One major concern about our approach is its performance since dynamic architectural reconfiguration during runtime is usually resource-intensive. Will it be practically possible to deploy such a dynamic system in a contemporary internationally operating website? In this section, we describe four variant implementations of our system and an in-depth performance evaluation under realistic workload conditions. Our work stands in the tradition of similar attempts in the past to gauge the performance of user modeling tools through simulation experiments (such as (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 constraint of each individual user.

### 5.1.1 Experimental design and procedures

**Controlled variables**

Since we suspected that the XML-based Myx implementation described in Section 4.2.4 would perform poorly, we aimed at contrasting it with the two optimization methods described in Section 4.2.4 and 4.2.4 through the following 2-factorial design: (Myx vs. Customized) × (Non-caching vs. Caching).

**Simulation parameters**

Since we anticipated that a very large network of machines will be needed to handle real-world large-scale applications that was unavailable to us, we identified a reasonable number of 3000 maximum users per host in pre-trials 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 et al., 2006a; Wang and Kobsa, 2006, 2007), as well as the user modeling literature:

- Total number of UMCs in the PLA: 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; Chlebus and Brazier, 2007) has empirically shown that the arrival of new user sessions at a website largely follows a Poisson process[1]. 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.

[1] 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.

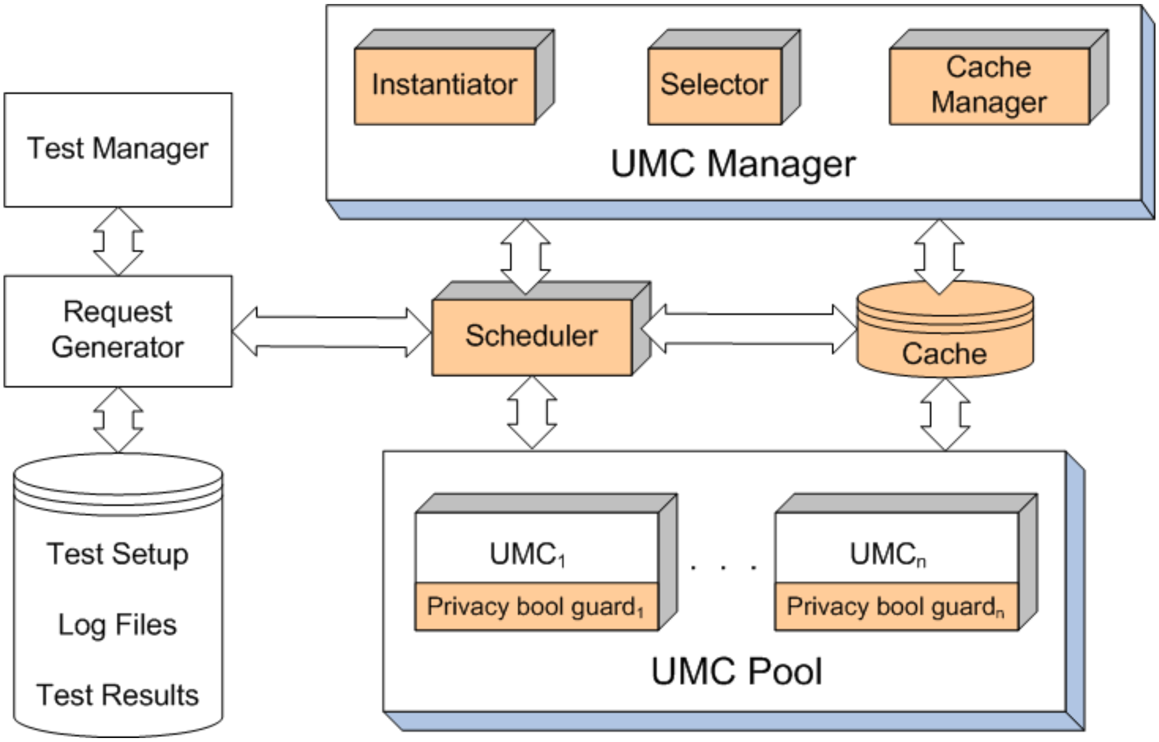


Figure 5.1: Testbed architecture

### Testbed

Figure 5.1 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 anyways when deployed in practice. We added a Test Manager to control 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.

### 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 4.2.2). Once the session has been assigned to a runtime system instance, the assignment is written into the cache if a cache is used. When 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, has a range of [0, 1) and is calculated as $\frac{Total\ Sessions\ -\ Total\ Instances}{Total\ Sessions}$

**Performance improvement (percentage),** 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 $\frac{\sum TotalHandlingTimeOriginalVersion\ -\ \sum TotalHandlingTimeEnhancedVersion}{\sum TotalHandlingTimeOriginalVersion}$

### 5.1.2 Evaluation results

#### Handling Time per User Session

Figure 5.2 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 multi-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.

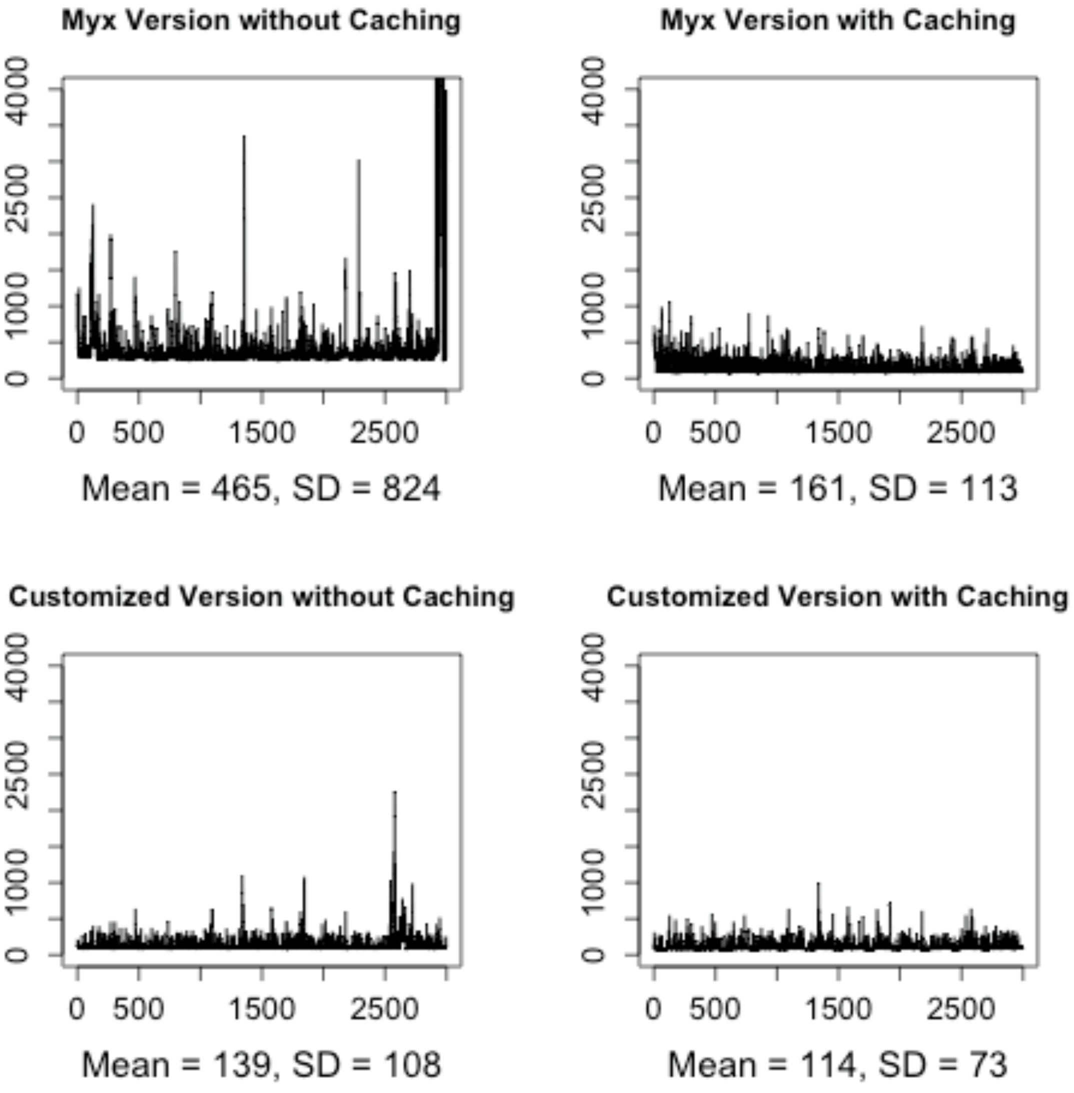
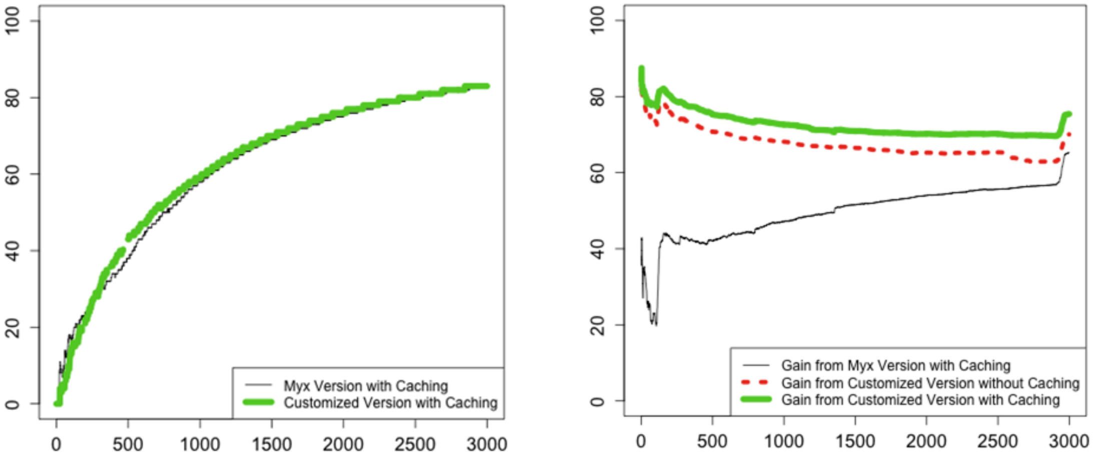


Figure 5.2: Handling time for each user session (milliseconds)

We also analyzed the spikes of the handling time in Fig. 5.2 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 delay lies in Java's indeterministic thread scheduling. Requests to handle a new session, select an architecture, and instantiate an architecture each creates 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, high handling times increase towards the end of the experiment. This is because the machine almost ran out of heap space, and the Java Virtual Machine kept switching threads. A good remedy for these effects of indeterministic thread switching is to shorten the processing time, which is confirmed by the substantial decrease of such delays in the conditions in which the customized version and/or caching have been used.

### Runtime Instance Reuse Rate

Figure 5.3(a) plots the runtime instance reuse rates for the two caching versions (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.



(a) Runtime Instance Reuse Rate

(b) Performance Improvement

Figure 5.3: Instance reuse and performance improvement (both in %), by cumulative number of users

### Performance Improvement

Figure 5.3(b) plots the performance gain of our three improved versions in comparison to the baseline Myx version without caching. The curve at the bottom (gain from Myx version with caching) goes up as expected: the cache size increases with an increasing number of users, and hence the hit rate and thus the performance gain increase. The curve in the middle (gain from 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 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, it is unfortunately not additive. While with increased number of users the gains through caching increase, each hit "cancels out" the gains through customization which will not be invoked in such a case. Larger cache sizes still cause performance gains as is demonstrated by the slightly increasing distance between the middle and upper curve. 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.

### 5.1.3 Discussion

*Performance Improvement.*

The evaluation results show that both our customization and caching improve the performance. 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 multi-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. In contrast, Yahoo.com which Alexa currently ranks No. 1 worldwide in terms of traffic seems to have a daily reach of close to 30 million unique visitors (Alexa, 2009). This roughly translates into an average arrival rate of 350 users 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). If we continue using our average arrival rate of 0.5 visitors for each node, then we can handle Yahoo-sized traffic with a cloud that consists of 700 nodes on average. 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 the Web server.

*Limitations of the Evaluation.*

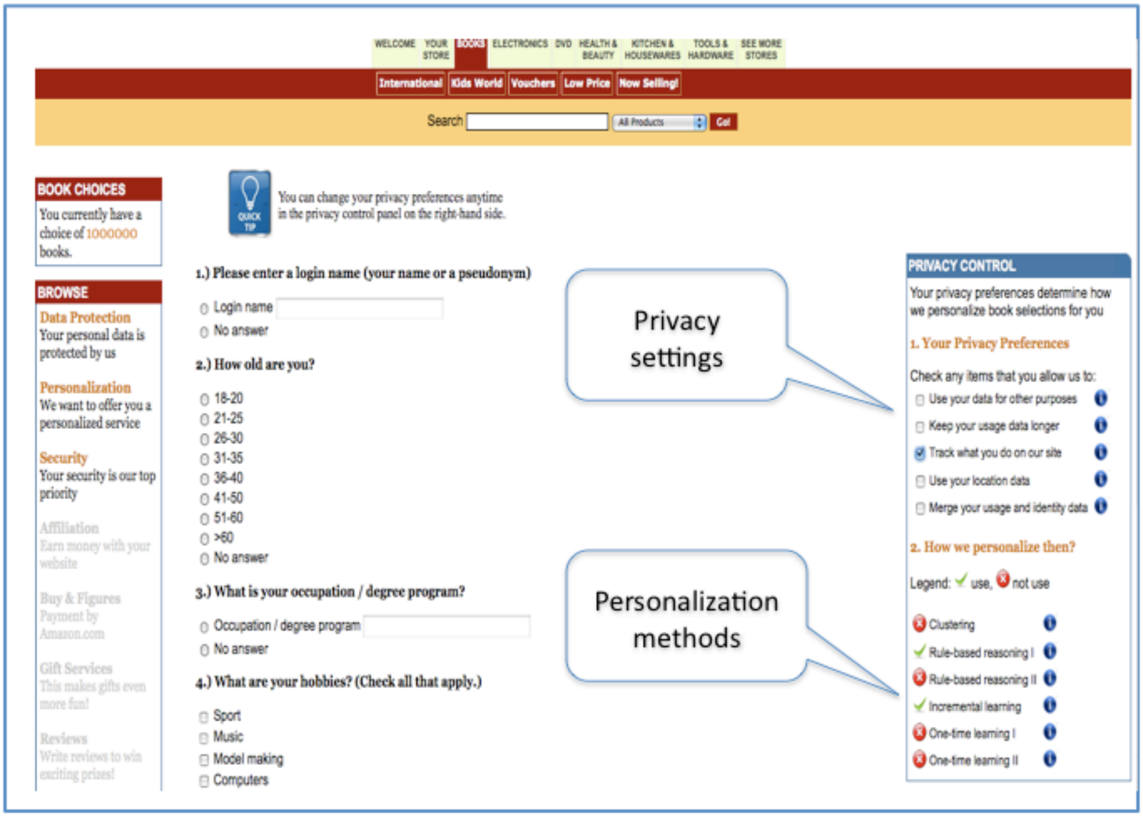
Privacy bindings are randomly assigned to sessions in our simulation, and hence their variations are evenly distributed across users. In reality though, users' individual privacy preferences are likely to gravitate towards "typical preferences", countries may have typical combinations of privacy bindings, and visitors from certain countries may be more frequent than from others. The hit rate in the privacy binding cache is likely to be higher in this more realistic scenario with uneven distribution, and the number of generated different instances lower than in our simulation, both of which reduces the memory load. Another limitation is that the 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.

### 5.1.4 Summary

Our performance evaluation shows that our light-weight customized implementation performs better than the original PLA implementation (the Myx version), that our multi-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 version. Overall, our results demonstrate that with a reasonable number of networked hosts in a cloud computing environment, an internationally operating website can use our dynamic PLA-based user modeling 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.

## 5.2 User Experiment

In section 4.1 we described the proposed privacy enhancement mechanism, i.e., the underlying personalized system selects and uses different personalization methods according to each individual user's personal privacy preferences as well as applicable privacy regulations. The main purpose of this controlled user experiment is to evaluate the effects of this privacy enhancement mechanism on users' experience with a personalized website.



BOOK CHOICES

You currently have a choice of 1000000 books.

BROWSE

Data Protection
Your personal data is protected by us

Personalization
We want to offer you a personalized service

Security
Your security is our top priority

Affiliation
Earn money with your website

Buy & Figures
Payment by Amazon.com

Gift Services
This makes gifts even more fun!

Reviews
Write reviews to win exciting prizes!

You can change your privacy preferences anytime in the privacy control panel on the right-hand side.

1.) Please enter a login name (your name or a pseudonym)

○ Login name
○ No answer

2.) How old are you?

○ 18-20
○ 21-25
○ 26-30
○ 31-35
○ 36-40
○ 41-50
○ 51-60
○ >60
○ No answer

3.) What is your occupation / degree program?

○ Occupation / degree program
○ No answer

4.) What are your hobbies? (Check all that apply.)

□ Sport
□ Music
□ Model making
□ Computers

PRIVACY CONTROL

Your privacy preferences determine how we personalize book selections for you

1. Your Privacy Preferences

Check any items that you allow us to:

□ Use your data for other purposes
□ Keep your usage data longer
☑ Track what you do on our site
□ Use your location data
□ Merge your usage and identity data

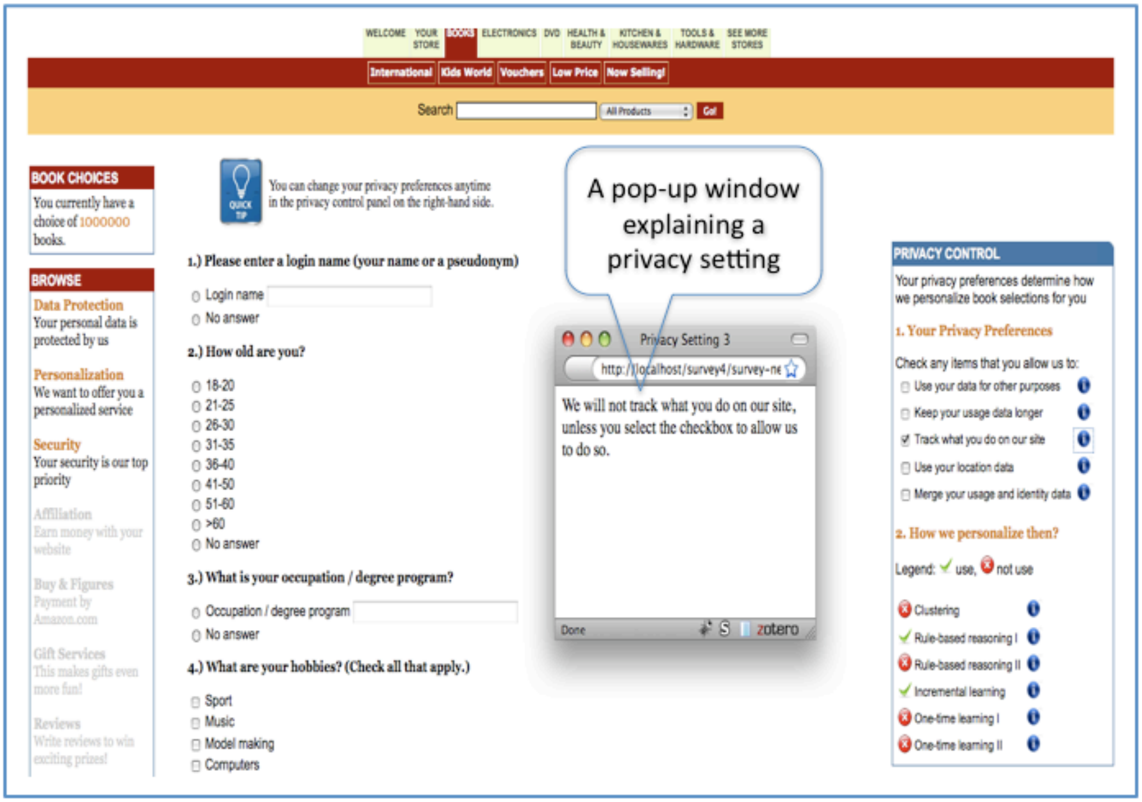
2. How we personalize then?

Legend: ✓ use, ☒ not use

☒ Clustering
✓ Rule-based reasoning I
☒ Rule-based reasoning II
✓ Incremental learning
☒ One-time learning I
☒ One-time learning II

Figure 5.4: User Interface for the Proposed Privacy Enhancement

We designed a control panel for the user interface that allows users to specify their personal privacy preferences on a personalized website. Fig. 5.4 shows this "Privacy Control" panel on the right-hand side of the screen. It has two components. The top component/subwindow contains a list of privacy preference settings that a user can specify. More specifically, they are permissions/consent that a user can give to the system. 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 on the site. By default, all the privacy options are unchecked. The bottom component/sub-window contains a list of personalization methods. The second method “Rule-based reasoning I” is a minimum personalization method that the system always uses. Other methods will be 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 turned on as well. If the user changes her privacy settings, the personalization methods will be re-evaluated for selection. Only the selected personalization methods will be used to provide personalized services to this user. It is important to note that there were no actual personalization methods running in the system at all. We used deception as a research tool to simplify the experiment setup with minimum impact on the validity of the results (see Section 5.2.8 for a methodological discussion of deception). More specifically, we told all the subjects that their personalized book recommendations will be generated by these personalization methods based on the answers they provide to the questions in the experiment. But in fact, regardless of their answers, every subject was presented with the same set of 50 recommended books that were hand picked by the researchers in advance. Because of the use of deception in our experiment, subjects’ behavior should reasonably match their counterpart in the situation that there are actual personalization methods running.



BOOK CHOICES

You currently have a choice of 1000000 books.

BROWSE

Data Protection
Your personal data is protected by us

Personalization
We want to offer you a personalized service

Security
Your security is our top priority

Affiliation
Earn money with your website

Buy & Figures
Payment by Amazon.com

Gift Services
This makes gifts even more fun!

Reviews
Write reviews to win exciting prizes!

You can change your privacy preferences anytime in the privacy control panel on the right-hand side.

1.) Please enter a login name (your name or a pseudonym)

○ Login name
○ No answer

2.) How old are you?

○ 18-20
○ 21-25
○ 26-30
○ 31-35
○ 36-40
○ 41-50
○ 51-60
○ >60
○ No answer

3.) What is your occupation / degree program?

○ Occupation / degree program
○ No answer

4.) What are your hobbies? (Check all that apply.)

□ Sport
□ Music
□ Model making
□ Computers

Privacy Setting 3

http://localhost/survey4/survey-ne

We will not track what you do on our site, unless you select the checkbox to allow us to do so.

Done

PRIVACY CONTROL

Your privacy preferences determine how we personalize book selections for you

1. Your Privacy Preferences

Check any items that you allow us to:

□ Use your data for other purposes
□ Keep your usage data longer
☑ Track what you do on our site
□ Use your location data
□ Merge your usage and identity data

2. How we personalize then?

Legend: ✓ use, ☒ not use

☒ Clustering
✓ Rule-based reasoning I
☒ Rule-based reasoning II
✓ Incremental learning
☒ One-time learning I
☒ One-time learning II

Figure 5.5: A pop-up window explaining a privacy option

There are blue "i" icons next to each of these privacy options as well as personalization methods. If a user clicks on one of them, a pop-up window will display with more details for the particular privacy option or personalization method. Fig. 5.5 shows an example of a pop-up window. It is important to note that this privacy UI (we use privacy UI and privacy control panel interchangeably) is shown persistently in all the pages of this web site. In addition, a "quick tip" that reminds users can change their privacy settings anytime using the Privacy UI is persistently shown on the top of the page.

### 5.2.1 Background

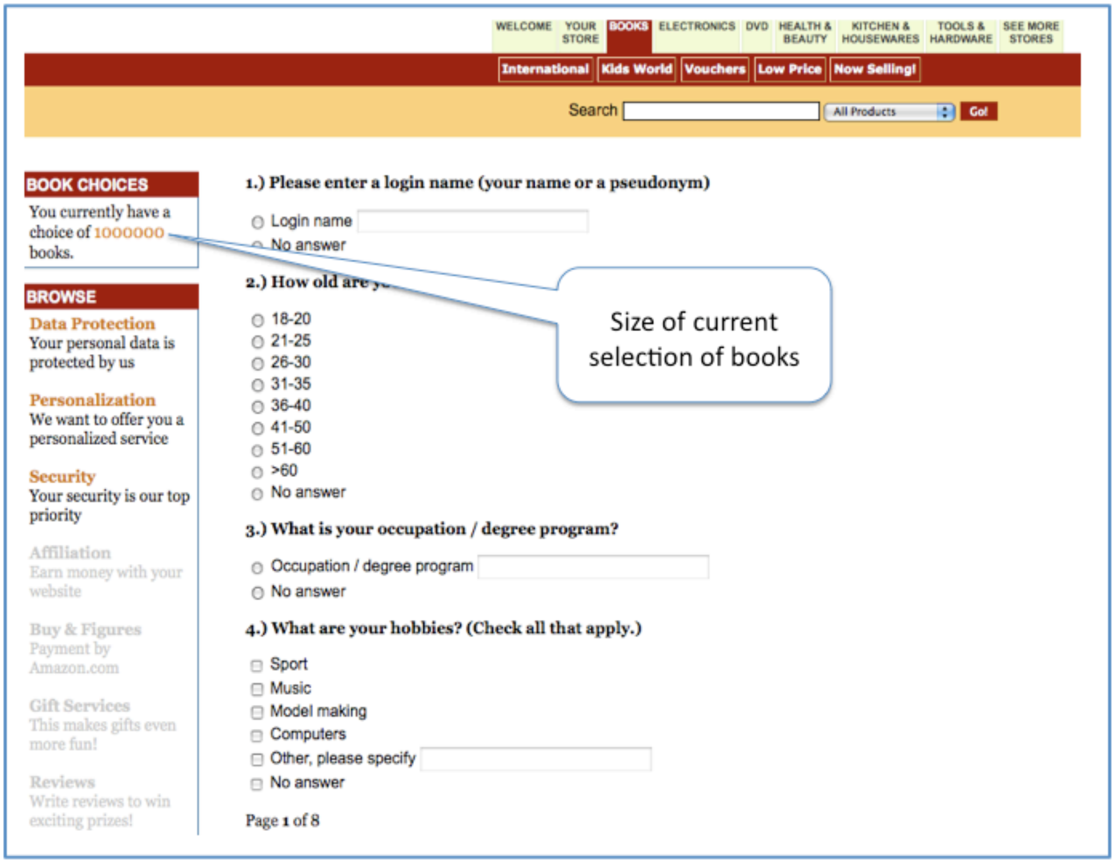
There are two methodological approaches to study users' reactions to different UI designs: inquiry-based and observational approaches. In the first approach, users would be interviewed about their opinions of the respective UI design. Users are often provided with representations of the UI designs, from paper-based sketches to fully functioning UI. In the second approach, users are being observed while carrying out some tasks using the UI. 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-reported behavior.

This latter problem seems to prevail in the area of privacy. Spiekermann et al. (2001a); Berendt et al. (2005) found that users' stated privacy preferences deviate significantly from their actual behavior. Also, an enormous discrepancy can be observed between the number of people who claim to read privacy policies and the actual access statistics of these pages. Solely relying on interview-based techniques for analyzing privacy impacts on users, as is currently nearly exclusively the case, must therefore be viewed with caution. Our empirical study therefore gravitated towards an observational approach, which we complemented by questionnaires and brief informal interviews.

### 5.2.2 Experiment design

To test the effects of the proposed mechanism, we designed a between-subjects experiment with two conditions. The subjects were randomly assigned to one of these two conditions/groups:

- The first group ("control" group) used the standard interface of the personalized website (i.e., without the privacy panel built into it). (see Fig. 5.6)



You currently have a choice of 1000000 books.

Data Protection
Your personal data is protected by us

Personalization
We want to offer you a personalized service

Security
Your security is our top priority

Affiliation
Earn money with your website

Buy & Figures
Payment by Amazon.com

Gift Services
This makes gifts even more fun!

Reviews
Write reviews to win exciting prizes!

1.) Please enter a login name (your name or a pseudonym)

○ Login name
○ No answer

2.) How old are y

○ 18-20
○ 21-25
○ 26-30
○ 31-35
○ 36-40
○ 41-50
○ 51-60
○ >60
○ No answer

3.) What is your occupation / degree program?

○ Occupation / degree program
○ No answer

4.) What are your hobbies? (Check all that apply.)

□ Sport
□ Music
□ Model making
□ Computers
□ Other, please specify
□ No answer

Page 1 of 8

Size of current selection of books

Figure 5.6: User Interface for the Control Group

- The second group ("enhanced" group) used the version with the privacy panel. The researchers also gave them verbal explanation of the mechanism. The privacy panel was constantly present on the website and subjects could interact with it while carrying out the experiment tasks. (see Fig. 5.4)

The experiment was designed to determine whether subjects exhibit different data sharing behavior and purchase behavior depending on the condition they receive. Our hypothesis was that users in condition 2 would be more willing to share personal data and view sites more favorably than users in condition 1. We treat the condition as an independent variable and users' perceptions and behaviors in the website (measured in term of the number of questions they answer about themselves, whether they purchase books, and their perception of the site's privacy practices and personalization benefits) as dependant variables.

### 5.2.3 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 were created for the two aforementioned conditions. Fig. 5.6 shows an excerpt of the first variant which was used in condition 1, while Fig. 5.4 shows an excerpt of the second variant which was used in condition 2.

A counter was visibly placed on each page that purported to represent the size of the currently available selection of books. Initially the counter is set to 1 million books. Data entries in web forms (both via checkboxes and radio buttons and through textual input) decrease the counter after each page by a random amount. Its behavior was designed to give study participants the feeling that the more information they provide, the smaller the set of books being selected. The web forms ask a broad range of questions relating to users' interests. A few sensitive questions on users' political interests, religious interests and adherence, their literary sexual preferences, and their interest in certain medical subareas (including venereal diseases) are also present. All questions "make sense" in the context of filtering books in which users may be interested. For each question, users have 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 are presented. Ten questions allow multiple answers, and seven questions have 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 are encouraged to review their entries and then to retrieve books that purportedly match their interests. The website then displays fifty predetermined and invariant books (see Appendix C for the list of 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 are visibly marked down by 70%, resulting in 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 are given the opportunity to buy only one book from the set of 50 recommended books at a 30% price. They are free to choose whether or not to make that purchase. Those who do are being asked for their shipping and payment data (a choice of debit card or credit card charge is offered).

### 5.2.4 Subjects

We used the following criteria for screening participants: previous online shopping experience and owning a credit card or a debit card that can be used for online purchases. We conducted a series of pilot tests and made changes accordingly. 65 subjects participated in the experiment. They were undergraduate, graduate students and staff from a wide range of departments in a large public university in the U.S.. The data of 7 subjects were eventually not used since the researchers knew them by sight and felt they may not feel truly anonymous.

### 5.2.5 Experiment procedures

Recruitment of study participants were taken place through posters, and email announcements in various campus distribution lists. Participants were promised a $10 coupon for a nearby popular coffee shop as a compensation for their participation, and the option to purchase a book with a 70% discount. Prospective participants were asked to bring their IDs and credit or debit cards 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.

Each subject was given a copy of the study information sheet and informed that he or she would test an experimental new version of the online bookstore with an intelligent book recommendation engine inside. Users were told that the system would ask them a number of book-related questions and then generate 50 books as personalized recommendations based on their answers to these questions. Users were also told that the more and the better answers they provided to these questions, the better would generally be the quality of their personalized book recommendations. 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 a credit card or a debit card.

A between-subjects design was used for the subsequent experiment, with the experiment conditions as the independent variable (see Section 5.2.2 for details). Subjects were randomly assigned to one of the two conditions (we will abbreviate them by "control" and "enhanced" in the following). After searching for books and possibly buying one, subjects filled in a post-questionnaire. This questionnaire uses the Concern for Information Privacy (CFIP) scale (Smith et al., 1996) to gauge subjects' privacy concerns and other questions to assess subjects' perceived benefits of personalization. In the debriefing phase of the study, the data of those users who had bought a book were compared with the credentials that subjects had brought with. These subjects were also informed that their credit cards will not be charged, instead when their books arrive at an on-campus location (instead of the shipping addresses they entered) they can pick up their books in person and pay the book prices in cash. The researchers then deleted all of their payment information except their names and email addresses from the system in front of the subjects.

### 5.2.6 Results

We had a total of 58 valid subjects and conducted analysis based on their results. There are 29 participants in each of the two conditions: "control" and "enhanced". There is no statistically significant difference between the two groups (i.e., subjects in the two experiment conditions) in terms of their general privacy concerns based on the CFIP scale (Smith et al., 1996).

**Data Sharing Behavior.**

*Number of questions answered.*

We first dichotomized their responses by counting whether a question received at least one answer or was not answered at all. Every question has a "no answer" option, and we treated the selection of this option as the question not being answered at all. On average, 87% of the questions were answered in the "control" condition, while this rose to 91% in the "enhanced" condition (see Table 5.1). A Chi-Square test on a contingency table with the total number of questions answered and not answered in each condition showed that the difference between conditions was statistically significant (p=0.012). We also tested whether there is a difference for more privacy-sensitive questions (Question 5, 16, 19, 20, 21, 22, 23, 24, 25, and 31) and irrelevant questions[2] (Question 12, 13, and 14) and we found no statistically significant results (see Appendix A for all the questions asked in the

[2] The answers to these questions do not help the book recommendations, e.g., "Do you go on package vacations or rather on individual travel?"

experiment).

Table 5.1: Data sharing behavior and results of Chi-Square tests

|  | “control” group | “enhanced” group | df | Chi-Square | p | N |
|---|---|---|---|---|---|---|
| % Questions answered | 87% | 91% | 1 | 6.34 | 0.012 | 1856 |
| % Answers given | 59% | 63% | 1 | 5.71 | 0.017 | 3712 |

*Number of answers given.*

The two conditions also differed with respect to the number of answers given (see Table 5.1). The maximum number of answers that any subject 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 Chi-Square contingency test showed again that the difference between the two conditions was statistically significant (p=0.017). We also tested whether there is a difference for privacy-sensitive questions (Question 5, 16, 19, 20, 21, 22, 23, 24, 25, and 31) and irrelevant questions (Question 12, 13, and 14) and we found no statistically significant results.

### Purchases

Table 5.2 shows that the purchase ratio in the “enhanced” condition is 60% higher than in the “control” condition (note that 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).

Table 5.2: 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 |

We regard this as an important confirmation of the success of our proposed privacy enhancement mechanism. In terms of privacy, the decision to buy is a significant step since at this point users reveal personally identifiable information (name, shipment and payment data) and risk that previously pseudonymous information may be linked to their identities. The privacy UI that allows users to set their privacy preferences and to view the personalization methods that the underlying system uses to generate personalized recommendations in accordance with their privacy settings seemingly alleviates such concerns.

**Summary on data sharing and purchase behavior.**

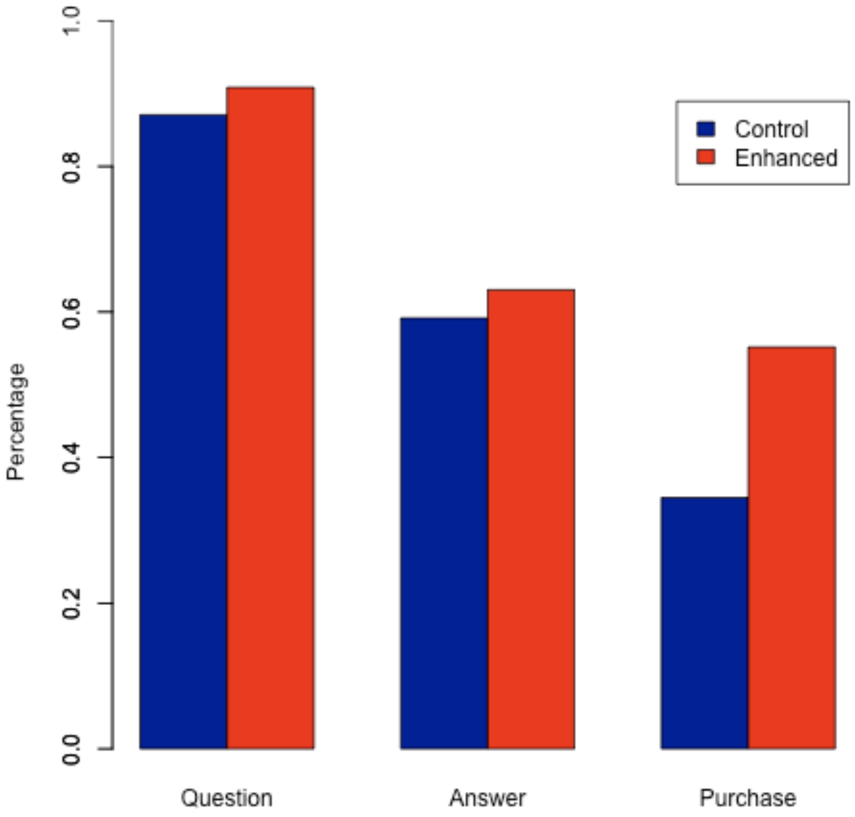


Figure 5.7: Summary of data sharing and purchase behavior

Fig. 5.7 summarizes of the above results regarding data sharing and purchase behavior. More specifically, it shows the average percentages of questions being answered, the average percentages of answer options being given, and the percentage of subjects who made a book purchase, for the "control" and the "enhanced" conditions, respectively. The above results demonstrate that our privacy enhancement mechanism in the form of the privacy UI has positive effects on users' willingness to share personal data and to make online purchases.

### Rating of Privacy Practices and Perceived Benefits Resulting from Data Disclosure

The post-experiment questionnaire that was administered to each subject at the end of the study includes a number of Likert questions (whose possible answers range from “strongly disagree” to “strongly agree”). It examines how users perceive the level of privacy protection at the website as well as the expediency of their data disclosure in helping the company recommend better books. The responses to the Likert questions were encoded on a one to five scale. A one-tailed t test revealed that 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 between the two conditions in the statement "the new book website uses my data in a responsible manner" approached significance (p<0.12). However, subjects’ perceptions of whether their data disclosure helped the bookstore in selecting interesting books for them were not significantly different between the two conditions (p=0.36). Note again that all subjects were offered the same set of books. More details about these results can be found in Table 5.3.

Table 5.3: Users’ perception of privacy practice and benefit of data disclosure. 1: strongly disagree, 2: disagree, 3: not sure, 4: agree, 5: strongly agree.

| Item | “control”<br>(mean) | “enhanced”<br>(mean) | df | t | p | N |
|---|---|---|---|---|---|---|
| Privacy has priority | 3.34 | 3.79 | 55 | -2.09 | 0.02 | 58 |
| Data is used responsibly | 3.59 | 3.83 | 56 | -1.16 | 0.12 | 58 |
| Data helped select interesting books | 3.00 | 2.89 | 55 | 0.37 | 0.36 | 58 |

### Self-Reported Practices and Perceived Usefulness of the Privacy Control

Subjects in the “enhanced” condition were asked six additional questions regarding the privacy control in the post-experiment questionnaire. 83% of the subjects said that they paid attention to the privacy control during the study, while only 38% of the subjects expressed that they clicked at "information" icon(s) in the privacy control to obtain more information about the privacy preferences and/or personalization methods. 66% of the subjects said that they set privacy options in the privacy control in order to change their privacy preferences, while 41% of the subjects did that in order to try out what happens. Table 5.4 summarizes these results.

Table 5.4: Users' self-reported practices of the privacy control panel

| Item | "enhanced" group | 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 control panel in order to change my privacy preferences | 66% | 29 |
| I set options in the privacy control panel in order to try out what happens | 41% | 29 |

The remaining two questions were attitudinal, Likert (whose possible answers range from "strongly disagree" to "strongly agree") questions. Again, the responses to the Likert questions were encoded on a one to five scale. Table 5.5 shows the results of these two attitudinal questions.

Table 5.5: Users' perception of the usefulness of the privacy control panel. 1: strongly disagree, 2: disagree, 3: not sure, 4: agree, 5: strongly agree.

| Item | "enhanced" (mean) | N |
|---|---|---|
| Privacy control panel is useful in general | 3.97 | 29 |
| I would use a privacy control panel if a site offers one | 4.03 | 29 |

We regard the above results regarding the privacy control as further evidence of the success of our proposed privacy enhancement mechanism. The majority of the subjects paid attention to the privacy control, found it useful, used it in setting their privacy, and would use it if a site offers one.

### Comments about the Privacy Control from Informal Interviews

We also conducted brief informal interviews after the experiment. We asked the subjects in the “enhanced” condition if they have any comments about the privacy control and any suggestions to improve it. In general, users liked the idea of the privacy control. 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 and further explanations (from the “Info” icons) of the privacy options and personalization methods were difficult to understand. For instance, what does “other purposes” mean in the first privacy option of “user your data for other purposes”, or what is the “clustering” personalization method? They suggested providing more concrete explanation 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 in food 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 privacy control, or quickly played with the privacy options to select the most powerful personalization. Some other users largely focused on the privacy options and ignored the personalization methods since they did not quite understand the later. There was one subject in the “enhanced” condition said that “I liked the idea (privacy control), but I didn’t play with it. It reminds me privacy concerns, I chose ’no answer’ for many questions.”

### 5.2.7 Discussion

Our experiment was designed so as to create an online shopping experience as realistic as possible, and thereby to increase its ecological relevance. The incentive of a highly discounted book and the extremely large selection set that visibly decreased with more answers given was designed to incite users to provide ample and truthful data about their interests. The claim that all data would be made available to the website meant that users really had to trust the privacy policy that the website promised when deciding to disclose their identities.

The results demonstrate that our proposed privacy enhancement mechanism has a significant positive effect on users' data sharing behavior, and on their perceptions of the website's privacy practices. The additional finding that this mechanism also leads to more purchases was approaching statistical significance. We believe that our privacy control panel offers end users more transparency and control of their privacy and correspondingly the personalized system. The adoption by web retailers of interface designs that contain such a privacy control panel therefore seems clearly advisable. However, the concrete design of the privacy control panel still needs further exploration and verification.

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.

In an earlier experiment conducted in Germany (we denote it as the German experiment thereon), our collaborators assessed the effects of a contextualized privacy policy interface design 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 in the experiment. Note that we had to choose a different set of 50 books in our experiment since many of the 50 books in the German experiment were in German and thus not appropriate for our study. However, we tried to keep the range of book prices and types of book topics the same. In a nutshell, our results reveal similar trends discovered in the German experiment. Both their and our privacy enhancement mechanisms have shown 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 the subjects in our experiment were from various disciplines such as engineering, mathematics, chemistry, biology, medicine, social sciences and law. Choosing 50 books that may potentially interest a heterogenous group is seemingly more difficult than doing the same task for a homogenous group. Indeed, many of our subjects said they did not find the recommended books interesting. This may explain why we did not get significant result on the perceived usefulness of disclosing data because our pre-selected 50 books failed to interest many of our diversified group of subjects. Besides, our statistically significant results were less significant than the German counterparts (i.e., our results had larger P values). We suspect that this may be due to the fact that ordinary users have quite limited technical savviness or mental models of personalized systems. Therefore, they could not fully understand and take advantage of the privacy control panel, especially the part of personalization methods.

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 were unlikely to be favorable. Clear and readily available explanations might alleviate some of the unfavorable speculation". One may speculate that the opportunity offered by the Privacy UI in illustrating the relationship between users' data disclosure and the underlying personalization may alleviate some of the unfavorable speculation. 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 Privacy UI makes the personalization process more transparent and better yet, more controllable for the end users. The Privacy UI in our experiment aligns with the “openess” 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, for instance, the reputation of a website. We chose a webstore 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); Xie et al. (2006)). Our high response rates of 87% without and 91% with the Privacy UI 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 will experience an even stronger effect of our privacy enhancement mechanism.

In our experiment, the Privacy UI 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, replacing the Privacy UI with a link or an icon that symbolizes the availability of such privacy control panel? If so, how can the privacy control panel be presented so that users can easily access them and at the same time will not be distracted by them? 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 learnt that many subjects were not really trying to understand or even paying attention to the part of personalization methods. This raises the question that by default whether we should present the part permanently or show it as a link or an icon. Several subjects also asked for more concrete textual explanations of the privacy options and personalization methods. This remains as a UI design challenge for us because we are constrained by the size of the panel and cannot put too much verbiage on it. Developing a succinct and yet clear representation of these options and methods is one of our future directions. One idea is to explore the "calorie" metaphor that one of our subjects suggested. More specifically, the subject pointed to the idea that any food sold in the US market would have a calorie chart that enumerates the ingredients and their respective energy information (in calories). The food calorie chart helps consumers scrutinize the food energy information before purchase and/or consumption. Similarly, we can design a privacy "calorie" chart for privacy options and personalization methods illustrating what type of personal data (e.g., usage log), how much of the data (e.g., a single session or 1 year), and how the data will be used (e.g., 1-year usage data combined with your demographic data such as age). In a similar attempt in representing privacy policy, the idea of privacy label seemed to be a promising direction (Kelley, 2009). Some subjects also requested a clearer representation of the relationship between their data disclosure and the quality of the personalization they get. Instead of just showing the turn-on and turn-off of personalization methods based on users' privacy settings, we can potentially show examples of books that would otherwise not be recommended if a certain privacy option is not checked. In other words, the effect of setting a privacy option is now reflected in terms of recommended books rather than personalization methods which are presumably more difficult to grasp for end users. All of these design ideas need to be tested with additional user studies.

### 5.2.8 Limitations

Despite our efforts to create an online shopping experience as realistic as possible, the experiment setup still did not completely replicate the authentic online shopping experience. We asked the subjects to conduct the experiment in our laboratory instead of their home where people usually do online shopping. One may argue that people may behave differently in the lab environment. A lab experiment was deemed necessary because (1) we needed to verify whether a subject provides his or her true information when buying a book; and (2) privacy is highly situated and contingent and thus the physical environment in which a subject conducts the experiment may affect their behavior. Therefore, we chose to use the same physical environment (our lab) for all the experiment subjects. In principle, we champion the idea of studying people's privacy attitudes and behaviors in their authentic context of usage (e.g., Nguyen et al. (2009)). However, we were constrained by practical difficulties of taking this approach.

We use deception in our experiment where we told the site is for Amazon and we did not mention this is a privacy study. While deception as a methodological tool can cause ethical and moral issues, it can be a powerful tool for research when its use is justified and appropriate (Christensen, 1988). We told the subjects that this is an Amazon site because the German experiment (Kobsa and Teltzrow, 2005) that we planed to compare results with used the Amazon design. We did not tell subjects that the experiment was designed to study their privacy attitudes and behavior because people may become more privacy concerned if they are reminded about privacy. This would bias our results. We told all the subjects the "truth" in the debriefing. We believe our usage of deception in the experiment is justified and appropriate.

### 5.2.9 Summary

In summary, our controlled user experiment show that our proposed privacy enhancement mechanism has a significant positive effect on users' data sharing behavior, and on their perceptions of the website's privacy practices. The additional finding that this mechanism also leads to more purchases was approaching statistical significance. We believe that our Privacy UI offers end users more transparency and control of their privacy and correspondingly the personalized system.

# Chapter 6

# Conclusion and Future Work

## 6.1 Verification of Hypotheses

- **Hypothesis 1**: *If a web-based personalized system respects applicable privacy constraints for each user in its usage of personalization methods, and informs its users about this privacy-aware practice, then users will have higher regard for the privacy practices of the system.*

  This hypothesis is verified by the evidences we got from the user experiment (see Section 5.2.6). 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 between the two conditions in the statement "the new book website uses my data in a responsible manner" approached significance ($p<0.12$).

  Users' privacy concerns negatively impact the adoption of web personalization. If users perceive the personalized system's privacy practices better, then they are more likely to embrace and engage with web personalization.

- **Hypothesis 2**: *If a web-based personalized system respects applicable privacy constraints for each user in its usage of personalization methods, and informs its users about this privacy-aware practice, then users will disclose more information about themselves to the system.*

  This hypothesis is verified based on the results from the user experiment (see Section 5.2.6). Subjects in the "enhanced" condition answered significantly more questions (p=0.012) than their counterparts in the "control" condition. Similarly, subjects in the "enhanced" condition provided significantly more answers (p=0.017) than their counterparts in the "control" condition.

  This finding has important implications for web personalization. In general, the more information users disclose about themselves, the better personalized systems know about the users, and then the better quality of the personalized services.

- **Hypothesis 3**: *If a web-based personalized system respects applicable privacy constraints for each user in its usage of personalization methods, and informs its users about this privacy-aware practice, then users will be more likely to exhibit other privacy-sensitive behaviors, e.g., make online purchases.*

  While we do not have strong evidence to verify this hypothesis, we suspect there is a trend and we would observe statistically significant results if we had much more subjects. In our experiment, the purchase ratio in the "enhanced" condition is 60% higher than in the "control" condition (note again that all subjects saw the same set of 50 books in both conditions). This difference approaches significance (p<0.09). (see Section 5.2.6).

  This hypothesis has important practical implications. In order to make purchases, subjects have to provide their accurate credit card or debit card information (including card number, security number, expiration date) as well as their shipping address and e-mail address. These pieces of information are considered highly personal and sensitive (Ackerman et al., 1999). We also checked the accuracy of the financial information they provided after the experiment (we asked them show their credit/debit cards again and compared with the information they entered in our system) and did not find any case of discrepancy. This finding suggests that our privacy enhancement mechanism may positively affect users' online purchase behavior, which is of great interests to E-commerce sites or applications.

- **Hypothesis 4**: *If a web-based personalized system respects applicable privacy constraints for each user in its usage of personalization methods, and informs its users about this privacy-aware practice, then this will not compromise users' perceived personalization benefits.*

  This hypothesis is important because the goal of this research is to reconcile privacy and personalization. If the privacy-aware user modeling does unduly compromise users' perceived benefits of web personalization, then it fails to strike a good balance between privacy and personalization. In our experiment, subjects' perceptions of whether their data disclosure helped the bookstore in selecting interesting books for them were not significantly different between the two conditions (p=0.36). Strictly speaking, we failed to reject the null hypothesis (privacy-aware user modeling does compromise how users perceive the benefits of web personalization) and nor can we confirm Hypothesis 4. Practically speaking, we can say that we did not find evidence that falsifies Hypothesis 4. The difference in perceived benefits is too small to be statistically significant. Future research (e.g., large-scale experiments) needs to further investigate this hypothesis.

- **Hypothesis 5**: *Respecting applicable privacy constraints for each user in the usage of personalization methods is technically possible with reasonable computing resources even for contemporary personalized sites with heavy traffic.*

  We prototyped our privacy-aware user modeling framework (see Section 4.2.4) and show that it is technically possible. Our performance evaluation (see Section 5.1) shows that the average handling time per user session is less than 0.2 seconds with our performance enhancement (including light-weight PLA representation, computation distribution, and multi-level caching). We can distribute the handling of incoming user sessions to a network of machines in a Cloud computing environment. According to simplified calculations (see Section 5.1.3) based on average arrival rate of incoming user sessions, even for sites that have heavy traffic and/or a large number of users/visitors, e.g., Yahoo, our approach can scale with a reasonable number of networked machines (about 700 machines for Yahoo's average arrival rate of 350 users per second). Therefore, we believe that not only is our approach technically possible but also is scalable to meet with the demands of sites with highest traffic.

## 6.2 Summary of Contributions

The main contributions of this research are:

- Our analysis of privacy laws reveal that they affect the usage of user modeling methods used to make inferences about users in personalized systems.
- For the first time, individual users' privacy is treated as a first-class design requirement and expressed as part of the system design specifications for personalized systems.
- Our PLA-based framework provides a flexible, extensible, and enforceable approach in modeling and addressing privacy constraints in web personalization.
- Our evaluations of the framework show that users value user-tailored privacy enforcement in personalization, and that it is technically feasible.

## 6.3 Future Research Directions

Finally, we summarize the following future directions that we believe in:

- We advocate more recognition of the importance of privacy in web personalization research and practice, and argue that privacy needs be treated as first-class design requirements since (1) regulatory privacy requirements and users' privacy concerns have significant impacts on personalization and its possible benefits, and (2) privacy, like security and usability, is extremely difficult if not impossible to achieve after a system has already been built. Therefore, privacy should be taken into serious consideration from the early onsets of the development process. How do we develop good privacy processes, and weave them into the system development process from a software engineering standpoint? Can we develop tools to help support system designers and architects model, manage, and verify privacy requirements?

- We currently use a Boolean guard language to express privacy constraints. With regard to the expression of privacy constraints, two things are desirable. First, a formal language is needed that can sufficiently express potential privacy constraints. If it is a standard language (such as P3P (Cranor et al., 2006) or XACML (OASIS, 2005)), the adoption and integration with other systems would be easier. As discussed in Section 3.2, XACML seems to come close to this vision. Very recently, Adam Barth (Barth, 2008) proposed a formal language based on linear temporal logic to model the "Contextual Integrity" privacy framework (Nissenbaum, 2004). He showed that the language can express many legal privacy requirements found in the sector-specific privacy laws such as the Health Insurance Portability and Accountability Act of 1996 (HIPAA) (HHS, 1996) and Children's Online Privacy Protection Act of 1999 (COPPA) (FTC, 1999) in the US. One advantage of this language over other formal languages such as P3P is that it explicitly models the temporal aspect of privacy regulations. For instance, past and future are first-class modeling concept in this language. Further studies need to test this language by using it to express privacy constraints in other representative privacy laws such as the European Union directives and the German privacy laws. Secondly, potential privacy constraints should be captured and expressed as they arise, preferably in real time. Users' privacy concerns usually emerge as they interact with a web-based personalized system. Designers of privacy enhanced web personalization should not assume that users can and would express their privacy concern in a formal privacy language. A hybrid approach of "user empowerment" and "expression and enforcement" might be promising in which users become empowered to act on their contingent privacy needs and possibly also express them in a user-friendly fashion (e.g., in natural language). Thereafter, the system would compile this information into formal expressions that can be executed and enforced. Systematic enforcement is also largely neglected in privacy enhancement in web personalization. Solutions like the IBM Tivoli Privacy Manager (IBM, 2003a) need to be adopted.

- Our research focuses on the data processing aspect of personalized systems and its privacy implications (see Section 1.2.1). We assume that users' data has been lawfully collected. In the future, we will plan to build a comprehensive privacy-aware user modeling framework that covers both the data collection and data processing/reasoing perspective. We envision access control between the external user-adaptive systems and the user models such that user-adaptive applications may have access to different information from the same user model depending on their nature. For instance, a personalized healthcare application may retrieve a user's health-related information while a personalized banking applications can not.

- Users' privacy needs have been studied predominately in the domain of E-commerce. However, web personalization can also take place in, e.g., E-learning or Ubiquitous

Computing, and research is needed to uncover users' privacy needs in these domains as well. Besides, since users' privacy needs and preferences are inherently dynamic and contingent, users' *individual* privacy needs must be taken into account. Our studies of privacy attitudes and perceptions have been conducted through surveys and a lab experiment. In the future, we seek to study people's privacy attitudes and behaviors in their authentic, real-life situations just as Ubiquitous Computing researchers do (e.g., (Iachello et al., 2006)). Additional methodologies such as experience sampling (e.g., (Iachello et al., 2006)) and field deployment (e.g., (Hayes et al., 2008)) will be considered.

- Another promising future direction is usable personal privacy management tools that can help users manage and keep track of the disclosure and usage of their personal information (e.g., by indicating which organization knows what about the user and employs this information for what purposes). Besides, while compliance has long been technically framed and treated as a server-side problem, solutions that follow the user empowerment strategy (such as Personis (Czarkowski and Kay, 2002)) bear great potential. How to appropriately empower users in the context of web personalization is still an open question, e.g. in light of the fact that the users may not be technically savvy. Techniques such as visualization may be useful in this regard.

- In recent years, cross-system personalization (CSP) is gaining much traction both in academia and industry. CSP refers to "personalization that shares information across different systems in a user-centric way" (Mehta et al., 2005). In a converged service environment, CSP enables services or applications that adapt to each user based on the user's service consumption data from multiple service domains (e.g., music and news) and multiple service platforms (e.g., IPTV and mobile phone) (Aghasaryan et al., 2008). Imagine the personalized radio (e.g., Pandora) on your smart phone playing music that is (partially) based on what news and shows you watched on your

IPTV, and/or the Youtube videos you saw on your laptop. CSP has the potential to strengthen the benefits of personalization: further engage and retain end users, help select targeted ads, etc. However, since CSP usually relies on collecting, merging and mining user data gleaned from multiple applications/platforms, it may cause more intense privacy issues than general web personalization. We plan to study people's privacy attitudes and perceptions towards CSP and design novel privacy enhancement mechanisms for CSP.

# Bibliography

Ackerman, M. and Cranor, L. (1999). Privacy critics: UI components to safeguard users' privacy. In *CHI '99 Extended Abstracts on Human Factors in Computing Systems*, pages 258–259.

Ackerman, M. S. (2000). The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human-Computer Interaction*, 15(2):179–203.

Ackerman, M. S., Cranor, L. F., and Reagle, J. (1999). Privacy in e-commerce: Examining user scenarios and privacy preferences. In *First ACM Conference on Electronic Commerce*, pages 1–8, Denver, CO.

Aggarwal, G., Feder, T., Kenthapadi, K., Khuller, S., Panigrahy, R., Thomas, D., and Zhu, A. (2006). Achieving anonymity via clustering. In *Symposium on Principles of Database Systems*, pages 153 – 162.

Aghasaryan, A., Betgé-Brezetz, S., Senot, C., and Toms, Y. (2008). A profiling engine for converged service delivery platforms. *Bell Lab. Tech. J.*, 13(2):93–103.

Agrawal, R., Kiernan, J., Srikant, R., and Xu, Y. (2003). An XPath-based preference language for P3P. In *the 12th Int'l World Wide Web Conference*, pages 629 – 639, Budapest, Hungary.

Alexa (2009). Yahoo traffic details. `http://www.alexa.com/data/details/traffic_details/yahoo.com`.

Allen, J. F. (1979). A Plan-Based approach to speech act recognition. Technical Report 131/79, Dept. of Computer Science, University of Toronto.

Altman, I. (1975). *The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding.* Brooks/Cole Publishing Company, Monterey, California.

Anderson, A. (2005). A comparison of two privacy policy languages: EPAL and XACML. Technical report, Sun Microsystems Laboratories.

APEC (2005). APEC privacy framework. Report APEC#205-SO-01.2, Asia-Pacific Economic Cooperation.

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

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

Arlein, R. M., Jai, B., Jakobsson, M., Monrose, F., and Reiter, M. K. (2000). Privacy-Preserving global customization. In *2nd ACM Conference on Electronic Commerce*, pages 176–184, Minneapolis, MN. ACM Press.

Ashley, P., Hada, S., Karjoth, G., Powers, C., and Schunter, M. (2003). *Enterprise Privacy Authorization Language (EPAL 1.2). W3C Member Submission 10 November 2003*.

Ashrafi, N. and Kuilboer, J. (2005). Privacy protection via technology: Platform for privacy preferences (P3P). *International Journal of E-Business Research*, 1(2):56–69.

Atkinson, C., Bayer, J., Bunse, C., Kamsties, E., Laitenberger, O., Laqua, R., Muthig, D., Paech, B., Wüst, J., and Zettel, J. (2002). *Component-based product line engineering with UML*. Addison-Wesley Longman Publishing Co., Inc.

Barth, A. (2008). *Design and Analysis of Privacy Policies (Ph.D. Dissertation)*.

Batory, D. (2005). Feature models, grammars, and propositional formulas. In *SPLC'05: Proceedings of the 9th International Sofware Product Line Conference (SPLC 2005)*, pages 7–20.

Bell Labs Lucent Technologies (1997). Sablime v5.0 user's reference manual. Technical report.

Bellotti, V. and Sellen, A. (1993). Design for privacy in ubiquitous environments. In *The Third European Conference on Computer-Supported Cooperative Work (ECSCW'93)*, pages 77–92, Milan, Italy. Kluwer.

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

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

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

Bhole, Y. and Popescu, A. (2005). Measurement and analysis of http traffic. *Journal of Network and Systems Management*, 13(4):357–371.

Billsus, D. and Pazzani, M. (2007). Adaptive news access. In *The Adaptive Web*, pages 550–570.

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

Boyle, M. and Greenberg, S. (2005). The language of privacy: Learning from video media space analysis and design. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):328 – 370.

Brusilovsky, P. and Millán, E. (2007). User models for adaptive hypermedia and adaptive educational systems. In *The Adaptive Web*, pages 3–53.

Buffett, S., Jia, K., Liu, S., Spencer, B., and Wang, F. (2004). Negotiating exchanges of P3P-Labeled information for compensation. *Computational Intelligence*, 20(4):663–677.

Buyya, R., Yeo, C. S., and 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.*, pages 5–13. IEEE Computer Society.

Byers, S., Cranor, L., Kormann, D., and McDaniel, P. (2004). Searching for privacy: Design and implementation of a P3P-Enabled search engine. In *the 2004 Workshop on Privacy Enhancing Technologies (PET2004)*, pages 26–28, Toronto, Canada.

Cannon, J. C. (2005). *Privacy: What Developers and IT Professionals Should Know*. Addison-Wesley.

Canny, J. (2002a). Collaborative filtering with privacy. In *Proceedings of the 2002 IEEE Symposium on Security and Privacy*, page 45. IEEE Computer Society.

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*, pages 238–245, Tampere, Finland. ACM.

Carmichael, D. J., Kay, J., and 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.

Cawsey, A., Grasso, F., and Paris, C. (2007). Adaptive information for consumers of healthcare. In *The Adaptive Web*, pages 465–484.

Chappell, D. (2006). Introducting windows CardSpace. http://msdn.microsoft.com/en-us/library/aa480189.aspx.

Chaum, D. (1982). Blind signatures for untraceable payments. In *CRYPTO 82*, pages 199–203.

Chellappa, R. K. and Sin, R. G. (2005). Personalization versus privacy: An empirical examination of the online consumer's dilemma. *Inf. Technol. and Management*, 6(2-3):181–202.

Chen, P., Critchlow, M., Garg, A., van der Westhuizen, C., and van der Hoek., A. (2003). Differencing and merging within an evolving product line architecture. In *The Fifth International Workshop on Product Family Engineering*, pages 269–281, Siena, Italy.

Chlebus, E. and Brazier, J. (2007). Nonstationary poisson modeling of web browsing session arrivals. *Information Processing Letters*, 102(5):187–190.

ChoiceStream (2005). ChoiceStream personalization survey: Consumer trends and perceptions. Technical report.

Christensen, L. (1988). Deception in psychological research: When is its use justified? *Pers Soc Psychol Bull*, 14(4):664–675.

Clark, J. and DeRose, S. (1999). *XML Path Language (XPath) Version 1.0, W3C Recommendation 16 November 1999.*

Clements, P., Northrop, L., and Northrop, L. (2001). *Software Product Lines : Practices and Patterns*. Addison-Wesley Professional.

ConneXion, C. (1996). *Anonymous Surfing*.

Conradi, R. and Westfechtel, B. (1998). Version models for software configuration management. *ACM Computing Surveys*, 30(2):232–282.

Cooley, T. (1888). *Cooley on Torts*. 2nd edition.

Cooperstein, D., Delhagen, K., Aber, A., and Levin, K. (1999). Making net shoppers loyal. Technical report, Forrester Research.

Cottenier, T., van den Berg, A., and Elrad, T. (2006). The motorola weavr: Model weaving in a large industrial context. *International Conference on Aspect-Oriented Software Development, Industry Track*, Vancouver, Canada.

Coyle, K. (1999). P3P: pretty poor privacy? a social analysis of the platform for privacy preferences (P3P). http://www.kcoyle.net/p3p.html.

Cranor, L., Dobbs, B., Egelman, S., Hogben, G., Humphrey, J., Langheinrich, M., Marchiori, M., Presler-Marshall, M., Reagle, J., Schunter, M., Stampley, D. A., and Wenning, R. (2006). *The Platform for Privacy Preferences 1.1 (P3P1.1) Specification. W3C Working Draft 10 February 2006.*

Cranor, L., Langheinrich, M., and Marchiori, M. (2002). *A P3P Preference Exchange Language 1.0 (APPEL1.0): W3C Working Draft 15 April 2002.*

Cranor, L. F. (2003). 'I didn't buy it for myself': Privacy and ecommerce personalization. In *2003 ACM Workshop on Privacy in the Electronic Society*, Washington, DC. ACM Press.

Cranor, L. F. and Reidenberg, J. R. (2002). Can user agents accurately represent privacy notices? In *30th Research Conference on Communication, Information and Internet Policy*, Alexandria, VA.

Culnan, M. J. and Bies, R. J. (2003). Consumer privacy: Balancing economic and justice considerations. *Journal of Social Issues*, 59(2):323–342.

CZ (2000). *Czech Republic Act of 4 April 2000 on the Protection of Personal Data and on Amendment to Some Related Acts*, volume 101.

Czarkowski, M. and Kay, J. (2002). A scrutable adaptive hypertext. In *Proceedings of the Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 384–387. Springer-Verlag.

Dashofy, E., Asuncion, H., Hendrickson, S., Suryanarayana, G., Georgas, J., and Taylor, R. (2007). Archstudio 4: An architecture-based meta-modeling environment. In *ICSE 2007: Intl. Conf. on Softw. Eng.*, pages 67–68. IEEE Computer Society.

Dashofy, E. M., van der Hoek, A., and Taylor, R. N. (2005). A comprehensive approach for the development of XML-Based software architecture description languages. *ACM Transactions on Software Engineering and Methodology*, 14(2):199–245.

DE (2006). Federal data protection act, as of 15 nov. 2006. Technical report.

de Paula, R., Ding, X., Dourish, P., Nies, K., Pillet, B., Redmiles, D. F., Ren, J., Rode, J. A., and Filho, R. S. (2005). In the eye of the beholder: A visualization-based approach to information system security. *International Journal of Human-Computer Studies*, 63(1-2):5–24.

DE-TML (2007). *German Telemedia Law, as of 1 March 2007.*

DePallo, M. (2000). *AARP National Survey on Consumer Preparedness and E-Commerce: A Survey of Computer Users Age 45 and Older*. AARP.

DiGioia, P. and Dourish, P. (2005). Social navigation as a model for usable security. In *Symposium on Usable Privacy and Security SOUPS 2005*, pages 101–108, Pittsburgh, PA.

Disney (2002). Personal communication, chief privacy officer, disney corporation.

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

Dourish, P., Grinter, R. E., Dalal, B., de la Flor, J. D., and Joseph, M. (2004). Security Day-to-Day: user strategies for managing security as an everyday, practical problem. *Personal Ubiquitous Computing*, 8(6):391–401.

Earp, J. B. and Baumer, D. (2003). Innovative web use to learn about consumer behavior and online privacy. *Commun. ACM*, 46(4):81–83.

Egelman, S., Cranor, L., and Chowdhury, A. (2006). An analysis of P3P-Enabled web sites among top-20 search results. In *the 8th international conference on Electronic commerce*, pages 197 – 207.

EPIC and Junkbusters (2000). *Pretty Poor Privacy: An Assessment of P3P and Internet Privacy*. Electronic Privacy Information Center and Junkbusters.

Estublier, J., Leblang, D., van der Hoek, A., Conradi, R., Clemm, G., Tichy, W., and Wiborg-Weber, D. (2005). Impact of software engineering research on the practice of software configuration management. *ACM Trans. Softw. Eng. Methodol.*, 14(4):383–430.

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.

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.

Forrester Research (1999). The privacy best practice. Technical report, Forrester Research.

FTC (1973). Fair information practice principles.

FTC (1998). Privacy online: A report to congress.

FTC (1999). Children's online privacy protection act of 1999.

FTC (2000a). Online profiling: A report to congress. Technical report.

FTC (2000b). *Privacy Online: Fair Information Practices in the Electronic Marketplace. A Report to Congress*. Federal Trade Commission. Five main principals: (1)Notice/Awareness; (2) Choice/Consent; (3) Access/Participation; (4) Integrity/Security; and (5) Enforcement/Redress.

FTC (2000c). *The Seven Safe Harbor Principles*. Federal Trade Commission.

Gabber, E., Gibbons, P. B., Matias, Y., and Mayer, A. (1997). How to make personalized web browsing simple, secure, and anonymous. In *Financial Cryptography'97*, volume 1318 of *Lecture Notes in Computer Science*, pages 17–31. Springer Verlag, Berlin - Heidelberg - New York.

Garfinkel, S. and Cranor, L. (2002). *P3P: Privacy Primer*. O'Reilly Network.

Garg, A., Critchlow, M., Chen, P., Westhuizen, C. V. d., and Hoek, A. v. d. (2003). An environment for managing evolving product line architectures. In *ICSM '03: Proceedings of the International Conference on Software Maintenance*, pages 358–367, Washington, DC, USA. IEEE Computer Society.

Georgas, J. C., van der Hoek, A., and Taylor, R. N. (2005). Architectural runtime configuration management in support of dependable Self-Adaptive software. In *The 2005 Workshop on Architecting Dependable Systems*, pages 1–6, St. Louis, Missouri. ACM Press.

Gideon, J., Cranor, L., Egelman, S., and Acquisti, A. (2006). Power strips, prophylactics, and privacy, oh my! In *Second Symposium on Usable Privacy and Security*, pages 133–144, Pittsburgh, Pennsylvania. ACM Press.

Gilburd, B., Schuster, A., and Wolff, R. (2004). k-TPP: a new privacy model for Large-Scale distributed environments. In *10th ACM SIGKDD International Conference on Knowledge Discovery and Date Mining (KDD'04)*, pages 563–568, Seattle, WA.

Goldberg, I. (1997). Privacy-enhancing technologies for the internet, II: five years later. pages 103—109.

Goy, A., Ardissono, L., and Petrone, G. (2007). Personalization in E-Commerce applications. In *The Adaptive Web*, pages 485–520.

Gülcü, C. and Tsudik, G. (1996). Mixing email with BABEL. In *the 1996 Symposium on Network and Distributed System Security (SNDSS '96)*, page 2. IEEE Computer Society.

Hagen, P. R., Manning, H., and Souza, R. (1999). Smart personalization. Technical report, Forrester Research.

Hayes, G. R., Gardere, L. M., Abowd, G. D., and Truong, K. N. (2008). CareLog: a selective archiving tool for behavior management in schools. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 685–694, Florence, Italy. ACM.

Hendrickson, S. A. and van der Hoek, A. (2007). Modeling product line architectures through change sets and relationships. In *ICSE '07: Proceedings of the 29th international conference on Software Engineering*, pages 189–198, Washington, DC, USA. IEEE Computer Society.

Herlocker, J. and Konstan, J. (2001). Content-Independent Task-Focused recommendation. *IEEE Internet Computing*, 5(6):40 – 47.

HHS (1996). Health insurance portability and accountability act of 1996.

Hine, C. and Eve, J. (1998). Privacy in the marketplace. *The Information Society*, 14:253–262.

Hitchens, M., Kay, J., Kummerfeld, B., and Brar, A. (2005). Secure identity management for Pseudo-Anonymous service access. In Hutter, D. and Ullmann, M., editors, *Security in Pervasive Computing: Second International Conference*, pages 48–55, Boppard, Germany.

Hof, R., Green, H., and Himmelstein, L. (1998). Now it's YOUR WEB. *Business Week*, October 5:68–75.

Iachello, G., Truong, K. N., Abowd, G. D., Hayes, G. R., and Stevens, M. (2006). Prototyping and sampling experience to evaluate ubiquitous computing privacy in the real world. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 1009–1018, Montréal, Québec, Canada. ACM.

IBM (1999). IBM Multi-National consumer privacy survey. Technical report, IBM.

IBM (2003a). *IBM Tivoli Privacy Manager for E-Business*.

IBM (2003b). Personal communication, chief privacy officer, IBM zurich.

Ishitani, L., Almeida, V., and Wagner, M. (2003). Masks: Bringing anonymity and personalization together. *IEEE Security & Privacy Magazine*, 1(3):18–23.

Karat, J., Karat, C., Brodie, C., and Feng, J. (2005). Privacy in information technology: Designing to enable privacy policy management in organizations. *International Journal of Human-Computer Studies*, 63:153–174.

Kass, R. (1989). Student modeling in intelligent tutoring systems – implications for user modeling. In Kobsa, A. and Wahlster, W., editors, *User Models in Dialog Systems*, pages 386–410. Springer-Verlag, Berlin, Heidelberg.

Kay, J. (2006). Scrutable adaptation: Because we can and must. In *Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 11–19. Springer Berlin / Heidelberg.

Kay, J., Kummerfeld, B., and Lauder, P. (2003). Managing private user models and shared personas. In *Workshop on User Modelling for Ubiquitous Computing, 9th International Conference on User Modeling*, pages 1–11.

Kelley, P. G. (2009). Designing a privacy label: assisting consumer understanding of online privacy practices. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, pages 3347–3352, Boston, MA, USA. ACM.

Knuth, D. E. (1964). backus normal form vs. backus naur form. *Commun. ACM*, 7(12):735–736.

Kobsa, A. (1990). Modeling the user's conceptual knowledge in BGP-MS, a user modeling shell system. *Computational Intelligence*, 6(4):193–208.

Kobsa, A. (2001). Generic user modeling systems. *User Modeling and User-Adapted Interaction*, 11(1-2):49–63.

Kobsa, A. (2002). Personalization and international privacy. *Communications of the ACM*, 45(5):64–67.

Kobsa, A. (2003). A component architecture for dynamically managing privacy constraints in personalized Web-Based systems. In *Privacy Enhancing Technologies*, pages 177–188.

Kobsa, A. (2007a). Generic user modeling systems. In Brusilovsky, P., Kobsa, A., and Nejdl, W., editors, *The Adaptive Web: Methods and Strategies of Web Personalization*, pages 136–154. Springer Verlag, Heidelberg, Germany.

Kobsa, A. (2007b). Privacy-enhanced web personalization. In Brusilovsky, P., Kobsa, A., and Nejdl, W., editors, *The Adaptive Web: Methods and Strategies of Web Personalization*, pages 628–670. Springer-Verlag.

Kobsa, A. and Fink, J. (2003). Performance evaluation of user modeling servers under real-world workload conditions. In Brusilovsky, P., Corbett, A. T., and Rosis, F. d., editors, *User Modeling 2003: 9th Intl. Conf.*, pages 143–153. Springer Verlag.

Kobsa, A. and Fink, J. (2006). An LDAP-based user modeling server and its evaluation. *User Modeling and User-Adapted Interaction*, 16(2):129–169.

Kobsa, A., Koenemann, J., and Pohl, W. (2001). Personalized hypermedia presentation techniques for improving online customer relationships. *THE KNOWLEDGE ENGINEERING REVIEW*, 16:111–155.

Kobsa, A. and Schreck, J. (2003). Privacy through pseudonymity in User-Adaptive systems. *ACM Transactions on Internet Technology*, 3(2):149–183.

Kobsa, A. and Teltzrow, M. (2005). Contextualized communication of privacy practices and personalization benefits: Impacts on users' data sharing and purchase behavior. In Martin, D. and Serjantov, A., editors, *Privacy Enhancing Technologies: Fourth International Workshop, PET 2004, Toronto, Canada*, volume LNCS 3424, pages 329–343. Springer Verlag, Heidelberg, Germany.

Kobsa, A. and Wahlster, W. (1989). *User Models in Dialog Systems*. Symbolic Computation. Springer-Verlag, Berlin.

Langheinrich, M. (2002). A privacy awareness system for ubiquitous computing environments. In *the 4th International Conference on Ubiquitous Computing*, pages 237–245, Göteborg, Sweden.

LaRose, R. and Rifon, N. J. (2006). Your privacy is assured—of being disturbed: Comparing web sites with and without privacy seals. *New Media and Society*, 8(6):1009–1029.

Magee, J. and Kramer, J. (1996). Dynamic structure in software architectures. In *The 4th ACM SIGSOFT Symposium on Foundations of Software Engineering*, pages 3–14, San Francisco, California, United States. ACM Press.

Malin, B., Sweeney, L., and 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.

May Lwin, Jochen Wirtz, J. D. W. (2007). Consumer online privacy concerns and responses: a power–responsibility equilibrium perspective. *Journal of the Academy of Marketing Science*, 35(4):572–585.

Medvidovic, N. and Taylor, R. N. (2000). A classification and comparison framework for software architecture description languages. *IEEE Transactions on Software Engineering*, 26(1):70–93.

Mehta, B., Niederee, C., Stewart, A., Degemmis, M., Lops, P., and Semeraro, G. (2005). Ontologically-Enriched unified user modeling for Cross-System personalization. In *User Modeling 2005*, pages 119–123.

Mens, T. (2002). A state-of-the-art survey on software merging. *IEEE Transactions on Software Engineering*, 28(5):449–462.

Micarelli, A., Gasparetti, F., Sciarrone, F., and Gauch, S. (2007). Personalized search on the world wide web. In *The Adaptive Web*, pages 195–230.

Microsoft (2000). *Microsoft Announces Privacy Enhancements for Windows, Internet Explorer*. Number 21 June 2000. Microsoft Corporation.

Miller, B., Konstan, J., and Riedl, J. (2004). PockLens: toward a personal recommender system. *ACM Transactions on Information Systems*, 22(3):437–476.

Movielens (1997). Movielens - movie recommendations. http://www.movielens.org/.

NAI (2006). *Self-Regulatory Principles for Online Preference Marketing by Network Advisers*. Network Advertising Initiative.

Nakashima, E. (2006). *AOL Search Queries Open Window Onto Users' Worlds*. washingtonpost.com.

Nguyen, D. H., Marcu, G., Hayes, G. R., Truong, K. N., Scott, J., Langheinrich, M., and Roduner, C. (2009). Encountering SenseCam: personal recording technologies in everyday life. In *Proceedings of the 11th international conference on Ubiquitous computing*, pages 165–174, Orlando, Florida, USA. ACM.

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review Association*, 79:119–158.

OASIS (2005). *eXtensible Access Control Markup Language (XACML), Version 2.0; OASIS Standard, February 1, 2005.*

OECD (1980). *Recommendation of the Council Concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*. OECD.

Olsson, T. (1998). Decentralised social filtering based on trust. In *The AAAI-98 Recommender Systems Workshop (Working Notes)*, pages 84–88, Madison, Wisconsin.

OpenID (2006). *OpenID: An Actually Distributed Identity System.*

Oreizy, P., Medvidovic, N., and Taylor, R. N. (1998). Architecture-based runtime software evolution. In *The 20th International Conference on Software Engineering*, pages 177–186, Kyoto, Japan. IEEE Computer Society.

P3PEdit (2001). P3PEdit. https://p3pedit.com/.

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

Perry, D. E. and Wolf, A. L. (1992). Foundations for the study of software architecture. *ACM SIGSOFT Software Engineering Notes*, 17(4):40–52.

Personalization Consortium (2001). New survey shows consumers are more likely to purchase at web sites that offer personalization: Consumers willing to provide personal information in exchange for improved service and benefits. http://www.personalization.org/pr050901.html.

Pew (2008). Privacy implications of fast, mobile internet access.

Polat, H. and Du, W. (2003). Privacy-Preserving collaborative filtering using randomized perturbation techniques. In *IEEE International Conference on Data Mining*, volume 0, page 625, Los Alamitos, CA, USA. IEEE Computer Society.

Polat, H. and Du, W. (2005a). Privacy-Preserving collaborative filtering. *The International Journal of Electronic Commerce (IJEC)*, 9(4):9–35.

Polat, H. and Du, W. (2005b). SVD-based collaborative filtering with privacy. In *Proceedings of the 2005 ACM symposium on Applied computing*, pages 791–795, Santa Fe, New Mexico. ACM.

Preibusch, S. (2006). Personalized services with negotiable privacy policies. In Spiekermann, S., editor, *PEP06, CHI 2006 Workshop on Privacy-Enhanced Personalization*, pages 29–38, Montreal, Canada.

Rao, J. R. and Rohatgi, P. (2000). Can pseudonymity really guarantee privacy? In *9th USENIX Security Symposium*, page 85–96.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. (1994). GroupLens: an open architecture for collaborative filtering of netnews. In *ACM Conference on Computer Supported Cooperative Work*, pages 175–186, Chapel Hill, North Carolina, United States.

Rosenblatt, W., Trippe, W., and Mooney, S. (2001). *Digital Rights Management: Business and Technology*. Hungry Minds, Inc., Indianapolis, IN.

SEC (2002). Sarbanes–Oxley act of 2002.

SEI (2009). Software product line hall of fame. http://splc.net/fame.html.

Sinnema, M., Deelstra, S., Nijhuis, J., and Bosch, J. (2004). Covamof: A framework for modeling variability in software product families. In Nord, R. L., editor, *Third International Software Product Lines Conference (SPLC 2004)*, pages 197–213, Boston, MA, USA. Springer Berlin / Heidelberg.

Smith, H. J. (2001). Information privacy and marketing: WHAT THE U.S. SHOULD (AND SHOULDN'T) LEARN FROM EUROPE. *California Management Review*, 43(2):8–33.

Smith, H. J. (2004). Information privacy and its management. *MIS Quarterly Executive*, 3(4):201–213.

Smith, H. J., Milberg, S. J., and Burke, S. J. (1996). Information privacy: Measuring individuals' concerns about organizational practices. *MIS Quarterly*, 20(2):167–196. ArticleType: primary_article / Full publication date: Jun., 1996 / Copyright © 1996 Management Information Systems Research Center, University of Minnesota.

Soller, A. (2007). Adaptive support for distributed collaboration. In *The Adaptive Web*, pages 573–595.

Solove, D. J. (2006). A taxonomy of privacy. *University of Pennsylvania Law Review*, 154(3):477–560.

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

Spiekermann, S., Grossklags, J., and Berendt, B. (2001b). Stated privacy preferences versus actual behaviour in EC environments: a reality check. In *WI-IF 2001: the 5th International Conference Wirtschaftsinformatik - 3rd Conference Information Systems in Finance*, pages 129–148, Augsburg, Germany.

Stufflebeam, W., Anton, A. I., He, Q., and Jain, N. (2004). Specifying privacy policies with P3P and EPAL: lessons learned. In *the 2004 ACM Workshop on Privacy in the Electronic Society*, pages 35–36.

Sweeney, L. (2002). K-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570.

Tam, K. Y. and Ho, S. Y. (2003). Web personalization: Is it effective? *IT Professional*, 5(5):53–57.

Taylor, R.N., e. a. (1996). A component- and message-based architectural style for GUI software. *IEEE Trans. Softw. Eng.*, 22(6):390–406.

Teltzrow, M. and Kobsa, A. (2004). Impacts of user privacy preferences on personalized systems: a comparative study. In *Designing personalized user experiences in eCommerce*, pages 315–332. Kluwer Academic Publishers.

Tor (2004). Tor: anonymity online. http://www.torproject.org/.

Tsai, J., Egelman, S., Cranor, L., and Acquisti, A. (2007). The effect of online privacy information on purchasing behavior: An experimental study. In *Sixth Workshop on the Economics of Information Security*, Pittsburgh, PA.

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

US (1970). Fair credit reporting act of 1970.

US (1974). Privacy act of 1974.

US (1999). Gramm-Leach-Bliley act of 1999.

USA (2002). *The E-Government Act*.

USACM (2006). USACM policy recommendations on privacy. Technical report, U.S. Public Policy Committee of the Association for Computing Machinery.

van der Hoek, A. (2004). Design-Time product line architectures for Any-Time variability. *Science of Computer Programming, special issue on Software Variability Management*, 53(30):285–304.

van der Hoek, A., Mikic-Rakic, M., Roshandel, R., and Medvidovic, N. (2001). Taming architectural evolution. In *The Sixth European Software Engineering Conference (ESEC) and the Ninth ACM SIGSOFT Symposium on the Foundations of Software Engineering (FSE-9)*, pages 1–10, Vienna, Austria.

van Ommering, R., van der Linden, F., Kramer, J., and Magee, J. (2000). The koala component model for consumer electronics software. *Computer*, 33(3):78–85.

Waldo, J., Lin, H., and Millett, L. I. (2007). *Engaging privacy and information technology in a digital age*. National Academies Press.

Wang, H., Lee, M. K. O., and Wang, C. (1998). Consumer privacy concerns about internet marketing. *Commun. ACM*, 41(3):63–70.

Wang, Y., Chen, Z., and Kobsa, A. (2006a). A collection and systematization of international privacy laws, with special consideration of internationally operating personalized websites. `http://www.ics.uci.edu/~kobsa/privacy`.

Wang, Y., Hendrickson, S. A., van der Hoek, A., and Taylor, R. N. (2009). Modeling PLA variation of Privacy-Enhancing personalized systems. pages 71–80, San Francisco, CA.

Wang, Y. and Kobsa, A. (2006). Impacts of privacy laws and regulations on personalized systems. In Kobsa, A., Chellappa, R. K., and Spiekermann, S., editors, *PEP06, CHI06 Workshop on Privacy-Enhanced Personalization*, pages 44–46, Montréal, Canada.

Wang, Y. and Kobsa, A. (2007). Respecting users' individual privacy constraints in web personalization. In Conati, C., McCoy, K. F., and Paliouras, G., editors, *User Modeling 2007: 11th Intl. Conf.*, pages 157–166. Springer.

Wang, Y. and Kobsa, A. (2009). Privacy-enhancing technologies. In Gupta, M. and Sharman, R., editors, *Social and Organizational Liabilities in Information Security*, pages 203–227. IGI Global.

Wang, Y., Kobsa, A., van der Hoek, A., and White, J. (2006b). PLA-based runtime dynamism in support of Privacy-Enhanced web personalization. In *10th International Software Product Line Conference*, pages 151–162, Baltimore, MD. IEEE Press.

Warren, S. D. and Brandeis, L. D. (1890). The right to privacy. *Harvard Law Review*, 4(5):193–220.

Webb, G. I., Pazzani, M. J., and Billsus, D. (2001). Machine learning for user modeling. *User Modeling and User-Adapted Interaction*, 11(1-2):19–29.

Westin, A. (1967). *Privacy and Freedom*. Atheneum, New York.

Westin, A. and van Gelder, V. (2003). *Privacy & American Business: Special Issue on Consumer Privacy in Japan and the New National Privacy Law*.

Whitten, A. and Tygar, D. (1999). Why johnny can't encrypt: A usability evaluation of PGP 5.0. In *Ninth USENIX Security Symposium*.

Xie, E., Teo, H., and 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.

Xu, H. (2009). Consumer responses to the introduction of privacy protection measures: An exploratory research framework. *International Journal of E-Business Research (Special Issue on the Protection of Privacy in E-Business)*, 5(2):21–47.

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

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

# Appendices

## A Questions Asked in the Experiment

Page 1 of 8:

We offer you a selection of more than one million books. If you give us some information about yourself, we can adapt this selection to your needs.

1. Please enter a login name (your name or a pseudonym)
   - Login name ___________
   - No answer
2. How old are you?
   - 18-20
   - 21-25
   - 26-30
   - 31-35
   - 36-40
   - 41-50

- 51-60
- >60
- No answer

3. What is your occupation / degree program?

   - Occupation / degree program ___________
   - No answer

4. What are your hobbies? (Check all that apply.)

   - Sport
   - Music
   - Model making
   - Computers
   - Other, please specify ___________
   - No answer

   Page 2 of 8:

5. We would like to use cookies to record the order in which you visit our web pages (click stream). Do you agree with this?

   - Yes
   - No
   - No answer

6. Please enter your favorite author or book title:

   - Favorite author / book title ___________
   - No answer

7. When buying a book, do you pay more attention to the text on the back cover or to the name of the author?

    - Text on the back cover (short description)
    - Author
    - No answer

8. Please indicate which categories interest you most. (Check all that apply.)

    - Antique books
    - Bestseller
    - Business & Career
    - Stocks & Money
    - Computer & Internet
    - E-Books
    - Textbooks
    - Film
    - Culture & Comics
    - Mind & Knowledge
    - Audio books
    - Books for young people
    - Cooking & Lifestyle
    - Mysteries & Thrillers
    - Learning & Reference
    - Sheet Music
    - Science & Technology

- Politics
- Biographies & History
- How-to books & Self-help books
- Travel & Sports
- Religion & New Age
- Science Fiction
- Fantasy & Horror
- No Answer

Page 3 of 8:

9. Which kinds of books do you buy particularly often? (Check all that apply.) Note: Please rank them from 1 (most frequently bought) to 4 (least frequently)

    - Fiction
    - Scientific books
    - Technical books
    - Biographies
    - No Answer

10. Would you like to buy a book that complements your past preferences, or rather something completely different?

    - Something that complements my past preferences
    - Something completely different
    - No Answer

11. Do you borrow books from the library or from friends or colleagues?

- Yes
- No
- No Answer

12. What kind of vacation do you especially like? (Check all that apply.)

    - City trips
    - Beach vacations
    - Family trips
    - Adventure tours
    - Others, please specify ________________
    - No Answer

    Page 4 of 8:

13. Do you go on package vacations or rather on individual travel?

    - Package vacations
    - Individual travel
    - Neither
    - No Answer

14. Do you learn the basics of the local language before traveling to another country?

    - Yes
    - No
    - No Answer

15. Which nationality of authors do you prefer? (Check all that apply.)

- American
- English
- German
- French
- Spanish
- Other
- No Answer

16. Which political orientation do you prefer in the books you read?

    - Left-winged 1_____7 Right-winged
    - Not interested in politics
    - No Answer

    Page 5 of 8:

17. Do you like Marx or Machiavelli?

    - Marx
    - Machiavelli
    - I do not know
    - I like both
    - No Answer

18. Do you own health books?

    - Yes
    - No
    - No Answer

19. For which health topics do you seek answers? (Check all that apply.)

    - Allergies
    - Colds
    - Skin diseases
    - Chronic diseases
    - Venereal diseases
    - Other
    - No Answer

20. Are you interested in books on self-medication?

    - Yes
    - No
    - No Answer

    Page 6 of 8:

21. Are you also interested in alternative healing?

    - Yes
    - No
    - No Answer

22. Which faith interests you in books on religion?

    - Buddhism
    - Christianity
    - Hinduism
    - Islam

- Judaism
- Other
- No Answer

23. What religion do you believe in?
    - Buddhism
    - Christianity
    - Hinduism
    - Islam
    - Judaism
    - Other
    - I have no religious belief
    - No Answer

24. Do you like love stories?
    - Yes
    - No
    - No Answer

Page 7 of 8:

25. Which kind of erotic books do you like? (Check all that apply.)
    - Man/woman
    - Man/man
    - Woman/woman
    - No Answer

26. Why do you buy books? (Check all that apply.)

- For entertainment
- For advanced training
- To promote my career
- As a pastime
- To look up things
- No Answer

27. Which types of narrative do you like most?

- Short stories
- Fables
- Conversational novels
- No Answer

28. Do you like hardcover books or paperbacks?

- Hardcover books
- Paperbacks
- No Answer

Page 8 of 8:

29. How many books do you read per year?

- 1-3
- 4-5
- 6-10
- 11-15

- More than 15
- No Answer

30. How much money per year do you usually spend on books?

    - \$1 - \$50
    - \$51 - \$100
    - \$101 - \$200
    - \$201 - \$500
    - \$501 - \$1000
    - More than \$1000
    - No Answer

31. How much do you earn per month?

    - \$0 - \$500
    - \$501 - \$1000
    - \$1001 - \$1500
    - \$1501 - \$2000
    - More than \$2000
    - No Answer

32. Are you interested in good deals?

    - Yes
    - No
    - No Answer

# B Post-Experiment Questionnaire

Please tell us about your impressions of the website that you worked with:

(please circle one number for each item to show whether you agree or disagree with the statement):

1 (Strongly Disagree), 2 (Disagree), 3 (Neither Agree or Disagree), 4 (Agree), 5 (Strongly Agree)

1. I was satisfied with the book recommendations that I received from the new Amazon website.
2. I felt that my data are in good hands at the new Amazon website.
3. The data that I provided helped the new Amazon website select interesting books for me.
4. I understood how the new Amazon website used the data that I provided.
5. The new Amazon website handles my data in a responsible manner.
6. The new Amazon website assigns high priority to data protection.
7. I find the new Amazon website reliable.
8. The new Amazon website was able to provide me with book recommendations that I liked.
9. I am familiar with the web site of amazon.com.
10. I have previously purchased products from amazon.com.
11. If you did not buy any book, why not? (Check all that apply.)

- None of the books was interesting to me
- The books are still too expensive
- I didn't want to give away my personal data (e.g. address)
- I didn't want to give away my credit card number
- I don't read books anyway
- Others, please specify______________

12. The bookstore has a question if you decided to buy a book. The book is

    - for me
    - for someone else

    **Note that subjects in the "control" condition did not see Question 17 - 22, while subjects in the "enhanced" condition did.**

13. Did you pay attention to the privacy control panel on the right-hand side of the screen during the experiment?

    - Yes
    - No
    - I don't remember

14. Did you click at one or more of the blue, round "information" icons in the privacy control panel, to obtain more information on privacy preferences or personalization methods?

    - Yes
    - No
    - I don't remember

15. Did you check or uncheck one or more of the options in the privacy control panel, IN ORDER TO CHANGE YOUR PRIVACY PREFERENCES?

    - Yes
    - No
    - I don't remember

16. Did you check or uncheck one or more of the options in the privacy control panel, IN ORDER TO TRY OUT WHAT HAPPENS?

    - Yes
    - No
    - I don't remember

Please tell us whether you agree or disagree with the following statements.

(please circle one number for each item to show whether you agree or disagree with the statement):

1 (Strongly Disagree), 2 (Disagree), 3 (Neither Agree or Disagree), 4 (Agree), 5 (Strongly Agree)

17. I find controlling personalization methods with a privacy control panel useful in general.
18. I would myself use such a privacy control panel if a website offers one.
19. It usually bothers me when companies ask me for personal information.
20. All the personal information in computer databases should be double-checked for accuracy no matter how much this costs.

21. Companies should not use personal information for any purpose unless they have been authorized by the individuals who provided the information.

22. Companies should devote more time and effort to preventing unauthorized access to personal information.

23. When companies ask me for personal information, I sometimes think twice before providing it.

24. Companies should take more steps to make sure that the personal information in their files is accurate.

25. When people give personal information to a company for some reason, the company should never use the information for any other reason.

26. Companies should have better procedures to correct errors in personal information.

27. Computer databases that contain personal information should be protected from unauthorized access no matter how much it costs.

28. It bothers me to give personal information to so many companies.

29. Companies should never sell the personal information in their computer databases to other companies.

30. Companies should devote more time and effort to verifying the accuracy of the personal information in their databases.

31. Companies should never share personal information with other companies unless they have been authorized by the individuals who provided the information.

32. Companies should take more steps to make sure that unauthorized people cannot access personal information in their computers.

33. I'm concerned that companies are collecting too much personal information about me.

34. I consider myself a computer-savvy person.

Demographics: We would like to know just a little about you so we can see how different types of people feel about the issues we have been examining.

35. What is your gender?

    - Male
    - Female

36. Which of following best describe you?

    - Undergraduate student
    - Graduate student
    - University faculty/researcher/scientist
    - University staff

37. What is the highest level of education you've completed?

    - Less than high school
    - High school or equivalent (12th grade)
    - Technical degree (2 year)
    - Some college/associates degree
    - College Degree/Bachelor's degree (4 year)
    - Master's degree
    - PhD/Doctorate degree

38. How would you categorize your ethnical origin? (Choose the one that best applies.)

- African American/Black
- American Indian/Alaska Native
- Chinese/Chinese-American
- Mexican/Mexican-American/Chicano
- Pacific Islander (including Micronesia, Polynesian, other Pacific Islanders)
- Other Asian
- Other Spanish-American/Latino
- Filipino/Filipino-American
- Japanese/Japanese-American
- Korean/Korean-American
- White/Caucasian
- Other, please specify____________________________

39. How long have you been living in US?

- Less than 1 year
- 1 year – 2 years
- 2 years – 3 years
- 3 years – 4 years
- 4 years – 5 years
- 5 years – 8 years
- 8 years – 10 years
- 10 years – 15 years

- 15 years – 20 years
- Over 20 years

40. How many years of Internet experience do you have (your best estimate)?

    - Less than 1 year
    - 1 year – 2 years
    - 2 years – 3 years
    - 3 years – 4 years
    - 4 years – 5 years
    - 5 years – 8 years
    - 8 years – 10 years
    - Over 10 years

41. How many years of online shopping experience do you have (your best estimate)?

    - Less than 1 year
    - 1 year – 2 years
    - 2 years – 3 years
    - 3 years – 4 years
    - 4 years – 5 years
    - 5 years – 8 years
    - 8 years – 10 years
    - Over 10 years

# C List of Recommended Books in the Experiment

The format of book information: book title, book author(s), price, ISBN(10 digits)

1. The Science of Getting Rich, Wallace D Wattles, Ruth L Miller, 10.76, 1582701881
2. 1001 Things Every College Student Needs to Know, Harry H. Harrison Jr., 9.38, 1404104348
3. The 100 Best Business Books of All Time: What They Say, Why They Matter, and How They Can Help You, Jack Covert, 10.38, 1591842409
4. Judgment: How Winning Leaders Make Great Calls, Noel M. Tichy, 9.74, 1591841534
5. Thanks, Dad, Allen Appel, 5.24, 0312152213
6. Jan Karon Mitford Cookbook and Kitchen Reader: Recipes from Mitford Cooks, Favorite Tales from Mitford Books, Jan Karon, 11.69, 0670032395
7. The Dirty Dozen: How Twelve Supreme Court Cases Radically Expanded Government and Eroded Freedom, Robert A. Levy, 10.38, 1595230505
8. The Sistine Secrets: Michelangelo Forbidden Messages in the Heart of the Vatican, Benjamin Blech, 10.78, 0061469041
9. Wired for War: The Robotics Revolution and Conflict in the 21st Century, P. W. Singer, 11.98, 1594201986
10. I Am America And So Can You!, Stephen Colbert, 10.80, 0446580503
11. Irresistible Forces Thorndike Press Large Print African American Series, Brenda Jackson, 13.27, 1410417352
12. Lover Avenged Black Dagger Brotherhood, Book 7, J.R. Ward, 9.98, 0451225856

13. All Together Dead Southern Vampire Mysteries, Book 7, Charlaine Harris, 16.47, 0441014941

14. Open: An Autobiography [DECKLE EDGE], Andre Agassi, 11.75, 0307268195

15. Something Borrowed, Emily Giffin, 5.51, 0312321198

16. The Master Key System, Charles F. Haanel, 5.99, 1934451320

17. Glenn Beck Common Sense: The Case Against an Out-of-Control Government, Inspired by Thomas Paine, Glenn Beck, 7.19, 1439168571

18. Eat This Not That! Supermarket Survival Guide: The No-Diet Weight Loss Solution, David Zinczenko, 11.66, 1605298387

19. Pride and Prejudice and Zombies: The Classic Regency Romance - Now with Ultraviolent Zombie Mayhem!, Jane Austen, 8.47, 1594743347

20. The Last Song, Nicholas Sparks, 13.49, 0446547565

21. The Vortex: Where the Law of Attraction Assembles All Cooperative Relationships, Esther Hicks, 11.53, 1401918824

22. Three Cups of Tea: One Man Journey to Change the World... One Child at a Time, Sarah Thomson, 8.99, 0142414123

23. Tempted House of Night Novels, P. C. Cast, 10.47, 0312567480

24. Emergency: This Book Will Save Your Life, Neil Strauss, 11.55, 0060898771

25. Beautiful Creatures, Kami Garcia, 9.71, 0316042676

26. The Unit, Ninni Holmqvist, 10.17, 1590513134

27. How I Became a Famous Novelist, Steve Hely, 10.08, 0802170609

28. Tinkers, Paul Harding, 11.66, 193413712X

29. Genesis, Bernard Beckett, 13.60, 0547225490

30. Your Credit Score, Your Money & What Is at Stake (Updated Edition): How to Improve the 3-Digit Number that Shapes Your Financial Future, Liz Pulliam Weston, 12.91, 0137016611

31. How to Win Friends & Influence People, Dale Carnegie, 7.99, 0671723650

32. Deep Survival: Who Lives, Who Dies, and Why, Laurence Gonzales, 10.85, 0393326152

33. Heat Wave Heat Wave, Richard Castle, 11.69, 1401323820

34. The Girl with the Dragon Tattoo Vintage, Stieg Larsson, Reg Keeland, 8.97, 0307454541

35. The Shack, William P. Young, 8.99, 0964729237

36. Kindred in Death, J.D. Robb, 13.49, 0399155953

37. Push: A Novel, Sapphire, 7.15, 0679766758

38. Say You are One of Them Oprah Book Club, Uwem Akpan, 8.24, 0316086371

39. Clean Energy Common Sense: An American Call to Action on Global Climate Change, Frances Beinecke, Bob Deans, 9.95, 144220317X

40. The Elegance of the Hedgehog, Muriel Barbery, Alison Anderson, 9.00, 1933372605

41. Olive Kitteridge: Fiction, Elizabeth Strout, 8.40, 0812971833

42. Eclipse The Twilight Saga, Stephenie Meyer, 7.79, 0316027650

43. Twilight The Twilight Saga, Book 1, Stephenie Meyer, 6.59, 0316015849

44. New Moon The Twilight Saga, Book 2, Stephenie Meyer, 6.59, 0316024961

45. Three Cups of Tea: One Man Mission to Promote Peace . . . One School at a Time, Greg Mortenson, David Oliver Relin, 9.47, 0143038257

46. Freakonomics: A Rogue Economist Explores the Hidden Side of Everything P.S., Steven D. Levitt, Stephen J. Dubner, 9.35, 0060731338

47. The Time Traveler Wife, Audrey Niffenegger, 7.97, 015602943X

48. Ice: A Novel, Linda Howard, 11.00, 0345517199

49. The Tipping Point: How Little Things Can Make a Big Difference, Malcolm Gladwell, 7.97, 0316346624

50. Bed of Roses, Bride Quartet, Nora Roberts, 9.36, 0425230074