Automatic Privacy Policy Clustering

Mike Bergmann[1], Simone Fischer-Hübner[2], Andreas Pfitzmann[1], Marit Hansen[3], John Sören Pettersson[2]

[1] Technische Universität Dresden, Germany,
[2] Karlstad University, Sweden,
[3] Independent Centre for Privacy Protection (ICPP), Germany

Abstract. Every day users disclose various kinds of personal data using the Internet for daily activities. The disclosed data in summary may draw a perfect picture of its owner. Up to now it is difficult for end users to decide what to disclose and what to hide. We try to support the user in this task and propose a set of applicable privacy preferences settings to formalize the decision–procedure and to visualize the results.

Keywords: Privacy, Policy, eCommerce Application Scenario, Usability

1 Introduction

Nowadays electronic communication and electronic business gets more and more established. Every day the user discloses personal data using the Internet for daily activities like looking up timetable of public transport, shopping, translation services, etc. Even television gets a new face with the upcoming internet–based IPTV. Using TV by connecting to an Internet broadcast service may release the IP address and the user's watching behaviour. In summary, watching IPTV like any other internet application usage becomes possibly traceable and therefore a possible privacy issue.

To enforce the informational right of self–determination in the digital world, the importance of user–controlled identity management increases. However, privacy enhancing technologies, based on complex concepts (like credentials, certificates and pseudonyms) are often not easy understood by users. Additionally a user normally has to deal with various email accounts, public/private key pairs, (cell-) phones and credit card numbers, GPS positioning and movement track data and so on. This increases the complexity too.

In [BRP05] an approach is sketched to simplify privacy–enhancing identity management by using a town map–like approach to create a relationship between different kinds of privacy preferences and their representations in the topology of a artificial town map. There are various other methods, like role concept at [CK03], P3P policies [W3C02] etc. The PRIME project[4] is presenting a slightly

[4] The PRIME project receives research funding from the European Community's Sixth Framework Program and the Swiss Federal Office for Education and Science. For further information see https://www.prime-project.eu/

different approach, combining roles and policies[PFHD+05]. But anyhow the management of dedicated privacy preferences, assigned to dedicated activities remains difficult.

This paper proposes a set of predefined privacy preferences for online transactions based on existing application scenarios. We complement the analysis by interviewing users. Based on these results, we discuss the application of the proposed privacy preferences for typically application scenarios. Finally we decompose the complex ebusiness process into sub tasks to enforce privacy–friendly data handling according to a proposed "strict no transfer" policy. Finally a wizard approach will offer a decrease of complexity with further potential for simplification.

1.1 Application Scenarios

To investigate details of privacy preferences we performed an internal PRIME privacy preferences survey [Ber07]. 35 persons from various countries of the European Union took part in the survey[5]. 17 detailed questions related to private and business context were presented. The details are listed in [Ber07]. As a conclusion we summarize that about 68% of the participants mentioned ecommerce and ebanking applications as the most prominent privacy sensitive web applications. Obviously blogging and community scenarios are of less privacy relevant perception.[6].

Based on these results we focus our elaboration on web–based ecommerce scenarios as a starting point. We derived application scenario decomposition, explained in the next subsection. In detail, 46 dedicated online web services were considered with respect to their requested personal data, their promised privacy policy and the personal data really necessary for the service. We explored concrete use cases from the application areas *news service, email, Intranet, blogging, physical shopping, digital shopping, communication services, hosting, community applications and other services*. The result is that a limited set of data items is sufficient for most of the services. To cover 80% of the use cases, even in the new Web 2.0 landscape, we need only a few different data items. We extracted the following application scenarios[7]:

Business: Used in *professional* surroundings, as employee or administrative representative. The amount of *PII* to disclose depends on the services, usu-

[5] We are aware of the fact that this small number of participants in general and the PRIME project members with the different legal basics in particular are not fully representative for this kind of survey, but it holds as a starting point for further research.

[6] A dedicated survey about the community and per–to–per behaviour could help to figure out motivation and characteristic of these results

[7] This set of application scenarios could not be complete. We have to state that we do focus on todays web based application scenarios.

ally real resp. correct personal data[8] (and even sensitive data[9]) are required. The *real data* belonging to the user is used. The applied *privacy policy* limits PII usage to the stated *current purpose*[10] with *strict no transfer* to third parties outside the business. The business portals XING and LinkedIn are portal examples, ebanking and an ebay selling account ecommerce examples for this use case.

eShopping: Applicable in *semi–professional* surroundings, as buyer or seller. The amount of *PII* to disclose depends on the services, usually also authentic personal data are required. eShopping in general is used *pseudonymously*. The applied *privacy policy* limits PII usage to the stated *current purpose* with *transfer* to third parties to fulfil the business processes. Normal ebanking, amazon and ebay accounts are ecommerce examples for this use case.

SocialNetwork: Applicable for social networks in *non–professional* context, about music, photo, video sharing and so on. *PII* are not necessary, but expected to be released and authentic to increase the trustworthiness. In general a *pseudonymous* account is used. The applied *privacy policy* is less strict wrt. eShopping. Beside the stated *current purpose* the data are used for statistical, marketing and other purposes, too. The data could be *transferred* to business partners fulfilling the policy requirements of the service provider. Prominent examples are Skype, YouTube, MySpace, SecondLife etc.

Download: To download something in *non–professional* context when payment is not mandatory. *PII* are not required, but often requested. The data is not sensitive and possibly even not authentic (faked name, address data etc.) A *pseudonymous* account is used. The *privacy policy* may have some marketing aspect included, consequently the *purpose* is extended to marketing with possible *transfer* to third parties.

Blog: Read, edit, create new comments in a news forum or blog within *non–professional* context. Wa assume, that the released data is of less importance. It is personal data, but not identifying (*no PII*) or possibly not authentic. The usage is performed under *pseudonym or even anonymously*. The applied *privacy policy* often states *further purposes* beside the current purpose and allows *transfer* to third parties.

Email: To access a freemail/email account to write emails, to configure spam filter and to fill the address book in *non–professional* context. The release of *PII* is not required, but during usage the personal data could be accumulated to PII, especially address data, contact information, personal interests and

[8] We could require certified data, but this is less usual in the explored scenarios

[9] Sensitive in the meaning according to the definition of Art.8 [EC] and additionally according to the perception of the user. The preset should contain the corresponding sensitivity flag that a user could set.

[10] Current Purpose is defined [W3C02] as the usage for *Completion and Support of Activity For Which Data Was Provided: Information may be used by the service provider to complete the activity for which it was provided, whether a one-time activity such as returning the results from a Web search, forwarding an email message, or placing an order; or a recurring activity such as providing a subscription service, or allowing access to an online address book or electronic wallet.*

other[11]. The applied *privacy policy* allows *additional purposes and transfer* to third parties. The released personal data could be *sensitive*[12]. Prominent examples are Yahoo or Google mail for instance. Especially the Google mail privacy policy [Inc07] does not comply to our understanding what a user and privacy–friendly data handling policy is[13].

Membership: To get access to restricted resources like special web pages etc. in *semi–professional* surroundings. *PII* is required and may be *sensitive*. The access is provided *pseudonymously*. The *privacy policy* limits the PII usage to the *current purpose* with *no transfer* to third parties. Examples are automobile and sport club membership etc.

Further: All other application scenarios like *infrastructure, licensing, collaboration, news* are of less frequent usage and fit into one of the above mentioned scenarios.

In this section we defined and elaborated typical application scenarios wrt. identity management aspects. We discussed the necessary amount and quality of data and personal data for each scenario and the privacy policies a user usually expects.

Further on we will explore these application scenarios with respect to their required amount of personal data and the applied privacy settings.

2 Sets of Privacy Preferences

2.1 Managing Privacy Preferences

Table 1 structures the application scenarios mentioned in section 1.1. The aim is to predefine useful privacy preferences settings that could be offered to users to choose from (this helps to simplify as users could now choose from these predefined options instead of defining their preference settings by hand). We derive four elementary sets of privacy preferences:

No PII, transaction pseudonyms are used, i.e. user actions are not linkable, personal data are not released. The user may decide afterwards to connect transaction pseudonyms to become linkable. Reading a weblog or editing a Wikipedia entry anonymously [Wik07] (see the blog scenario description in section 1.1) are examples where such a preference setting could be applied[14].

No PII, but linkable personal data are not released. Transactions are linkable through the use of (role–) relationship pseudonyms as defined in [PK01] but communication is pseudonymous. Data about personal settings (and

[11] We assume that all content is encrypted so the content itself could not be read.

[12] could be for instance information about health status, about membership etc.

[13] Besides this, the Google privacy policy represents a completely different legal basis. But as the Google services are really widely distributed used and accepted we took such non–European application scenarios also into account.

[14] We assume the standard TCP/IP connection is anonymized (for instance using JAP [BFK00]) and no other identification (via cookies etc.) available

Scenario	*Personal Data*	*Purpose*	*Transfer*
Download	no data, not linkable		
Blog	no data, linkable		
Email	no data, linkable		
Membership	real data, linkable	current	not allowed
Business	real data, linkable	current	not allowed
eShopping	real data, linkable	current	not allowed
Social Network	(real) data, linkable	further	conditionally allowed

Table 1. The application scenarios data release policy items overview

pseudonyms) might be released (which are not directly identifying the user). Prominent examples are web mailers, various news panels (see the email application scenario in section 1.1) etc. The more data gets linkable, the more difficult it gets for the user to remain anonymous.

Disclose only necessary PII; Only a minimal amount of personal data are released for the purpose of the requested primary service. No "sensitive" data are released. Beyond that, strict no further transfer policy should be applied to other recipients to avoid data leaking, or at least data is only released to "trusted"[15] communication partners with the user's explicit consent only. The communication becomes linkable. A well known example is to buy a book online (see eshopping use case in section 1.1).

Disclose additional PII (related to above) personal data may be released also for additional services. Data are released only to "trusted" communication partners according to the user's trust policy. Transfer to other recipients should be controlled (e.g. only with the user's explicit consent) or only transfer to "trusted" recipients. "Sensitive" data are excluded. An example is a participation in a customer care program to get bonus points or other benefit. Since the additional service is a further purpose, it could be mapped into the setting above, but with dedicated additional purpose.

Table 2 gives an overview about the major features of the privacy preferences. The first column is numbering the sets. The second column contains a short description about the affected data, the third column states whether the user acts anonymous, pseudonymous or identifiable and the last column contains the data release policy applied to the mentioned data. The mentioned privacy policy sets do not cover all possible cases. There is no policy covering the "sensitive PII" or the "any transfer" case for instance. But as far as these cases are seldom and very critical we suggest to check the non–matching cases every time they occure very carefully.

The percentage column in Table 2 shows the mapping details about the application scenarios, elaborated in [Ber07] wrt. the concrete privacy preferences applied. It underlines that the privacy preferences set II with pseudonymous

[15] What trusted means a user defined in the privacy preferences. Valid measures could be personal reputation, privacy seals, trusted hardware etc.

	PII	Relationship	Purpose and Transfer	%
I	no PII	anonymous	none (not important)	2%
II	no PII, but user name, password, further additional non–identifying personal data	pseudonymous	only for current purpose and no transfer	65%
III	no sensitive PII	pseudonymously or real identity	only for current purpose and strict no further transfer	9%
IV	additional (but not sensitive) PII	pseudonymously or real identity	for additional purposes and strict no further transfer	24%

Table 2. Our four predefined privacy policy sets

relationship and without PII disclosure is currently the most often used configuration.

The privacy preferences could be applied for the standard web scenarios as shown in section 1.1. But what is about the applicable privacy preferences for each scenario? In set I and II personal data (PII) are not disclosed. However, a collection of search strings, a click stream, special personal preferences like favourite colour, interests etc. may lead to a significantly reduced anonymity set and therefore to an identification of the subject [BJ06]. It is necessary to pay attention to the privacy policy the communication partners have to agree on. In the next section, we will instantiate the discussed templates and privacy preferences to define concrete, applicable privacy preferences.

2.2 Clustering Privacy Preferences

A conventional web shopping scenario contains various sub tasks. Sub tasks for instance are *Order, Payment* and *Shipping*. If there are third parties involved, more sub tasks may be defined (address verification, certification etc.). To perform the shopping process quite a lot of data are necessary. But applying the discussed "no transfer" policy, as discussed in section **??**, setting III, we have to state that most of the current service scenarios are not feasible because of the need to transfer PII to third parties to fulfil the service.

However there is usually no need for the shop vendor to know the customers address or payment information. If we split the business process into the three mentioned separate parts *Order, Payment* and *Shipping*, every partner gets the data needed and is interconnected with the others by a so called transaction pseudonym. The benefit for the user is quite clear. Data disclosure is really limited to the party related to the business, the purpose, the data are released for, could be tailored for the very special business and there may be no need to agree on a third party transfer policy. Also from the service provider's point of view it becomes much easier to be legally compliant – if less data are seized less data have to be protected etc. In our example very few personal details at all

should be disclosed to the shop so the privacy policy may even be less strict. It could look like the following:

Ordering a book at www.bookstore.net:

Data	Communication Partner	Policy
Pseudonym (e.g. customer number), item of interest e.g. ISBN	Merchant/Service provider, in our example www.bookstore.ne	Only for "current" purpose, in our case selling a book; no further transfer, secure data storage as long as legally required

Payment with credit card

Data	Communication Partner	Policy
Credit card number, expiry date, real name, amount of money	Payment service provider, for instance www.cash.eu	Only for "current" purpose, in this case payment; no further transfer, secure data storage as long as legally required

Delivery of the good to a specified address

Data	Communication Partner	Policy
Address of the client and the (packed/hidden) good, i.e. the item of interest is not known	Shipping service provider, for instance www.ups.com	Only for "current" purpose, in our example to deliver the good; no further transfer, secure data storage as long as legally required

Table 3. Business process sub tasks

The Figure 1 shows the communication flow in case we split the data disclosure process as described. It shows that the user request is answered by the service provider, which may state a transaction id to link the processes and the accepted third parties to handle payment and shipping. The user now contacts the third parties directly, which use the transaction id and dedicated credentials to signal the status of the transaction to the service provider. In case of the shipping we sketched an additional communication flow to take into account variable shipping cost. The sipping provider asks for additional parameters, like size or weight for instance to calculate the shipping costs and sends it back to the initial service provider together with transactionId and a credential. The service provider now is able to update the payment information to request appropriate payment. If the delivered credentials satisfy the service provider the process is finalized by signalizing the final delivery of the good via the third party using the transaction id again.

In summary, splitting the business process into sub tasks offers the benefit to bind personal data, privacy policy and service providers to a dedicated transaction and purpose. It helps to fulfil the privacy principles in general and data minimization in particular and it lowers the complexity of data handling and release policy per transaction significantly.

2.3 Implementation and the "Wizard–like Approach"

In the introduction we promised to lower the complexity of identity management processes and to simplify the user interactions regarding privacy and identity management. However in the previous section, we have split the action buying a book into tree different actions with different service providers (see Figure 1). At first glance it seems to contradict to the idea of simplification.

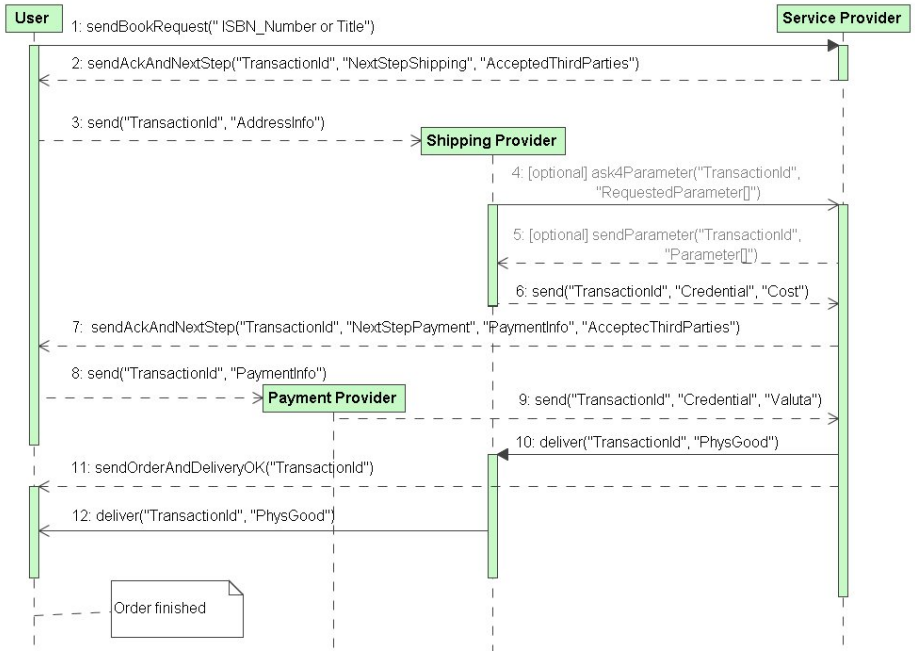

Fig. 1. A possible sequence diagram for our split scenario

We solve this issue by introducing a wizard–like approach assisting and guiding the user through the decision making process. The assistant presents a sequence of decision requests according to the privacy preferences to the end user to compile the final negotiated policy. As shown in Figure 2 the assistant should inform the user about the overall procedure, it allows to jump back to check previously stated preferences, to collect the required data items and to communicate to the service provider with respect to the requested certificates, seals, reputation, data handling policy and obligations.

In our example the sample dialogue contains three sections with the statements about data recipient, stated purposes and required personal data. The wizard in Figure 2 receives the template provided by the service provider. Using such presets the user could automatically handle the PII request if the corresponding privacy preferences are set to fulfil the service requirements and to

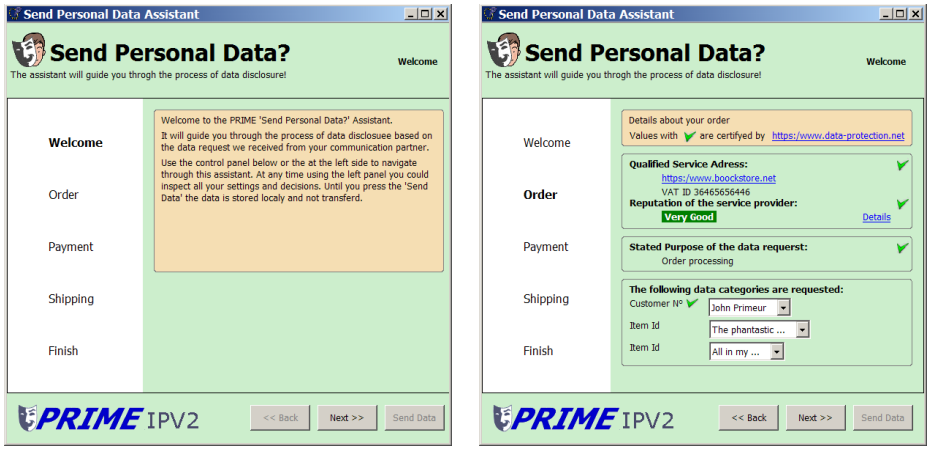

Fig. 2. An example "Send Personal Data" Assistant

maintain the privacy requirements. Thanks to the predefined template it becomes easy to gather and maintain the user's privacy requirements even for users not primarily interested in privacy.

3 Conclusion

In this paper we provide a set of predefined privacy preference settings, which cover most practical use cases. The user could simply choose from this set instead of defining preference settings by hand.

Beside this we discuss a new approach structuring privacy–related data disclosure in a privacy–friendly way. The presented approach to structure the privacy preferences using predefined templates allows to cover the ever increasing application fields, especially with respect to the upcoming Web 2.0 capabilities *without* increasing the numbers of roles and configurations a user has to deal with. The special focus to the strict "no transfer, only for the stated purpose" policy enhances the privacy of the user.

Further test of the user acceptance in the scope of PRIME and behind will help to improve and fine-tune the approach.

4 Acknowledgment

Thanks to our colleagues for the helpful comments. Special thanks to the members of the TU Dresden, namely Sebastian Clauß, Thomas Kriegelstein, to Lothar Fritsch from Johann Wolfgang Goethe–University Frankfurt, to Jan Schallaböck from the Independent Centre for Privacy Protection, Kiel, the PRIME internal survey participants and all others contributors for the very helpful discussion.

References

[Ber07] Mike Bergmann. PRIME internal privacy preferences survey about privacy concerns and conditions. In *Technical Report TUD-FI07-04-Mai 2007*, Technische Universität Dresden, Saxony, Germany, May 2007. Technische Universität Dresden. http://dud.inf.tu-dresden.de/~mb41/publications/TUD-FI07-04-Mai2007.pdf.

[BFK00] O. Berthold, H. Federrath, and M. Köhntopp. Project "Anonymity and Unobservability in the Internet". In *Proc. Workshop on Freedom and Privacy by Design / Conference on Freedom and Privacy 2000*, pages 57–65, Toronto/Canada, April 4-7 2000. ACM.

[BJ06] Michael Barbaro and Tom Zeller Jr. A face is exposed for AOL searcher No. 4417749. *New York times Online*, August 2006. http://www.nytimes.com/2006/08/09/technology/09aol.html?ex=1175400000&en=fd9b0c3b15c36970&ei=5070.

[BRP05] Mike Bergmann, Martin Rost, and John Sören Pettersson. Exploring the feasibility of a spatial user interface paradigm for privacy-enhancing technology. In *Proceedings of the Fourteenth International Conference on Information Systems Development*, Karlstad, August 2005. Springer-Verlag.

[CK03] Sebastian Claußand Thomas Kriegelstein. Datenschutzfreunliches Identitätsmanagement. *DuD Datenschutz und Datensicherheit*, 27:297, 2003.

[EC] Council of Europe EC. Convention for the protection of human rights and fundamental freedoms. Rome 4.XI.1950, online available at http://conventions.coe.int/treaty/en/Treaties/Html/005.htm (last visited 27th November, 2006).

[Inc07] Google Inc. Google Privacy Policy, Google Privacy Center, 2007. Online; accessed 20-May-2007; http://www.google.com/intl/en/privacypolicy.html.

[PFHD+05] John Sören Pettersson, Simone Fischer-Hübner, Ninni Danielsson, Jenny Nilson, Mike Bergmann, Sebastian Clauß, Thomas Kriegestein, and Henry Krasemann. Making PRIME usable. In *Symposium on Usable Privacy and Security*, Carnegie Mellon University, Pittsburgh, PA, USA, July 2005. Carnegie Mellon University.

[PK01] A. Pfitzmann and M. Köhntopp. Anonymity, unobservability, and pseudonymity - a proposal for terminology. In *Proceedings of WS on Design Issues in Anonymity and Unobservability*, Designing Privacy Enhancing Technologies, LNCS2009, Proceedings of the Fourteenth International Conference on Information Systems Development, Heidelberg, August 2001. LNCS. revised version http://www.koehntopp.de/marit/pub/anon/Anon_Terminology.pdf.

[PRI07] PRIME Project. PRIME PIIBase.owl, version v1.11, March 2007. online available at https://www.prime-project.eu/ont/PIIBase.

[W3C02] W3C. Platform for Privacy Preferences, version REC-P3P-20020416, April 2002. online available at http://www.w3.org/TR/P3P/.

[Wik07] Wikipedia. Help:contents/getting started — wikipedia, the free encyclopedia, 2007. [Online; accessed 19-May-2007] http://en.wikipedia.org/wiki/Help:Contents/Getting_started.