Comparing and Matching Privacy Policies Using Community Consensus

Yee, George; Korba, Larry

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

Comparing and Matching Privacy Policies Using Community Consensus *

Yee, G., and Korba, L.
May 2005

* published in Proceedings of the 2005 Information Resources Management Association International Conference (IRMA 2005). May 15-18, 2005. San Diego, California, USA. NRC 47430.

Canada

Comparing and Matching Privacy Policies Using Community Consensus[1]

George Yee and Larry Korba

Institute for Information Technology
National Research Council Canada
1200 Montreal Road, Building M-50
Ottawa, Ontario, Canada K1A 0R6
{George.Yee, Larry.Korba}@nrc-cnrc.gc.ca

ABSTRACT

The growth of the Internet is increasing the deployment of e-services in such areas as e-commerce, e-learning, and e-health. In parallel, service providers and consumers are realizing the need for privacy. Managing privacy using privacy policies is a promising approach. In this approach, an e-service consumer and an e-service provider each have separate privacy policies. Before an e-service is engaged, the consumer's policy must "match" the provider's policy. However, how is this matching defined? We propose a method for comparing consumer and provider privacy policies by comparing the privacy levels of privacy preferences in the policies. A "match" between consumer and provider privacy policies is then defined using this method. Since the notion of privacy is subjective and can vary from individual to individual, the privacy levels of individual preferences are obtained through community consensus.

1 INTRODUCTION

The rapid development of the Internet has been accompanied by a growth in the number of e-services available to consumers. E-services are available for banking, shopping, learning, healthcare, Government Online, and so on. However, each of these services requires a consumer's personal information which leads to concerns over privacy. In order for e-services to be successful, privacy must be protected. An effective and flexible way of protecting privacy is to manage it using privacy policies. In this approach, a provider of an e-service and a consumer of that e-service each have a privacy policy containing individual privacy preferences. The provider's policy expresses what private information it requires from the consumer. The consumer's policy expresses what private information she is willing to give the provider. Before the service can begin, the policies have to "match". The objectives of this paper are to a) propose a method for comparing privacy policies that can be applied to define this "match", and b) ensure that this comparison reflects community values, since the notion of privacy is subjective and can vary between individuals.

In the literature, one work related to the comparison of privacy policies is the matching of consumer privacy preferences to web site P3P privacy policies [1] using the AT&T Privacy Bird [11]. However, the Privacy Bird a) appears to match on only a predefined set of privacy preferences useful for web sites, and b) does only identical matching – it does not employ a continuous numerical comparison model that can allow a positive comparison even in the absence of an identical match up. In this work, the items of private information can be anything and privacy policies are compared based on the numerical privacy levels of the policy contents to give greater flexibility (e.g. a positive outcome can result even if the corresponding privacy preferences are not identical). Another work, by Backes et al [12], examines the comparison of enterprise privacy policies using a formal abstract syntax and semantics to express the policy contents. In this approach, one policy "matches" another if using the first policy automatically fulfills the second policy (the first policy is said to "refine" the second policy). These authors provide

formal definitions and rules under which refinement can occur, and incorporate them in an algorithm for checking refinement for privacy policies expressed in EPAL [13]. In this work, we take a simpler more pragmatic approach, preferring to compare 2 polices directly using levels of privacy as measures. Our approach is more likely to be understood by consumers thereby gaining their trust. Work related to privacy policies include the semi-automated derivation of personal privacy policies [6], privacy policy negotiation [7, 8, 9], and privacy policy compliance [10].

Section 2 describes the content of privacy policies. Section 3 presents our method for comparing policies. Section 4 proposes a community consensus approach for obtaining privacy levels required by our method. Section 5 gives conclusions and future research.

2 THE CONTENT OF PRIVACY POLICIES

To compare privacy policies we need to first define their content. We consider content requirements imposed by privacy legislation. This approach is realistic since e-service providers must comply with privacy legislation and hence such content must be implemented. Of course, other types of content may be expressed but they may not be part of a provider's privacy policy since they are not required by law.

Privacy Legislation and Directives

In Canada, privacy legislation is enacted in *PIPEDA (Personal Information Protection and Electronic Documents Act)* [2] and is based on the Canadian Standards Association's *Model Code for the Protection of Personal Information* [3], recognized as a national standard in 1996. This Code consists of ten Privacy Principles (Table 1) that we call CSAPP. CSAPP is a good choice to use for privacy policy content requirements because it is representative of principles behind privacy legislation in many countries, including the European Union's Data Protection Directive [4].

Table 1. CSAPP - The Ten Privacy Principles from the Canadian Standards Association

Principle	*Description*
1. Accountability	An organization is responsible for personal information under its control and shall designate an individual or individuals accountable for the organization's compliance with the privacy principles.
2. Identifying Purposes	The purposes for which personal information is collected shall be identified by the organization at or before the time the information is collected.
3. Consent	The knowledge and consent of the individual are required for the collection, use or disclosure of personal information, except when inappropriate.
4. Limiting Collection	The collection of personal information shall be limited to that which is necessary for the purposes identified by the organization. Information shall be collected by fair and lawful means.
5. Limiting Use, Disclosure, and Retention	Personal information shall not be used or disclosed for purposes other than those for which it was collected, except with the consent of the individual or as required by the law. In addition, personal information shall be retained only as long as necessary for fulfillment of those purposes.
6. Accuracy	Personal information shall be as accurate, complete, and up-to-date as is necessary for the purposes for which it is to be used.
7. Safeguards	Security safeguards appropriate to the sensitivity of the information shall be used to protect personal information.
8. Openness	An organization shall make readily available to individuals specific information about its policies and practices relating to the management of personal information.
9. Individual Access	Upon request, an individual shall be informed of the existence, use and disclosure of his or her personal information and shall be given access to that information. An individual shall be able to challenge the accuracy and completeness of the information and have it amended as appropriate.
10. Challenging Compliance	An individual shall be able to address a challenge concerning compliance with the above principles to the designated individual or individuals accountable for the organization's compliance.

Privacy Policy Content Requirements from CSAPP

We interpret "organization" as "provider" and "individual" as "consumer". We use CSAPP.n to denote Principle n of the CSAPP. Principle CSAPP.2 implies that there could be different providers requesting the information, thus implying a *collector* attribute. Principle CSAPP.4 implies that there is a *what* attribute, i.e. what private information is being collected. Principles CSAPP.2, CSAPP.4, and CSAPP.5 state that there are *purposes* for which the private information is being collected. Principles CSAPP.3, CSAPP.5 and CSAPP.9 imply that the private information can be disclosed to other parties, giving a *disclose-to* attribute. Principle CSAPP.5 implies a *retention time* attribute for the retention of private information. Thus, from the CSAPP we derive 5 attributes of private information collection, namely *collector*, *what*, *purposes*, *disclose-to*, and *retention time*. The remaining principles prescribe certain operational requirements that must be satisfied between provider and consumer, such as consent, and are discussed in [6].

We call the attribute grouping <collector, what, purposes, retention time, disclose-to> a *privacy rule*. A privacy policy then consists of a header section followed by one or more privacy rules. This header consists of the fields: *Policy Use* (for what e-service?), *Owner* (name of the provider or consumer who owns the policy), *Proxy* (Yes or no – yes if a proxy will act for the consumer to give the information), and *Valid* (period of time during which the policy is valid). Figure 1 shows example consumer and provider privacy policies for various e-services.

Privacy policies need to be expressed in a machine-readable policy language such as APPEL [5] (XML-based). The investigation of suitable policy languages for privacy policies is ongoing research and outside the scope of this paper.

3 METHOD FOR COMPARING PRIVACY POLICIES

We compare privacy policies by comparing the levels of privacy in individual privacy rules. Each rule has the attributes *collector*, *what*, *purposes*, *retention time*, and *disclose-to*. However, *collector* does not contribute to the level of privacy – it merely identifies who in the provider's organization will receive the consumer's information. Either the consumer agrees with who the *collector* is or she does not. Hence, we will let *collector* be compared between the two policies – either the *collector* is the same in both policies or the provider's policy has a specific name and the consumer's policy has "any" (necessary for consumer agreement); otherwise there is no agreement (need for negotiation). The attribute *disclose-to* behaves the same way as *collector*, and we treat it the same way as *collector*. Let the remaining attributes *what*, *purposes*, and *retention time* be represented by $w_{i,c}$, $p_{i,c}$, and $r_{i,c}$ respectively, for consumer privacy rule *i*. Similarly, we have $w_{i,p}$, $p_{i,p}$, and $r_{i,p}$ for provider privacy rule *i*. We wish to ascribe a function *v* that returns a numerical level of privacy from the consumer's point of view (since it is the consumer's private information that is requested) when applied to the attributes *what*, *purposes*, and *retention time*. A high *v* means a high level of privacy; a low *v* means a low level of privacy.

Definition 1
Let $V_{i,c}$, $V_{i,p}$ represent the privacy levels of consumer privacy rule *i* and provider privacy rule *i*, respectively. Then

$$V_{i,c} = v(w_{i,c}, p_{i,c}, r_{i,c}),$$
$$V_{i,p} = v(w_{i,p}, p_{i,p}, r_{i,p}).$$

Definition 2
There is a *match* between consumer and provider privacy policies if:

a) For corresponding privacy rules (same *what* in both policies) $V_{i,p} \geq V_{i,c} \quad for\ all\ i.$

Where there is no corresponding rule in one policy, we still consider the rule present and assign it a privacy level of ∞, since no rule means no information required, which is the highest level of privacy.

b) The values for *collector* and *disclose-to* attributes are the same for corresponding privacy rules in both provider and consumer policies, or the provider's policy has a specific name and the consumer's policy has "any".

Otherwise, there is a *mismatch.*

In Definition 2, part a), the level of privacy in the provider's rule is greater or equal to the level of privacy in the corresponding consumer rule (the provider is demanding less information (greater privacy) than the consumer is willing to offer).

Definition 3

The level of privacy P_p in a provider's privacy policy is $P_p = \sum_i V_{i,p}$. Similarly, the level of privacy P_c in a consumer's privacy policy is $P_c = \sum_i V_{i,c}$.

It is difficult to assign values to *v* (Definition 1) universally and consistently because privacy is a subjective notion, and one consumer's view of privacy may be different from another consumer's view. Section 4 will provide an approach for obtaining universal and consistent values for *v*. Then Definition 2 can be used to determine if provider and consumer policies match. We do not make use of definition 3 in this paper. We include it only for interest.

3.1 Policy Matching Shortcuts

In comparing policies, it is not always necessary to carry out the above evaluation. We mention two shortcuts.

Shortcut 1

Both policies are the same except one policy has fewer privacy rules than the other policy. Then according to Definition 2, there is a match if the policy with fewer privacy rules belongs to the provider. Otherwise, there is a mismatch.

Shortcut 2

Both policies are the same except one policy has one or more attributes that clearly lead to higher levels of privacy for the associated privacy rules. According to Definition 2, there is a match if the rules with resulting higher levels of privacy belong to the provider's policy. Otherwise, there is a mismatch. An example of such an attribute is retention time, where there is a match if the provider's policy has less retention time (higher privacy level) for corresponding items.

Consider the example policies in Figure 1. In these policies, there is a match for e-learning according to Shortcut 2, since the rule with lower retention time belongs to the provider. There is a mismatch for book seller according to Shortcut 1 since the policy with fewer privacy rules belongs to the consumer. There is a mismatch for medical help according to Shortcut 2, since the rule with the higher level of privacy is the one specifying a particular collector (Dr. Smith), and this rule belongs to the consumer's policy.

4 DETERMINING PRIVACY LEVELS USING COMMUNITY CONSENSUS

This section investigates how privacy levels can be assigned to privacy rules using community consensus for use in Definition 2 to determine a match. Community consensus overcomes the problem that what's private for one person may not be private for another person. Community consensus is obtained through surveys.

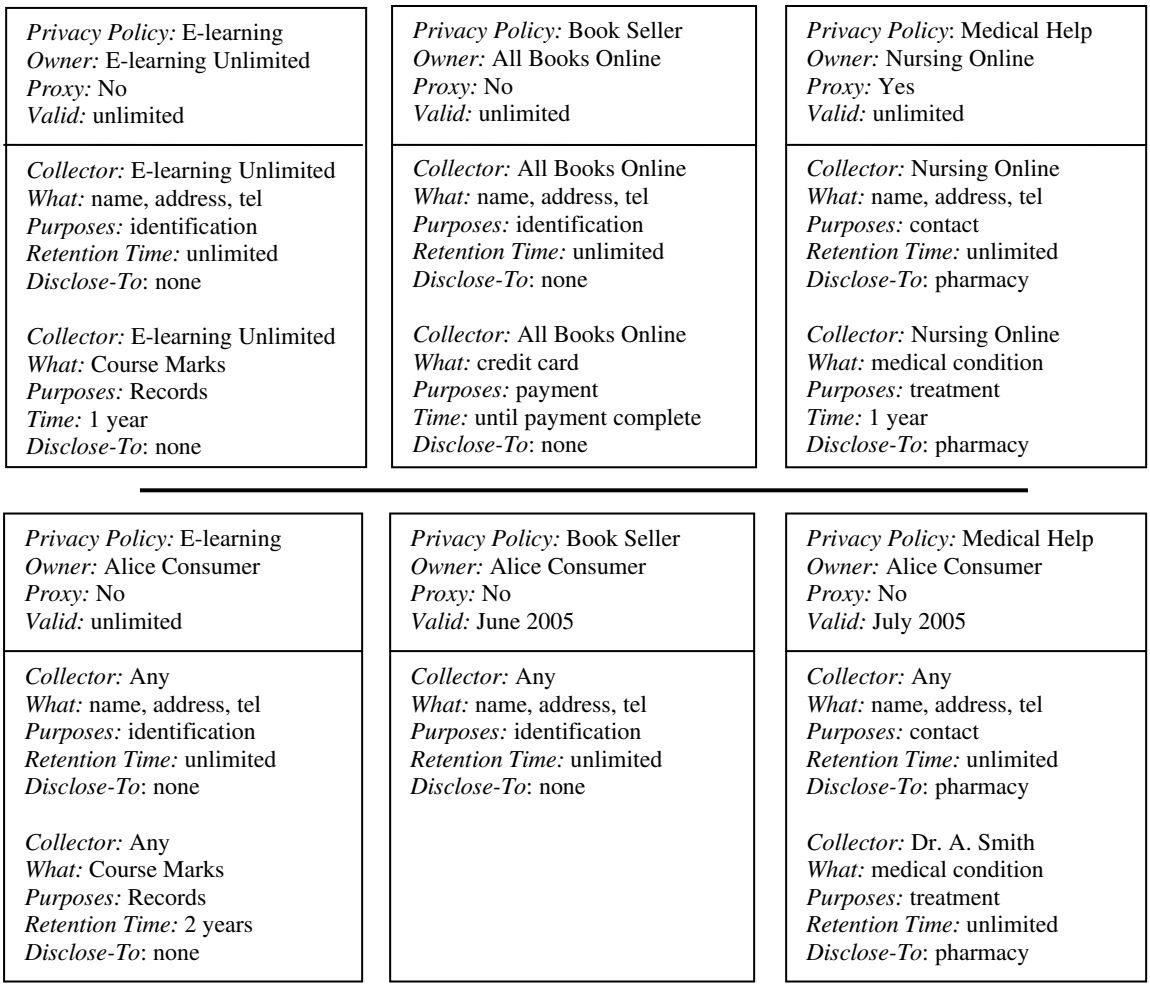

Figure 1. Example Provider Privacy Policies (top) and Corresponding Consumer Privacy Policies (bottom)

Privacy Levels Through Surveys

(a) A privacy rules evaluation provider makes use of third party surveys performed on a regular basis as well as those published in research literature to obtain user perceptions of the level of privacy for various items of private information (PI) separated according to their uses. This gives a sensitivity or range of privacy levels for different PI in different situations.

(b) Corresponding to a provider's privacy policy (which specifies what PI are required), the rules evaluation provider or a software application constructs and ranks partial privacy rules <*what, purposes, retention time*> for each use using the PI in (a), according to their sensitivity levels. The outcome of this process is a set of partial privacy rules, ranked by PI sensitivity or privacy level for different providers. There are different ways to do this ranking. One way is to assign a partial privacy rule the median of its sensitivity range as its privacy level (illustrated below).

(c) Providers and consumers obtain online from the rules evaluation provider the privacy levels for the rules and use in their policies. They do this by specifying the partial rules, the use for the rules, and the provider. In this way, large populations of providers and consumers may quickly obtain privacy levels for their rules to use in determining if their policies match.

This approach requires trust in the rules evaluation provider. Effectively the rules evaluation provider becomes a trusted third party. A certification process for this provider is probably required. For instance, in Canada, the offices for the provincial and federal privacy commissioners could be this certification body. They could also provide this privacy level determination service. Figure 2 illustrates this approach.

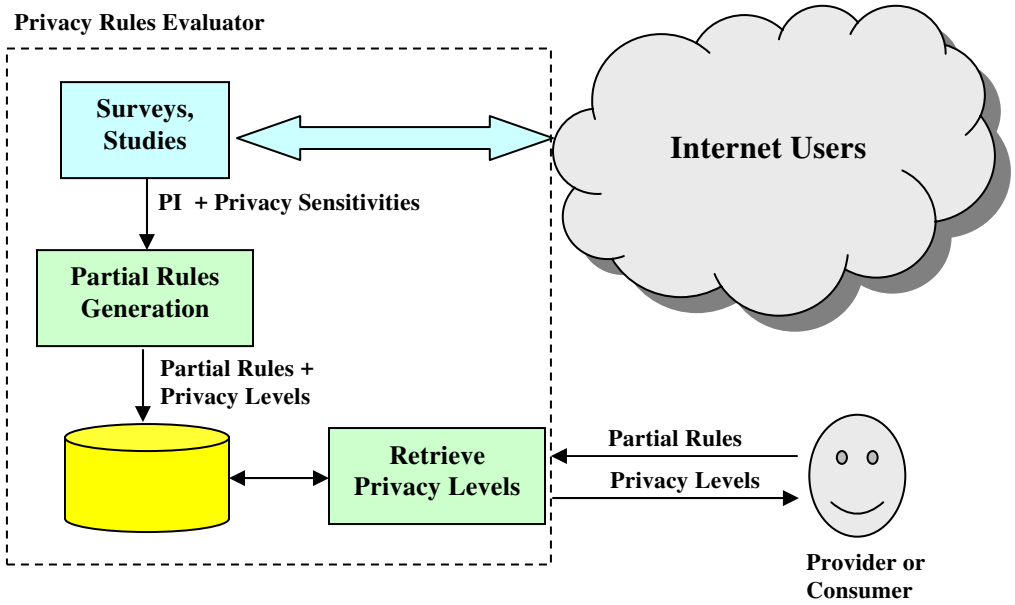

Figure 2. Determination of privacy levels from surveys

Example:

Suppose the item of PI for which we wish to derive a privacy level is "course marks retention time" from the e-learning privacy policy in Figure 1.

Then the above steps are implemented as follows:

a) The third party survey generates the following results for course marks retention time (the higher the privacy level, the higher the privacy; the highest level is 5, the lowest level is 1) for a use (or context) of e-learning.

PI	Privacy Level
course marks retention time 6 months	3
course marks retention time 6 months	4
course marks retention time 6 months	4
course marks retention time 6 months	5
course marks retention time 12 months	1
course marks retention time 12 months	1
course marks retention time 12 months	2
course marks retention time 12 months	3

Note that the other parameters in a partial privacy rule (i.e. *what, purposes*) may change too, not just retention time. We change retention time only to keep the example simple. Actually, each different combination of parameters represents a different privacy level. Also, the longer the marks are retained, the lower the privacy level.

(b) In this step, the rules evaluation provider constructs partial privacy rules from the PI in (a) and ranks them using the median value from the corresponding sensitivity range. Thus for the 4 course mark retention times of 6 months, the lowest value is 3, the highest value is 5, and the median is 4. Therefore the partial rule < course marks, records, 6 months > is ranked with privacy level 4. Similarly, the partial rule < course marks, records, 12 months > is ranked with privacy level 2.

(c) To obtain their privacy levels, providers and consumers specify the use (or context) as e-learning and their partial rules. Suppose one of the rules is < course marks, records, 6 months >. The provider or consumer then obtains a privacy level of 4 for this partial rule.

5 CONCLUSIONS AND FUTURE RESEARCH

We began by defining the content of a privacy policy using representative Canadian privacy legislation. The use of privacy legislation to derive requirements for privacy policy content is practical since e-service providers must comply with such content by law. We proposed a method for comparing privacy policies using the privacy levels of individual privacy preferences to determine a match. We then gave an approach for obtaining these privacy levels through community consensus using third party surveys. Our approach reflects the privacy sensitivities of the online community, accounting for the fact that what is private to one person may not be private to another. By specifying "use" for PI, we also account for the fact that the privacy level of PI depends on how the PI is used or its context. Our approach may be automated for fast policy matching in an e-commerce environment.

The methods provided in this paper also apply to privacy policy negotiation [7, 8, 9], since policies are compared and matched there as well. Further, our approach can be applied to other types of privacy policies that may differ from ours in terms of privacy provisions, so long as the corresponding privacy levels can be ascertained. As well, our approach is applicable to privacy policies that are expressed in particular implementation languages, since we have not specified any particular implementation language in the above development.

For future research, we plan to investigate other ways of matching privacy policies easily. For instance, since e-business is a global affair, there may well be differing ways in which the "what" and "purposes" attributes may be expressed. These may vary on a jurisdiction basis, or depend on different policy variants. Approaches such as fuzzy comparisons may be possible in some instances. In others, dictionaries designed to translate "what" and "purposes" amongst different jurisdictions or services may be helpful. Helping the user understand who the collector or disclose-to organizations are, how they use the information gathered, and their reputation with other users would promote user trust. Exploring ways to gather this information in a world-wide context, and portray it effectively to a user is another research topic. We plan to construct simulations of the methods presented in this paper to assess effectiveness, especially regarding scalability and performance issues.

REFERENCES

[1] W3C, "The Platform for Privacy Preferences", http://www.w3.org/P3P/

[2] Department of Justice, Privacy Provisions Highlights, http://canada.justice.gc.ca/en/news/nr/1998/attback2.html

[3] Canadian Standards Association, "Model Code for the Protection of Personal Information", retrieved Sept. 5, 2003 from: http://www.csa.ca/standards/privacy/code/Default.asp?articleID=5286&language=English

[4] European Union, "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", unofficial text retrieved Sept. 5, 2003 from: http://aspe.hhs.gov/datacncl/eudirect.htm

[5] W3C, "A P3P Preference Exchange Language 1.0 (APPEL1.0)", W3C Working Draft 15 April 2002, http://www.w3.org/TR/P3P-preferences/

[6] G. Yee and L. Korba, "Semi-Automated Derivation of Personal Privacy Policies", Proceedings, The IRMA International Conference 2004 (IRMA 2004), New Orleans, May 23-26, 2004.

[7] G. Yee and L. Korba, "Bilateral E-services Negotiation Under Uncertainty", Proceedings, The 2003 International Symposium on Applications and the Internet (SAINT2003), Orlando, Florida, Jan. 27-31, 2003.

[8] G. Yee and L. Korba, "The Negotiation of Privacy Policies in Distance Education", Proceedings, 14th IRMA International Conference, Philadelphia, Pennsylvania, May 18-21, 2003.

[9] L. Korba, "Privacy in Distributed Electronic Commerce", Proc. of the 35th Hawaii International Conference on System Science (HICSS), Hawaii, January 7-11, 2002.

[10] G. Yee and L. Korba, "Privacy Policy Compliance for Web Services", Proceedings, 2004 IEEE International Conference on Web Services (ICWS 2004), San Diego, California, July 5-9, 2004.

[11] L. F. Cranor, M. Arjula, and P. Guduru, "Use of a P3P User Agent by Early Adopters", Proceedings of WPES'02, November 2002, Washington, DC, USA.

[12] M. Backes, G. Karjoth, W. Bagga, and M. Schunter, "Efficient Comparison of Enterprise Privacy Policies", Proceedings, 2004 ACM Symposium on Applied Computing, Nicosia, Cyprus, March 2004.

[13] P. Ashley, S. Hada, G. Karjoth, C. Powers, and M. Schunter, "Enterprise Privacy Authorization Language (EPAL), Research Report RZ 3485, IBM Research, March 2003.

---

[1] NRC Paper Number: NRC 47430