Selective Privacy in a Web-Based World: Challenges of Representing and Inferring Context

K. Krasnow Waterman[1], Deborah L. McGuinness[2], Li Ding[2]

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge MA 02139, USA
[2]Tetherless World Constellation, Rensselaer Polytechnic Institute, 110 8th St, Troy, NY12180, USA
kkw@mit.edu, dlm@cs.rpi.edu, dingl@cs.rpi.edu

Abstract

There is a growing awareness and interest in the issues of accountability and transparency in the pursuit of digital privacy. In previous work, we asserted that systems needed to be "policy aware" and able to compute the likely compliance of any digital transaction with the associated privacy policies (law, rule, or contract). This paper focuses on one critical step in respecting privacy in a digital environment, that of understanding the context associated with each digital transaction. For any individual transaction, the pivotal fact may be context information about the data, the party seeking to use it, the specific action to be taken, or the associated rules. We believe that the granularity of semantic web representation is well suited to this challenge and we support this position in the paper.

From "Privacy" to "Selective Privacy"

When coined, apparently by scrivener error nearly five hundred years ago [1], "privacy" meant to seclude or keep out. Over the many years since, its definition has expanded to include the concepts of being free from public attention, being free from intrusion or interference, and having personal volition about that freedom. Today, with voluminous information about each individual already having crossed the digital privet hedge, it is too late to consider digital privacy as the ability to hold back one's personal information. Now, it must mean the ability to selectively control the use of that data based upon the context of the transaction.

This is not a wholly new concept. In the physical world, we provide far less detail about our social, family, and medical lives with our business associates than we do with family and friends. Conversely, we generally provide less financial details of our life to our social network than our commercial one. And, we sub-divide these decisions much more granularly – providing more detailed financial information when seeking a mortgage than when making a deposit or providing medical information to a supervisor in order to obtain a work variance – according to context. Our personal rules for privacy are kaleidoscopic, changing as the situation changes.

So, too, in the commercial and governmental environments, the written rules for privacy are heavily laced with contextual terms and conditions. Medical professionals may look at patient physical health records if they are treating the individual or providing institutional oversight (actor context); they are permitted more limited review if they are addressing insurance or other financial issues (event context) or if the records contain mental health information (data context). Financial professionals are permitted to seek and review more personal information when opening an account than when engaging in day-to-day transactions (event context). Government agencies must apply the Privacy Act [2] to the use of any data; this requires the application of some rules universally, but also the application of different rules per data repository (rule context). As implied by these examples, we take the broad view of privacy rules, meaning any rule which seeks to limit access to information about an individual or an identifiable group for the purpose of protecting them from harm.

We previously focused on the importance of accountability and transparency in the arsenal of privacy protection. [3] This paper focuses on the importance of context in successfully modeling the decision process for privacy protection. It also describes how decision systems that rely solely on actor's role, data content, and action, while incrementally useful – and possible to make transparent – fail to meet the nuanced requirements and expectations of privacy policy – and, therefore, can't be fully accountable. We believe Web-based systems can meet the need for obtaining decision-relevant data from beyond the locus of the transaction and further posit that semantic web based systems can provide the level of

granularity and inference that context-laden, selective privacy rules require.

Context Types and Their Challenges

The laws and policies for handling information have the base (most elementary) form

A(n) [actor] is [P/P/R□] to [action] [data]

"P/P/R" indicates the rule's operative instruction – that an action is "Permitted," "Prohibited," or "Required." And, an "action" is any possible use of data, including collect, retain, access, merge, copy, or delete.

For example,

A [police officer] is [permitted] to [access] [criminal history information]

In the base form, the rule is relatively straightforward to code and the necessary data to fire the rule is relatively easy to locate and retrieve. It requires only the highest level summary information – a job role and the data category of the target data. This is the incremental rule form that many access control systems have adopted. However, this form is fundamentally flawed, because, for example, it would allow the officer to look up his daughter's new boyfriend without any just cause, which is exactly the sort of abuse of authority that our privacy rules seek to curtail.

Real data handling rules add context requirements to one or more of the variables, significantly increasing the system design and semantic challenge. For example, a rule of the form

A(n) [actor] [actor context] is [p/r/r] [p/r/r context] to [action] [action context] [data] [data context]

might look like this:

A [police officer] [who is conducting a criminal case interview] is [permitted] [before or during the interview] to [access] [directly or by request to an authorized person] [criminal history information] [about the person being interviewed].

This rule requires the ability to access information about what the actor is doing or why; to calculate time, to determine relationships, and to associate the target data with information outside that dataset. This richer form of rule permits the sort of selective privacy we expect by taking context into account. Properly implemented in a system, this rule would have denied the officer's attempted boyfriend look-up.

Privacy rules, as they exist today, raise many more context-related challenges. This paper describes a number of them we have encountered in our research, and discusses our solutions and some remaining challenges. While the particular examples modeled in our work were largely governmental, it is important to remember that the same issues apply to commerce and social environments. We believe that expressing, finding, and reasoning over the nuances of context are critical to the satisfactory digital implementation of privacy rules and the ability to be sufficiently transparent and accountable to engender trust in such a system.

Data Context

Chain of custody is a familiar example of an institutional legal method for establishing trust about data. We trust it, not because it has a particular name or contains particular information – its content - but because through careful reconstruction of its provenance – which is a form of context - we can confirm that it has remained intact. So, too, many privacy rules require knowledge of the contextual information about data.

There are many data context challenges that need to be resolved to produce a fully accountable system. The first challenge we address relates to capturing the data path information – location of data, the path it has traveled since inception, transfer dates, etc. This is typically refered to as (one type of) provenance.

The Privacy Act requires an agency to log to whom it has released data when and for what purpose; it also requires the agency to provide that log information to the person who is the subject of the information upon request. One place this sort of information might be found is "header data". If we can capture the header data from a transfer and semantically tag it, we can make available to any reasoner the details of which data, to whom, and date/time of transfer. Inferring the purpose may be achievable in some cases and in others we will need a person to assert the "purpose" for the transfer in order to be able to capture and reuse it. The Privacy Act also permits agencies to use data in ways which are "compatible with the purpose for which it was collected." If we capture the "purpose" of collection, we can make it available to any subsequent holder of the information no matter how many hands it has passed through since.

A federal regulation for the handling of criminal intelligence information by certain state systems that have received federal funding [4], requires that the information be checked periodically for relevance and importance and that all misleading, obsolete or otherwise unreliable information be deleted. It requires the agency to retain the name of the review, date of the review, and reasons for changes. It also requires the agency to notify all agencies that previously received the information.

The Privacy Act also permits an agency to write different usage rules -- known as Routine Uses – for data in each of its repositories – known as Systems of Records. In order to apply the correct rule set, the system must be able to tell which repository the data is sitting in – the

formal name of the system and the organization which owns it.

We are using these sorts of requirements as the basis of our knowledge model. We are grounding the application of these requirements in use cases describing information exchanges that are typical in a variety of public and quasi-public entity scenarios..

Action Context

In one of our first scenarios, we modeled an MIT academic discipline proceeding. A student is accused of lying about being ill in order to obtain an extension. The only proof is the data showing that the student used her "prox" card to enter a campus lab within an hour of calling in sick. However, the MIT privacy policy [6] regarding the use of such surveillance data is limited to criminal proceedings. We recognized that "purpose" of use was the critical contextual factor and added it to the rule pattern:

```
air:pattern
            {
                <mit-policy#U>
                    air:data
                       <mit-policy#D>;
                    air:purpose
                       <mit-policy#P1>;
                    a    air:UseEvent.
                <mit-policy#D> a
mit:ProxCardEvent. };
    air:rule
       <mit-policy#MITRule2>;
    a    air:BeliefRule.
<mit-policy#MITRule2>
    air:alt
       [ air:rule <mit-policy#MITRule3> ];
    air:assert
       { <mit-policy#U> air:compliant-with <mit-
policy#MITProxCardPolicy>. };
    air:description
       ( <mit-policy#P1> " is same as a criminal
activity" );
    air:pattern
       { <mit-policy#P1> <mit-policy#sameAs>
pur:criminal-law-enforcement. };
    a    air:BeliefRule.
<mit-policy#MITRule3>
    air:assert
       { <mit-policy#U> air:non-compliant-with
<mit-policy#MITProxCardPolicy>. };
    air:description
       ( <mit-policy#P1> " is different from a
criminal activity" );
    air:pattern
       { };
```

When run against our hypothetical Committee on Discipline request for the prox card data, the reasoner properly determines that the prox card transaction record cannot be used for this purpose.

Most of our scenarios have focused on the context of use. We modeled a case in which a phone company couldn't use an inference of possible patient illness to deny a service visit, although it could have used that same inference to release the customer's records to the CDC for an epidemic investigation. In another, the FBI could use information sent to their attention for the purpose of terrorism investigation but not for the purpose of pursuing a "Deadbeat Dad." And, in one more hypothetical, the use of data was found non-complaint with the Geneva Convention's human rights policies when used as the basis for a counterterrorism search but would have been compliant if used for relief work.

More difficult challenges remain. Most notable among them may be the legal context known as "condition subsequent," the rule structure which says something is permissible only if it meets the requirement of something occurring afterwards. Conditions subsequent, such as deletion of the data after a specified number of days or a requirement that the subject of the data be notified after use, are common in privacy-protecting policies.

Actor Context

Much work is being done to better identify individuals to systems. The survey of the current state of the field by Halperin and Backhouse [6] includes the importance of context in distinguishing identities of an individual. This is important to our problem because, for example, different privacy rules apply to the acts of a private citizen and those of a law enforcement officer, even if it is the same person at different times of the day. However, the rules we are considering also want to know the context within the identity.

By "context within the identity" we mean that a person who is acting in a particular role is subject to different strictures depending upon the context of the moment. At a simple level, a person's ability to use information may be geographically dependent. For example, most jurisdictions restrict law enforcement access to criminal investigative data to personnel working within that jurisdiction and it is relatively easy to locate a person's assigned work location in a personnel file and determine whether the person falls within the stricture. However, in emergencies, people are detailed to other locations and the information often doesn't get timely recorded in the personnel file. A decentralized system could retrieve the relevant temporary assignment information from a travel reimbursements file, if the representations were sufficiently granular.

The Privacy Act presents a more difficult requirement, when it says that a federal agency may release information if it has received a written request for the information from the head of a law enforcement agency. This requires us to be able to determine that the actor is aware, or has received instructions from someone who is aware, that the agency has received such a request. If the request was received digitally, the system could determine from log files if the actor had received or accessed that written request.

In one of our scenarios, a public utility customer service representative refuses to send a serviceman to a customer she thinks might have a highly infectious form of tuberculosis. In the first version we modeled, she expressly stated the belief. At the time, we were not yet representing actor context and opted for a simplified version of disability law that made any denial of service illegal if based on health information. The Americans with Disabilities Act, though, prohibits this sort of discrimination based upon a "perception of" disability. This requires the more difficult ability to collect and represent the actor's belief as the context within which she is attempting to use information. (This is an important capability because it is the same problem as capturing whether an officer is acting with or without probable cause, often the determinative factor for whether the access of private information was appropriate or abusive.) In the second version, we infer her belief based upon a showing that she contacted another individual in the company, he accessed information indicating an earlier CDC request for the customer's toll records as part of an infectious disease investigation, and he then transferred that information to the customer service representative:

```
<http://tw.rpi.edu/proj/tami/Special:URIResolver/
Xphone_Database_discloses_Alex_some_records_about
_Bob_Same_(Event)>
      a       swivt:Subject , wikic:Event ;
      wikip:Coordinator
              wiki:Xphone_Company ;
      wikip:Datetime
              "2007-08-
17T11:09:00"^^<http://www.w3.org/2001/XMLSchema#d
ateTime> ;
      wikip:Description
              "Xphone database sends the records
to        Betty        as        her        query
result"^^<http://www.w3.org/2001/XMLSchema#string
> ;
      wikip:Operation
              wiki:Disclose_Data ;
      wikip:Output_data

<http://tw.rpi.edu/proj/tami/Special:URIResolver/
Xphone_record_no.2892-
3A_Xphone_disclosed_record_351_to_CDC_for_TB_inve
stigation> ;
      wikip:Participant
              wiki:Alex_Bialoski .

<http://tw.rpi.edu/proj/tami/Special:URIResolver/
Alex_reply_Betty_that_-22Bob_Same_is_blacklisted-
22_(Event)>
      a       swivt:Subject , wikic:Event ;
      rdfs:label  "Alex  reply  Betty  that  \"Bob
Same is blacklisted\" (Event)" ;
      wikip:Coordinator
              wiki:Alex_Bialoski ;
      wikip:Datetime
              "2007-08-
17T11:10:00"^^<http://www.w3.org/2001/XMLSchema#d
ateTime> ;
      wikip:Description
              "reply  Betty  that  \"Bob  Same  is
blacklisted\""^^<http://www.w3.org/2001/XMLSchema
#string> ;
      wikip:Operation
              wiki:Disclose_Data ;
      wikip:Output_data

<http://tw.rpi.edu/proj/tami/Special:URIResolver/
Xphone_record_no.3015-3A_Bob_Same_is_blacklisted>
;
      wikip:Participant
              wiki:Betty_Jo_Bialoski .
```

Environment Context

Within the law, there have always been some rules that address changing standards depending upon something else in the environment, such as the level of emergency. Not only are police given certain additional leeway when imminent death or severe injury is at risk, but so, too, are there exceptions in the accommodation of disability if it poses a significant health risk to others. Generally, these rules will require an ability to infer something directly related to the situation for which information is being used.

Since 9/11, government agencies have raised the issue of "breaking the glass" scenarios in which different data handling rules do or should apply. This is the concept of adapting to risk by seeking inference of environmental conditions beyond the particular data transaction. Some situations, such as the predicted arrival of a gale force hurricane, the declaration of a military environment ("DEFCON 4"), or the aftermath of a terrorist attack should be inferable in a web environment.

An everyday example of this sort of this environment-aware adaptability is the Freedom of Information Act waiver excusing agency's from revealing criminal case information while the investigation is ongoing and the subject is unaware:

> "(1) Whenever a request is made which involves access to records described in subsection (b)(7)(A) and—(A) the investigation or proceeding involves a possible violation of criminal law; and (B) there is reason to believe that (i) the subject of the investigation or proceeding is not aware of its pendency, and (ii) disclosure of the existence of the records could reasonably be expected to interfere with enforcement proceedings, the agency may, during only such time as that circumstance continues, treat

the records as not subject to the requirements of this section." [7]

The waiver lasts only so long as there is a risk that the release of information would disrupt the criminal case. This is much harder to determine than simply whether a case is open or closed and may always require a human assertion. To understand the environmental context, it is important to remember that the request for information may be entirely unrelated to the case which is causing the decision to grant or deny access.

In one of our previous scenarios,, we used the Florida Sunshine Law [8] to decide whether or not to release information. Under that rule the decision hinged on whether a criminal investigation was active, which we inferred by determining whether the case was open or closed.

```
    air:rule [
      air:description ( "It is permitted that
" :P " disclose " :D " to " :ACCESSOR ",
because" :P " is a Florida Government Agency" );
      air:pattern {  :E tamip:Coordinator :P . };
      air:assert {  :E  air:compliant-with
tami:FS_119_01_1.
                    :E  tamip:applicable
tami:FS_119_01_1.   }
    ];
    air:rule [
      air:description ( "It is permitted that
" :P " disclose " :D " to " :ACCESSOR ",
because" :P " is affiliated with a Florida
Government Agency" );
      air:pattern {  :E tamip:Coordinator [
tamip:Affiliation :P ]  };
      air:assert {  :E  air:compliant-with
tami:FS_119_01_1.
                    :E  tamip:applicable
tami:FS_119_01_1. }
    ]
  ].
@forAll   :EVENT, :DATA , :CASE, :EO, :EC .

:FS_119_071_2_c_1 a  air:Policy;
  rdfs:label "Fla. Stat. Ch. 119.071(2)(c)1";
  air:rule [
    air:pattern {
       :EVENT  tamip:Output_data :DATA  ;
               tamip:Operation tami:Disclose_Data
.

       :DATA  a
tamic:Criminal_Intelligence_Information.
    };
    air:rule[
      rdfs:label "Fla. Stat. Ch. 119.011(3)(d)";
      air:pattern {
       :EVENT tamip:Relation :CASE ;
              tamip:Antecedent :EO.

       :CASE a tamic:Criminal_Investigation.

       :EO a tamic:Event;
          tamip:Operation  tami:Open_Case;
          tamip:Relation :CASE .

       :EC a tamic:Event;
              tamip:Operation  tami:Close_Case;
              tamip:Antecedent :EVENT;
              tamip:Relation :CASE .
      };
      air:assert {
            :DATA air:compliant-
with :FS_119_071_2_c_1 .
            :EVENT air:non-compliant-
with :FS_119_071_2_c_1 .
             :EVENT tamip:applicable
 :FS_119_071_2_c_1 .
       }
    ]
  ].

:FS_119 a  air:Policy;
  rdfs:label "Fla. Stat. Ch. 119";
  air:variable  :E,
 :POLICY1, :POLICY2, :PROPERTY;
  air:rule [
    air:pattern {
       :E  tamip:applicable  :POLICY1 .
       :E  :PROPERTY :POLICY1 .
    };
    air:rule [
      air:pattern {
       :E  tamip:applicable  :POLICY2 .
       :POLICY2 tamip:Overrides :POLICY1  .
      } ;

      air:alt [
         air:assert {  :E :PROPERTY  :FS_119 . }
      ]
    ]
  ].
```

Though in a perfect world that fact would always be tagged by the person responsible for the case, we know that this doesn't always happen. A case is held open pending some other event and then can be forgotten upon the transfer of the responsible officer or the long passage of time. In future, we will work to model an inference of "inactive" based upon the passage of time. With such added data context inference, the reasoner could properly determine whether to release the information.

Rule Context

Rule context may be the hardest to address. An accountable system needs not only to be aware of potentially relevant policies but to be able to determine which apply.

At a conceptual level, we do not limit our definition of privacy rules to those marked "privacy." Instead, we consider that privacy protection is about limiting the use of personal information and so our view encompasses rules that intend to keep information about individuals out of inappropriate hands or from being used inappropriately. From that perspective, we include the broad spectrum from privacy to anti-discrimination and from sunshine laws to grand jury rules.

An effective policy aware system must be aware of when policies are in effect. In a real-time enforcement system, it must be able to apply the "current" policies. In an audit system, it must be able to use the time of the data event to find the version of the policy with an effective date range including that time. Temporal reasoning must be applied to the granular representation of effective date ranges.

A properly functioning accountable system must be able to address rule conflicts. These may occur when a legislative body enacts a new law, or an executive issues a new policy, without explicitly rescinding a piece of an older one. This can also occur when two jurisdictions each have an interest in the same transaction – an issue which arising increasingly in the context of cross-border eDiscovery. In one of our scenarios, we observed the Florida law [9] which grants Florida law enforcement the ability to supersede Florida disclosure laws with the more restrictive law of another jurisdiction, if that is a condition of receipt for Florida. Or, it can appear to occur when an in-house counsel opinion provides an interpretive overlay to an existing law or regulation. The ability to represent laws and policies in the appropriate sub-classes (level of government, branch of government, etc), will make it possible to apply the standard conflict resolution logic that lawyers apply mentally today.

Conclusion

Privacy is different from seclusion. The latter is simply a cutting off of contact, while the former is the selective elimination of contact. Throughout this paper we have shown that today's rules for creating and respecting privacy are selective, picking and choosing who can see or use what information based upon the context of the proposed use. We have shown that to be accountable for the implementation of such rules requires recognizing the relevant context. And, we have described why semantic technologies provide enhanced ability to find, and sometimes infer, complex context information.

Acknowledgments

This work is a result of joint projects and many illuminating discussions with Harold Abelson, Tim Berners-Lee, Chris Hanson, James Hendler, Lalana Kagal, Joe Pato, Gerald Jay Sussman, and Daniel J. Weitzner. It has been partially funded by NSF Cybertrust Award #0524481, DTO NICECAP Award #FA8750-07-2-0031, and IARPA (DHS Fusion Center Award).

References

[1] Oxford English Dictionary 2009 (online) (citing J. Imrie et al., *Burgh Court Bk. Selkirk*, 141 (1960)).
[2] 5 USC § 552a.
[3] Weitzner, Abelson, Berners-Lee, Hanson, Hendler, Kagal, McGuinness, Sussman, Waterman, Transparent Accountable Data Mining: New Strategies for Privacy Protection,; MIT CSAIL Technical Report MIT-CSAIL-TR-2006-007 [DSpace handle] (27 January 2006).
[4] 28 CFR part 23.
[5] "Am I Being Tracked?", Card Services, Privacy Policy, MIT (2009)
[6] Halperin, Ruth; Backhouse, James (2009), "A roadmap for research on identity in the information society", *Identity in the Information Society* (Springer) **1** (1), at 81.
[7] 5 USC 552(c)(1).
[8] FSL § 119.01, *et seq*.
[9] FSL § 119.071(2)(b).