

Content-based Privacy Management on the Social Web

Michal Jakob, Zbyněk Moler, Michal Pěchouček
Agent Technology Center, Dept. of Cybernetics
Faculty of Electrical Engineering, Czech Technical University
Praha, Czech Republic
{jakob, moler, pechoucek}@agents.fel.cvut.cz

Roman Vaculín
IBM T.J. Watson Research Center
Hawthorne, NY 10532 USA
vaculin@us.ibm.com

Abstract—Protection of privacy is a major concern for users of social web applications, including social networks. Although most online social networks now offer fine-grained controls of information sharing, these are rarely used, both because their use imposes additional burden on the user and because they are too complex for an average user to handle. To mitigate the problem, we propose an intelligent privacy manager that automates the assignment of sharing permissions, taking into account the content of the published information and user’s high-level sharing policies. At the core of our contribution is a novel privacy policy language which explicitly accounts for social web concepts and which balances the expressive power with representation complexity. The manager employs named entity recognition algorithms to annotate sensitive parts of published information and an answer set programming system to evaluate user’s privacy policies and determine the list of safe recipients. We implemented a prototype of the manager on the Facebook platform. On a small test scenario, the manager reached the F-measure value of 0.831 in correctly recommending safe recipients.

Keywords—privacy protection, social web, information extraction, policies, Facebook

I. INTRODUCTION

With the increasing volume of information shared through social networks, maintaining user privacy becomes a major problem. Among the causes of the problem are inadequate mechanisms for the control of information propagation in social networks. The majority of current systems restrict access based on the datatype, category or instance of the published information, combined with the social distance of the recipient (i.e. friends, friends of friends etc.). This is not well-aligned with human notion of privacy control – in reality, decisions about who should have access to a particular piece of information is greatly influenced by the content of the published information as well as the situation and the social context of the publishing user [1]. Another orthogonal problem is the complexity of sharing control settings, which are too complex for average users (e.g., Facebook has tens of privacy setting). Together, this contributes to wide-spread underuse of existing privacy control mechanisms on the side of users, and consequently to risky sharing behavior.

We address the problem by proposing a mechanism that allows the user to define *sharing policies*, which explicitly consider information content and social context, and apply

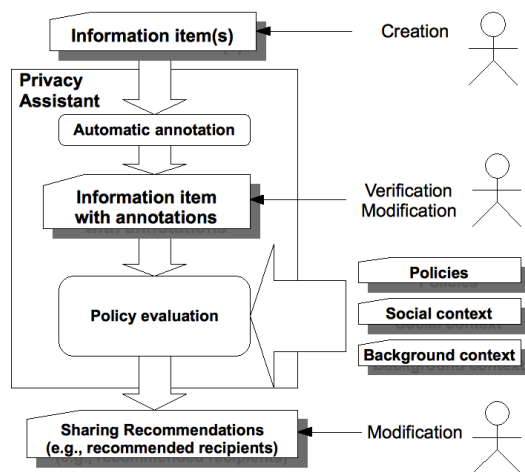


Figure 1. Overview of the proposed privacy management mechanism

these policies to automatically determine sharing permissions, thus reducing the need for manual privacy control on a case-by-case basis.

II. OVERVIEW OF THE APPROACH

Figure 1 shows the overall schema of the proposed mechanism. The mechanism takes three inputs: (1) *information item* – the information the user aims to publish on the social network; (2) *social and background context* – information about the user, her social relationships, and other information affecting sharing decisions; (3) *sharing rules/policies* – sharing policies representing user’s sharing preferences and restrictions. Each input is represented using the SWPL language as described in Section III.

Given the inputs, the mechanism determines a set of users that should be allowed to see the published information. The recommendation process consists of two main steps. The first step, *automated privacy annotation* takes the information item, analyses its content, and provides a set of (semi-)automatically extracted annotations that represent the sensitive aspects of the information item. In the second *policy evaluation* step, the reasoner takes as input the annotated information item, all information about the user’s social

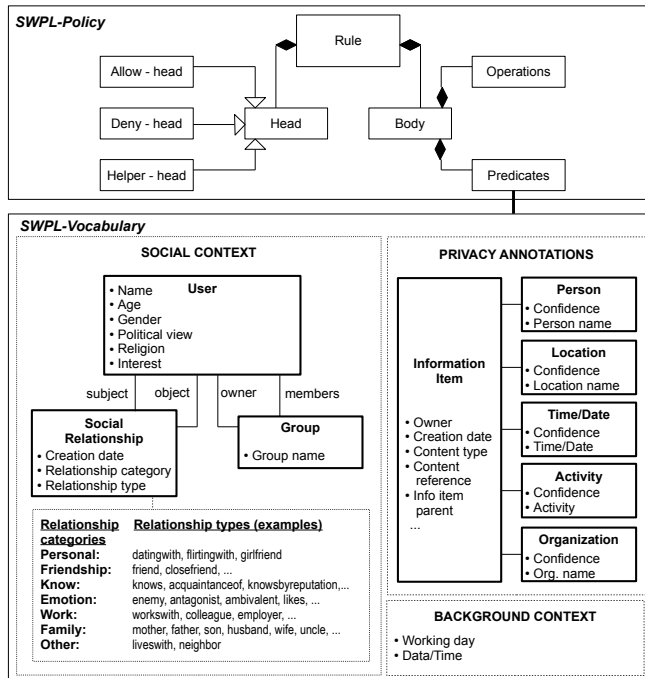


Figure 2. Overview of the SWPL language structure and constructs

and background context, and the user’s sharing policies, and it produces recommended sharing permissions for the information item.

III. SOCIAL WEB PRIVACY LANGUAGE

The *Social Web Privacy Language (SWPL)* is a key part of our contribution. It is used for (1) representing all inputs affecting the decision with whom a particular information item should be shared and (2) representing user’s sharing preferences in a form of machine interpretable, extensible and reusable policies. The SWPL language has two modules, the *SWPL-Vocabulary* and *SWPL-Policy*, that respectively address the two purposes.

A. *SWPL-Vocabulary: Vocabulary for Social Web Privacy*

The *SWPL-Vocabulary* defines a set of concepts by which privacy-affecting aspects of published content and context can be described. We provide a small set of concepts capable of capturing most relevant aspects affecting user’s decisions about information sharing. We define three groups of concepts: (1) privacy annotations, (2) social context and (3) background context.

1) *Privacy Annotations of Information items*: As an abstraction of various pieces of information that users publish (e.g., status updates, photos, comments, etc.), the *SWPL-Vocabulary* uses the concept of an information item. To capture the sensitive aspects of information items in a unified, structured way, we define several types of *privacy annotations*. Privacy annotations can be referenced from

sharing policies and thus enable to include the content of published information in automated sharing decisions – this is feature not supported by existing privacy languages. Privacy annotations themselves can either be extracted automatically from the content or assigned manually by users.

Inspired by the concept of *Five Ws (who, what, when, where, why)*¹ employed in journalism and police investigation as basics of information gathering, we define the following types of privacy annotations:

- *location* (where) – annotates information items that refer to a specific location
- *person* (who) – annotates items that refer to a specific person
- *time/date* (when) – annotates items that refer to a particular time
- *activity* (what) – annotates items that refer to a particular activity
- *organization* (“who”) – annotates items that reference to an organization

Generally the more privacy annotations associated with an information item, the higher the potential sensitivity of the information. Note that when speaking about sensitivity, our interest goes beyond personally identifiable information (such as names, telephone numbers, addresses, credit card numbers etc.), which is typically the subject of protection in information security.

2) *Social Context/Graph*: The second group of *SWPL-Vocabulary* concepts is used to describe the social context in which the user publishes information. Support for social context is the second novel feature of *SPWL-Vocabulary*. The choice of supported concepts is based on a comprehensive study of social constructs supported by systems, standards and languages related to the social web². Users, social relationships and user groups are represented with focus on attributes that may affect sharing decisions (see Figure 2 for a list).

3) *Background Context*: The third group includes concepts providing additional context related to the information being published and to the user(s). Example background concepts include generic notions of time or location of publication as this may affect sharing decisions.

Detailed description of the *SWPL-Vocabulary* including all concepts, attributes and predicates can be found in [2].

B. *SWPL-Policy: Policies for Social Web Privacy*

Building on the vocabulary for social web privacy, we define a rule-based language for expressing user’s sharing policies. The language builds on the established notions of *permissions* and *prohibitions* known from deontic logics and

¹See e.g. http://en.wikipedia.org/wiki/Five_Ws

²We analyzed constructs from the FOAF ontology <http://www.foaf-project.org/>, vCard standard <http://www.w3.org/TR/vcard-rdf/> and its RDF variant <http://www.w3.org/TR/vcard-rdf/>, Facebook, and Google data and open social API <http://code.google.com/apis/opensocial/>.

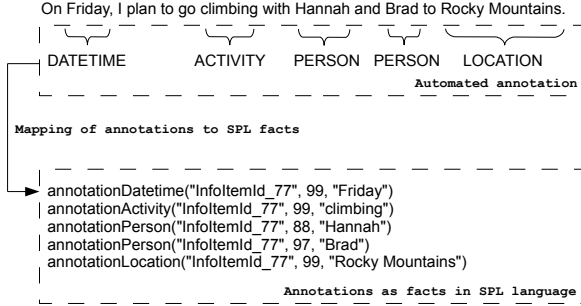


Figure 3. Example annotated information item. Annotation predicates and confidence levels at the bottom.

employ rule priorities for conflict resolution. The head of a rule is either of the form *allow*, *deny*, *helper-predicate*. The body of a rule can include any predicate defined in the SWPL-Vocabulary or additional, user-defined helper predicates, which allow simplifying complex and/or repeated expressions, including recursion. Predicates in the rule body can be negated and combined using *and* and *or* logical connectives. Arithmetic operations (addition, multiplication), comparison and aggregation operators (count, sum, times, min, max) can also be used. Negation has the negation-as-failure semantics, though depending on the reasoner used, true negation can be also supported.

IV. AUTOMATED ANNOTATION

To reduce the need for direct user input, the privacy manager employs automated privacy annotation of text information items. For a text input, each generated privacy annotation consists of a substring of the original text and a privacy annotation predicate (as defined in Section III-A) assigned to the substring. Figure 3 shows an example sentence with associated privacy annotations and their confidence levels as calculated by the annotation module.

We cast the extraction of privacy annotations as a *named entity recognition (NER)* problem and use established NER algorithms and tools to implement it. To support all required annotation types and to reduce annotation errors, we combine outputs from multiple NER algorithms into a single, more reliable output using two ensemble classification aggregation methods— *majority voting* for NER algorithms that only provide annotations and the more informative *Bayesian model averaging* for NER algorithms that also provide confidence levels. The integration of individual NER algorithms was done within the *GATE framework* [3] which provides a uniform interface to various NER algorithms. We used the following NER systems in our implementation: Name Finder from OpenNLP, Stanford NER, NER from LingPipe and Gazetteer from the GATE framework.

V. EVALUATION

In order to evaluate the proposed mechanism, we implemented a prototype of the *Privacy Assistant* as a Facebook

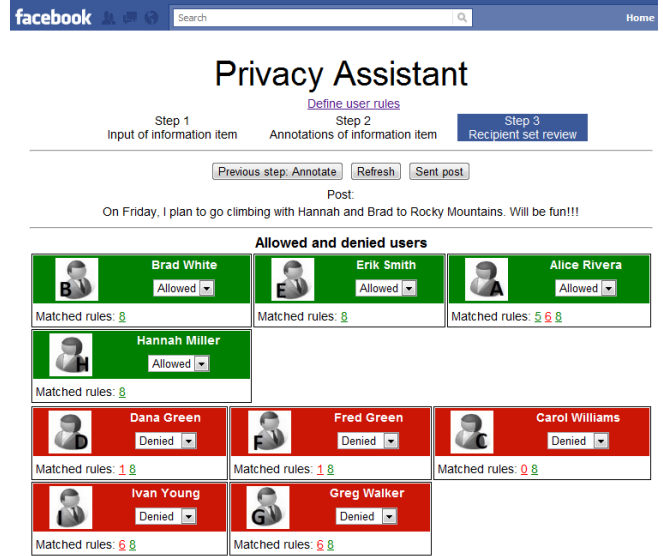


Figure 4. Output of the recommendation step of the Privacy Assistant. Allowed (green) and denied (red) recipients for the user’s post were selected according to the user-defined sharing policy.

application³ (Figure 4). DLV reasoner [4] was used for policy evaluation. More information about the prototype is available in [5].

To evaluate the mechanism, complete information about a user’s social network, information she publishes, and sharing settings she assigns to published posts is needed. As there is currently no available Facebook dataset that would provide such information, we created a synthetic dataset based on a realistic scenario and samples of real posts on Facebook⁴. The scenario involves a social network with 10 users and 110 status updates with different sensitivity levels. For each update, we recorded the correct complete set of privacy annotations and the set of recipients with whom the publishing user would like to share the update. The scenario contained 8 sharing rules for the publishing user (i.e. the user from whose perspective the system is evaluated).

We used standard metrics from machine learning and information retrieval to evaluate the performance of the system: *recall*, *precision* and *F-measure*. As ground truth served the information from the test dataset.

First, we evaluated performance of automated annotation alone – the results for individual annotation types and the overall average are given in Table A of Figure 5. Second, we evaluated the mechanism as a whole. Recommendation performance was measured by comparing the recipients suggested by the mechanism with the recipients specified as correct in the testing data. This was done for two operation modes: (1) fully automated operation (i.e. only automated

³A demonstration video of the application is available at project’s web site <http://agents.felk.cvut.cz/projects/privacy20/>

⁴<http://www.facebook.com/group.php?gid=51097603092&v=wall>

(A) Annotations performance			
Type of annotation	F-measure	Recall	Precision
LOCATION	0.720	0.643	0.818
PERSON	0.745	0.729	0.761
ORGANIZATION	0.400	0.333	0.500
ACTIVITY	0.644	0.633	0.655
DATETIME	0.541	0.526	0.555
Average (all types)	0.680	0.655	0.705
(B) Recipient recommendation performance			
Operation mode	F-measure	Recall	Precision
Semi-automated	0.823	0.764	0.893
Fully automated	0.831	0.738	0.952

Figure 5. Performance on posts from the test scenario

annotation was used), and (2) semi-automated automation where the user manually corrected the proposed privacy annotations (and they were thus entirely correct). The results for both cases as evaluated on the test dataset are given in Table B in Figure 5 and show very good agreement between recommended recipients by the mechanism and user’s expectations.

VI. RELATED WORK

Several languages for describing privacy policies and controlling access to data have been proposed [6], [7]; none of them, however, provides any support for content references and social network concepts. Furthermore, the focus of existing languages is on preventing security issues and on controlling access to personally identifiable information. Our work extends the control also to information that does not directly lead to security risks but whose disclosure can harm user’s reputation and credibility.

Various approaches (e.g. [8]) have been explored to help protecting privacy on the social web; none of them, however, combines information extraction and content-sensitive information access/sharing policies. In privacy protection, information extraction tools have been used for automated de-identification/anonymization (e.g. [9]) and content-based privacy protection for document sharing/e-mails (e.g. [10]). The key difference to our work is the lack of consideration of any social concepts.

VII. CONCLUSION

We proposed a mechanism that aims to simplify handling of potentially sensitive information on the social web by (partially) automating decisions with whom given information should be shared. As part of the solution, we proposed a novel language which supports content annotation based on the five W’s concept and explicitly represents user’s social context in the specification of sharing policies. We implemented a Facebook prototype of the mechanism and showed that it can achieve good performance on a small test set. The policies that can be defined by the proposed

mechanism can cover a wide range of situations. The danger, however, is that the number and complexity of rules required to capture preferences of an average user can be high. In the longer term, we therefore aim to complement the explicit policy specification with a learning mechanism which would discover user preferences from his past sharing decisions.

ACKNOWLEDGMENT

Supported by a Google Research Award and by the Czech Ministry of Education, Youth and Sports under grant "Decision Making and Control for Manufacturing III" (grant no. MSM 6840770038).

REFERENCES

- [1] D. Boyd, "Making sense of privacy and publicity," in *South by Southwest (SXSW 2010) – transcription of the talk*, March 2010.
- [2] Z. Moler, M. Jakob, and R. Vaculín, "Social web privacy language," Czech Technical University, Tech. Rep., 2011, available at <http://agents.felk.cvut.cz/cgi-bin/docarc/public.pl/document/331/SocialWebPrivacyLanguage-Report.pdf>.
- [3] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "GATE: A framework and graphical development environment for robust NLP tools and applications." in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA*, 2002.
- [4] N. Leone, G. Pfeifer, W. Faber, T. Eiter, G. Gottlob, S. Perri, and F. Scarcello, "The DLV system for knowledge representation and reasoning," *ACM Transactions on Computational Logic*, vol. 7, no. 3, 2006.
- [5] M. Jakob, Z. Moler, R. Vaculín, and M. Pěchouček, "Intelligent content-based privacy assistant for facebook," in *Proceedings of the 2011 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 2011.
- [6] A. Matheus, "How to declare access control policies for XML structured information objects using OASIS’ eXtensible Access Control Markup Language (XACML)," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*. Washington, DC, USA: IEEE Computer Society, 2005.
- [7] P. Ashley, S. Hada, G. Karjoth, C. Powers, and M. Schunter, "Enterprise privacy authorization language (EPAL 1.2)," W3C, Tech. Rep., 2003.
- [8] L. Fang and K. LeFevre, "Privacy wizards for social networking sites," in *Proceedings of the 19th international conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010.
- [9] B. Medlock, "An introduction to NLP-based textual anonymization," in *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 2006.
- [10] N. Boufaden, W. Elazmeh, Y. Ma, S. Matwin, N. El-Kadri, and N. Japkowicz, "Peep- an information extraction base approach for privacy protection in email." in *Proceedings of the 2005 Conference on E-mail and Anti-spam*, 2005.