Private Data Management in Collaborative Environments[1]

Larry Korba, Ronggong Song, George Yee, Andrew S. Patrick,
Scott Buffett, Yunli Wang, Liqiang Geng

Institute for Information Technology, National Research Council of Canada
Building M-50, Montreal Road, Ottawa, Ontario K1A 0R6
{Larry.Korba, Ronggong.Song,
George.Yee, Andrew.Patrick, Scott.Buffett, Yunli.Wang, Liqiang.Geng}
@nrc-cnrc.gc.ca
http://iit-iti.nrc-cnrc.gc.ca

Abstract. Organizations are under increasing pressures to manage all of the personal data concerning their customers and employees in a responsible manner. With the advancement of information and communication technologies, improved collaboration, and the pressures of marketing, it is very difficult to locate personal data is, let alone manage its use. In this paper, we outline the challenges of managing personally identifiable information in a collaborative environment, and describe a software prototype we call SNAP (Social Networking Applied to Privacy). SNAP uses automated workflow discovery and analysis, in combination with various text mining techniques, to support automated enterprise management of personally identifiable information.

Keywords: Privacy, compliance, workflow, social network analysis.

1 Introduction

The quantity of personal data that organizations must manage is increasing at a phenomenal rate. The main reason is the dramatic increase in the exploitation of communication and network technologies for collaboration, marketing, and sales. Other contributing factors include competitive pressures, as well as inexpensive computers and mass storage. While the amount of personal data is increasing, the prevalence of computer and network-based collaborations has made it very difficult for organizations to know where all the private data is stored and exactly how it is being used. This can occur despite attempts at controlling access to the data through centralization.

In the reality of the collaborative environments of today, the prospect of assuring privacy compliance, as may be required by legislation, regulations, or best practices, has become almost impossible. Our approach towards a solution is to combine several technologies in a manner that would allow organizations to understand and manage the life cycle of private data. The different technologies include: private data discovery, social network (workflow) analysis, knowledge visualization, and effective hu-

[1] National Research Council Paper Number 49356

man-computer interaction operating within a policy enforcement framework. Private data discovery involves text and data mining techniques to determine the location and use of personally identifiable information (PII) in different contexts across an organization. Social network analysis produces an understanding of workflow activities related to PII, providing measures for assessment of compliance with privacy policies, and a means for performing forensics analyses on the activities related to private data. In this paper we detail the challenges organizations have with privacy compliance, describe our progress in the development of a prototype for an automated privacy compliance system, outline early results, and detail the further challenges we are exploring.

2 Problem Description

The challenges for businesses today in the handling private data can be understood by referring to Figure 1. Key factors that put pressures on increasing the amount of private data collected include:

- Cheap Storage. Storage costs continue to drop so there are few impediments to collecting growing amounts of data of all sorts.
- Expanded Services. As new services are put in place, often more PII in different forms is collected and stored in order to assess product or service quality, and to facilitate follow-on sales.
- Marketing and Competition. Marketing pressures stemming from the desire to improve existing products and services or from competitive market conditions often lead to the collection and retention of more PII, e.g. e-service personalization where the consumer's contextual product selections are tracked for service improvement.
- Increasing Client Numbers. As an organization provides more products and services, and they become popular, more clients are garnered, leading to the collection and retention of larger amounts of PII.
- Computers Everywhere. Within organizations, desktop, portable, and handheld computers are being deployed at a high rate, due to their convenience and decreasing costs. The computers enable staff to share the work of creating and delivering services and products, which can lead to distributed, local storage of PII.

The pressures to decrease the amount of PII stored and managed within an organization include the following:

- Risk Management. Loss of customer PII is injurious not only to the customer but also to the organization. Data breaches can lead to lost sales due to a decrease in client trust. These risks can be reduced by minimizing PII collection and retention.
- Regulations and Policies. An organization may operate in a regulated sector (healthcare, banking, legal services, gaming, etc.) where there are specific, mandatory requirements for PII handling. In addition, an organization may have its own policies to manage its business and to evoke a stronger level of client trust.
- Legislation Enforcement. Beyond regulatory requirements, some jurisdictions, such as the European Union, may have legislation in place specifying how different types of PII must be handled.

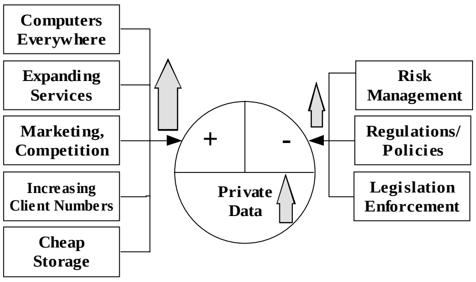


Figure 1. This is a schematic representation of the pressures on organizations concerning their collection and handling of personal data. The magnitude of the pressures to increase the collection of data (left side) is currently greater than pressures to control access and use (right side).

While the analysis depicted in Figure 1 illustrates the pressures that are leading to increases in the amounts and type of personal data collected and retained, another consequence has to do with how technologies are used. With the widespread availability of computers, networks, and collaboration tools, there is a dramatic increase in the numbers of work artifacts that may be shared amongst different staff. This may be by design (dictated by the organizational workflow), or by necessity (for instance, deferring to the shared experience of others or spreading the work among different geographic locations). The result is that PII may be readily shared amongst many different users, leading to challenges for the organization to understand fully where private data is stored, whether it is accurate, and how it has been used. It is within this context that we are developing technology to make it easier for organizations to manage personally identifiable information.

3 Our Approach

In our approach, which we call Social Network Applied to Privacy (SNAP), we have combined data mining techniques to discover PII, social network analysis to reveal workflows, automated analysis for decision making on PII usage, and data visualization techniques to improve understanding within in the organization, all mediated by electronically-readable policies. The system architecture is agent-based, with the SNAP agent, as described in [1], installed on every computer within the enterprise.

3.1 Overall Design

Figure 2 provides a high-level view of the SNAP system. The system consists of separate parts or modules that perform different functions:

- Machine-readable rules express privacy and security policies, and the prescribed workflow for the organization. This information is interpreted to control the operation of the SNAP system.

- Data collection is performed by a SNAP monitor agent that collects data from the host computer system by searching the file systems and monitoring system activities.

- Analysis involves the discovery and affirmation of personally identifiable information, examining any relationships among the personal data, and the determination of PII work flows based upon comparisons of local activities with the activities of others in the organization. Outputs from the analyses processes are displayed to the system operators and, depending upon the appropriate policies, can lead to automatic actions.

- The display functions present different aspects of the collected and analyzed data to the system operators. The displays include tables of raw data and graphical images of social networks based on the correlated operations by multiple users.

- Actions performed by the system are determined by policies. They can be prescriptive or reflexive. Prescriptive actions are, for example, warnings to an end user or a system operator about a policy breach involving some from of personal data. Reflexive actions include taking security and access control measures to prevent a breach of a privacy policy before it occurs.

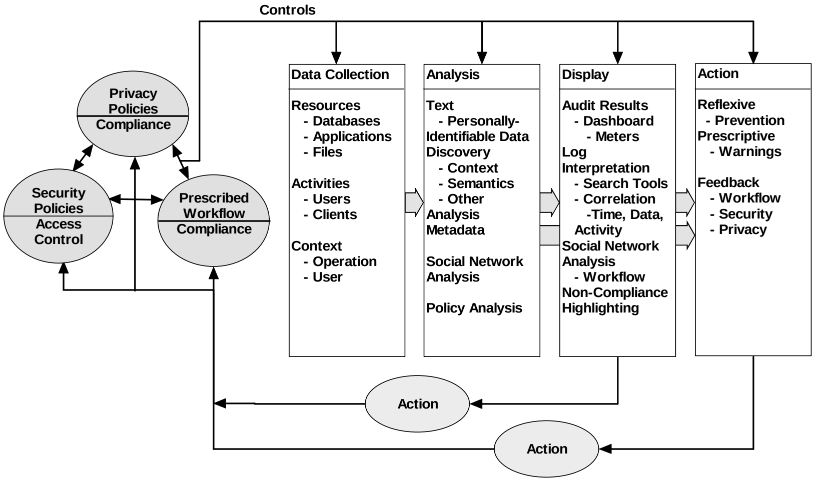


Figure 2. The overall framework for SNAP.

Figure 3 provides more details on the data collection and analysis portions of the system. At the lowest level, data is collected from two main sources: the file system and system hooks. When the SNAP agent is first installed, the file system is searched for PII. This allows for existing (historical) private data to be discovered, and for the data discovery to be updated should additional data be added to the system or the privacy policies changed.

Anytime that SNAP is running, system hooks are used to interrogate user activities (including current file activity) for the presence of PII. In order to lower the processing load and to limit the data collected, we restrict the data collection to certain privacy-sensitive contexts, such as when the user is using an email application or a web-based email account. In these cases, keystrokes and other information is passed onto the SNAP system for PII discovery. Discovery is currently done through regular expression matching and semantic analysis. The current prototype supports discovery of postal addresses, email addresses, titles, dates, dollar amounts, religion, and race.

Strings of characters that appear to be credit card or social insurance numbers are also assessed further using Luhn's algorithm [2]. Due to the potentially large and varied search scope, we have attempted to optimize the efficiency and order of the regular expression matching routines.

Local context discovery involves measuring the distance between locally discovered PII. This is done to assess whether disparate discovered PII should be grouped together (potentially as part of one record for one individual). For correlation discovery, the SNAP agents communicate with one another, on a peer-to-peer basis, to determine the PII events that are common among users. This information is used to determine the PII workflows within an organization.

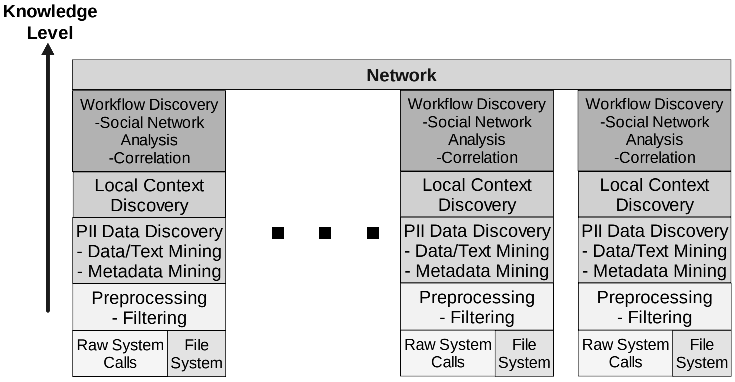

Knowledge Level
Network
Workflow Discovery -Social Network Analysis -Correlation
Local Context Discovery
PII Data Discovery - Data/Text Mining - Metadata Mining
Preprocessing - Filtering
Raw System Calls
File System

Figure 3. This figure presents the data collection and analysis hierarchy amongst a series of SNAP agents.

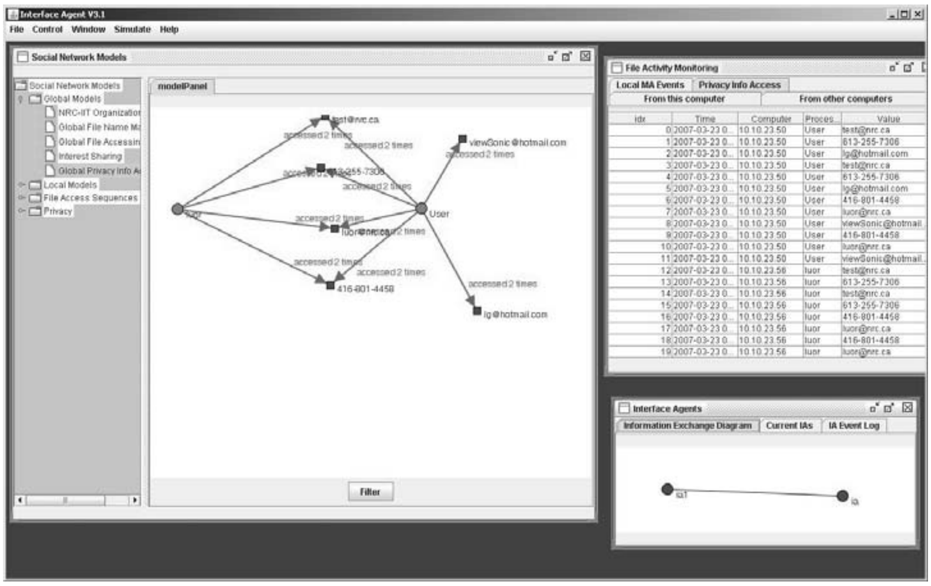
3.2 Implementation

Our SNAP prototype is implemented in C++, Java [3] and uses the Java Agent Development Environment (JADE) [4]. Each SNAP installation is comprised of a Monitor Agent, and an Interface Agent, and these agents are described in [1].

The discovery of PII, through file system searches or system monitoring, is captured in detailed event logs that are stored locally for later analysis and display. The event information is analyzed to build a model of local interactions with PII. These events can be checked against policies for permissible or prohibited authorizations, actions, types of data, or retention periods on a user by user basis. Additionally, in order to discover the workflow patterns, each SNAP agent communicates with the other agents operating on other computers in the enterprise eliciting requests for matches between with their locally discovered PII. If and when other agents find a match, it is reported back to the requesting agent.

A SNAP user can display reports about interactions on all local PII, or on all the interactions within the organization. The displays may be modified based upon time, type of data, agreement/disagreement with policy information, etc., or filtered to search for interactions of special interest.

4 Results



idx	Time	Computer	Proces...	Value
0	2007-03-23 0...	10.10.23.50	User	test@nrc.ca
1	2007-03-23 0...	10.10.23.50	User	613-255-7306
2	2007-03-23 0...	10.10.23.50	User	lg@hotmail.com
3	2007-03-23 0...	10.10.23.50	User	test@nrc.ca
4	2007-03-23 0...	10.10.23.50	User	613-255-7306
5	2007-03-23 0...	10.10.23.50	User	lg@hotmail.com
6	2007-03-23 0...	10.10.23.50	User	416-801-4458
7	2007-03-23 0...	10.10.23.50	User	luor@nrc.ca
8	2007-03-23 0...	10.10.23.50	User	viewSonic@hotmail...
9	2007-03-23 0...	10.10.23.50	User	416-801-4458
10	2007-03-23 0...	10.10.23.50	User	luor@nrc.ca
11	2007-03-23 0...	10.10.23.50	User	viewSonic@hotmail...
12	2007-03-23 0...	10.10.23.56	luor	test@nrc.ca
13	2007-03-23 0...	10.10.23.56	luor	613-255-7306
14	2007-03-23 0...	10.10.23.56	luor	test@nrc.ca
15	2007-03-23 0...	10.10.23.56	luor	613-255-7306
16	2007-03-23 0...	10.10.23.56	luor	416-801-4458
17	2007-03-23 0...	10.10.23.56	luor	luor@nrc.ca
18	2007-03-23 0...	10.10.23.56	luor	416-801-4458
19	2007-03-23 0...	10.10.23.56	luor	luor@nrc.ca

Figure 4. The SNAP interface when two individuals ("luor" and "User") access multiple forms of PII.

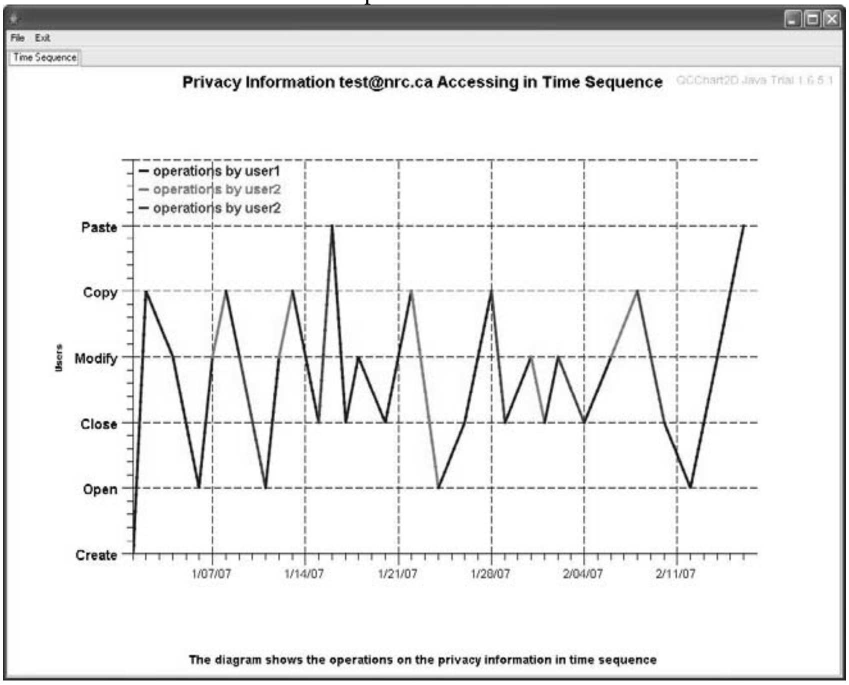


Figure 5. This figure illustrates a time sequence of operations on email address "test@nrc.ca" (private data in this case) over time, amongst three users.

The SNAP prototype features a multi-pane display, shown in Figure 4. A user can select from different display modes, filtering the data as required. Figure 4 shows the display of several different PII accessed by two users. In this example the users “luor” and “User” both accessed two email addresses and two telephone numbers, while “User” also accessed the email addresses “lg@hotmail.com” and “viewSonic@hotmail.com” separately. To the right of the graph is a table showing some of the data from the PII event logs. For privacy forensic purposes, it may be useful to see the different types of actions performed by different users at different times, as depicted in Figure 5. All of the graphical displays change dynamically as new PII is accessed, entered or discovered.

5 Related Work

Policy-based management of privacy has been under research for many years. For instance, IBM’s Enterprise Privacy architecture [5] and the SPARCKLE [6] project is an example where policies can be easily specified for user groups or information categories. IBM's system assumes that PII has a particular location where it can be monitored, such as a central database. Similarly, several smaller companies have begun to offer products that focus on various policy-based management approaches towards limiting insider threats (e.g., Vontu, Oakley Networks, and others) [7]. Our work assumes that PII may be anywhere in the enterprise, but builds on some of the techniques for insider threat detection, and focuses on the discovery of PII and social network analysis to discover privacy-related workflows to understand the context of work on PII and to support the overall enforcement of privacy policies.

In the academic research domain, systems like PRIVDAM [8] focus on protection of PII stored in centralized databases based upon analyzing access logs. Other work such as the “Trusted Privacy Manager” [9] deals with the management of outsourced private data, once again where all PII is stored in a centralized database and accesses are mediated by a trusted privacy manager. Our approach differs because we regard PII management as an enterprise-wide issue that requires the discovery of PII in many locations and an understanding of how it is being used throughout the enterprise. We accomplished this by distributed discovery of PII and workflow analysis to detect privacy violations.

While there has not been very much work published on scanning of documents for personally identifiable information, Aura et al. [10] describe a system that harvests text strings and applies regular expression matching to search for PII leakage in documents. Our work builds upon those ideas of scanning for PII, but includes a much wider scope of PII beyond user name and computer names to include street addresses, religion, etc. We also perform distributed searches for PII across all user activities within the organizations, as well as within files, correlating and presenting the findings across the enterprise, not just in specific documents.

6 Conclusions

Organizations are collecting more and more personal information, and facing increasing pressure to manage and to protect it. In this paper we have described the pressures on organizations regarding personally identifiable information and have introduced and described the development of our SNAP system, which can automate the management of personally identifiable information within an organization.

One of the technical challenges we face is to reduce the false positives and false negatives when discovering PII. We currently use regular expression matching and Luhn's algorithm to discover potential PII, and we measure the similarity between the discovered PII (time of occurrence, or location in files) to group disparate data that may be attached to the same user. We are currently assessing other means for discovering PII, including semantic feature analysis (looking for features around groupings of text that may indicate PII, even though it is not found with regular expression matching), and rule-based PII discovery (analysis employing dictionaries, domain ontologies, and rule-based matching). Additionally, we are expanding our use of context and workflow analysis to substantiate the discovery of PII.

Other technical challenges include PII contained within images. For instance, screen snapshots of PII records, or photographs of individuals, may represent sensitive information for individuals or organizations. It is a considerable technical challenge to search all image data in all file types within an organization for what might be personally identifiable information. A first step to this challenge is to use the location and creation time of picture objects to determine the possible links with PII being processed concurrently. Through our workflow analysis we can also use the context of linked files (images and text) but there is still a chance that images of significance may be missed.

Another challenge is presented by links or indirect references to PII. The information pointed to by links (e.g., hyperlinks or database references) may contain very sensitive information, yet without following and analyzing all of those links, their significance may be missed. With respect to metadata, data used to describe other data, there is a challenge in being able to accommodate all of the possible varieties of metadata, including both machine-readable and human-readable metadata types. The metadata problem is complicated because people working alone, or in collaboration, often use a variety of different tools, often in ways not originally envisioned by the developers. This leads to rather complicated metadata for analysis.

Appropriate visualization of any discovered knowledge and representing the relationships between workflows and PII handling is another area where there are interesting challenges. It is not sufficient (or even possible, in some cases) to display the multitude of instances of PII found within an organization. Analyzing workflow and comparing it against prescribed workflow or authorizations to highlight problem areas and risks is one approach we are exploring. Other approaches include display and navigation through multidimensional data spaces.

Other research underway in our team includes developing techniques to recommend corrections in privacy and security policies or workflow models based upon the measured workflow patterns. We are also exploring improved techniques for extracting workflows (represented as colored Petri nets), security techniques to protect

SNAP agents, their datasets, and communications, and instilling trust in the system operators and employees within an enterprise [11].

A clear social challenge is the fact that our technology has the potential to monitor and analyze many different user activities and behaviors as people work individually or in collaboration. We have attempted to address this by layering the functionalities within our system so it only tracks events that relate to PII processing in the context of particular operations. It is our intention to work with potential end users from different industries of this type of automated privacy management technology (e.g., the banking industry) in order to set functionality targets and research goals that will address the real challenges for managing private data in those domains.

6 Acknowledgements

The authors acknowledge the programming support of Luc Belliveau and the development contributions of Arlen Gallant and Rougu Lou.

References

1. Korba, L., Song, R., Yee, G., Patrick, A. Automated social network analysis for collaborative work, Proceedings of the Third International Conference on Cooperative Design, Visualization and Engineering (CDVE 2006). Palma de Mallorca, Spain. September 17-20, 2006.
2. Luhn's Algorithm on Wikipedia, last accessed: March 20, 2007, http://en.wikipedia.org/wiki/Luhn_algorithm
3. Java programming language. Available at: http://java.sun.com/ March, 2007.
4. Jade Platform available at: http://sharon.cselt.it/projects/jade/ March, 2007.
5. Ashley, P., Powers, C., Schunter, M. From privacy promises to privacy management: a new approach for enforcing privacy throughout an enterprise, Proc. of the 2002 New Security Paradigms Workshop, Virginia Beach, Virgina, pp. 43-50.
6. SPARCKLE (Server Privacy ARchitecture and CapabiLity Enablement) policy Workbench, IBM Watson Labs available at: http://domino.watson.ibm.com/comm/research.nsf/pages/r.security.innovation2.html, March, 2007.
7. Heck, M. Guard your data against insider threats, Oakley, Reconnex, Tablus and Vontu prevent costly data leaks, available at: http://www.infoworld.com/article/06/01/13/73680_03TCdataleak_1.html March, 2007.
8. Bhattacharya, J., Dass, R. Kapoor, Vishal, Chakraborti, D., Gupta, S.K. PRIVDAM: privacy violation detection and monitoring using data mining, available at: http://ideas.repec.org/p/iim/iimawp/2005-07-01.html March, 2007.
9. Carminati, B., Ferrari, E. Trusted privacy manager: a system for privacy enforcement of outsourced data, Proc. of the 21st workshop on Data Engineering, April 5-8, 2005, pp. 1195-1203.
10. Aura, T., Kuhn, T.A., Roe, M. Scanning Electronic Documents for Personally Identifiable Information, Proc. of the 5th ACM Workshop on Privacy in Electronic Society, Alexandria, Virgina, 2006, pp. 41-50.
11. Patrick, A.S., Briggs, P., Marsh, S. (2005). Designing systems that people will trust. In L. Cranor & S. Garfinkel (Eds.), Security and Usability: Designing Secure Systems That People Can Use, O'Reilly & Associates.