Automated Understanding Of Cloud Terms Of Service And SLAs

Nick Papanikolaou[1], Siani Pearson[1], Marco Casassa Mont[1]
Cloud and Security Lab
Hewlett-Packard Laboratories
[1]Bristol, UK / [2]Fusionopolis, Singapore
{nick.papanikolaou,siani.pearson,marco_casassa-mont}@hp.com

***Abstract*—We argue in favour of a set of particular tools and approaches to help achieve accountability in cloud computing. Our concern is helping cloud providers achieve their security goals and meeting their customers' security and privacy requirements. The techniques we propose in particular include: natural-language analysis (of legislative and regulatory texts, and corporate security rulebooks) and extraction of enforceable rules, use of sticky policies, automated policy enforcement and active monitoring of data, particularly in cloud environments. This is a position paper reporting our initial thinking and current progress.**

Keywords-cloud computing, accountability, natural language processing, policy enforcement

I. INTRODUCTION

For cloud services to be adopted on a wide scale by businesses and individuals, it is necessary for vendors to provide adequate security and privacy controls for the data stored in their systems. In order to ensure compliance with applicable law and standards, and adherence to particular customer requirements (e.g. "Certain types of data should not be stored beyond the national boundaries of Canada or in a public cloud"), vendors need to constantly monitor access and use of their infrastructure and protect against an increasing number of threats. The challenge of accountability is a central concern for vendors, and meeting this challenge means being able to trace the location, flows, instances and accesses of the data stored in their infrastructure.

There is currently no widely accepted methodology or toolset for technically achieving accountability in cloud computing, with potential solutions being heavily dependent on the particular platform and virtualization technology used by a vendor. What is clear is that a variety of mechanisms need to be put into place to protect against data leakage and to enforce legislation and other related restrictions on the storage and transfer of data, especially across national borders.

Our objective is to identify automated means for cloud service providers to provide accountability with regards to their data governance practices. In the context of this paper accountability is understood as the goal of preventing harm to a cloud provider's customers by enforcing adequate protections on these customers' data, and having available effective reporting and auditing mechanisms.

While accountability in the broadest sense can be guaranteed only through a combination of law, regulation and technical enforcement mechanisms (e.g. in the context of privacy [23], such mechanisms are Privacy Enhancing Technologies), our focus is on the technical aspects. As stated in the introduction, what is practically required for a cloud provider to be accountable is, among other things, a set of tools to track the location, flows, and accesses of its customers' data. As we shall see, this capability allows a provider to demonstrate compliance to the law and adherence to all relevant regulations and other restrictions. More importantly, this capability allows any instances of non-compliance to be detected effectively, so that suitable corrective action can be taken.

Of course, the capability to provide and demonstrate compliance needs to be founded on privacy law and secured based on best practices and industry standards. Any platform to provide accountability needs to be secured so that it cannot be exploited by attackers.

II. METHODOLOGY

We argue that it is possible to automate processes required to ensure that a provider is accountable, although we recognise the difficulty of mapping and linking legal and regulatory requirements - which are high-level and expressed in natural language - to technically enforceable policies on particular data items.

Key techniques that can be used to achieve a significant degree of automation include:

- **natural-language analysis**, in particular, extraction of policy rules from legislative and regulatory texts and corporate rulebooks; these rules should be represented in a form that can be interpreted by a technical enforcement mechanism (esp. a Policy Enforcement Point or PEP), but possibly also so that they can be incorporated into a compliance checker of information governance software (cf. Governance/Risk Management/Compliance (GRC) Platforms, widely used in industry). It is usually hard for humans to capture these policies (or the relevant part of them) and translate them into manageable/enforceable constraints. This mapping is highly error prone. Natural-language analysis aims at supporting the creation of

machine enforceable rules. It should be noted here that no natural-language processing system can operate with 100% accuracy, but use of such systems can help to reduce significantly the overall amount of human intervention in the process of policy creation and management.

- **use of sticky policies**: by strongly binding policies to the data they are associated with, it is easier for providers to control accesses to data within their cloud infrastructure and there is no need for a central policy repository. From the point of view of automating accountability, the use of sticky policies is a very useful technique. Sticky policies provide a means of data encryption, since the data which a policy is bound to cannot be accessed unless that policy is complied with. See references [30,33] for more details.
- **automated policy enforcement:** the deployment of control points throughout a cloud provider's infrastructure where policy rules can automatically be enforced, and human users only notified in case of failure or error is essential. We refer to the following current and future HP Labs European and TSB research projects for more related work on policy enforcement: EnCoRe [1] , Information Stewardship in the Cloud [35], and TrustDomains[2].
- **active monitoring for compliance:** we believe that it is fundamental for cloud providers to have in their infrastructure mechanisms for automatically detecting compliance problems and potential sources of such problems. It is possible to formulate and regularly check system invariants corresponding to conditions that should never occur at certain end points, such as links between a provider's data centres, and particularly cross-border links. An example of such monitoring mechanism can be found in the TrustCloud project [23].

In the following sections, we survey related work on extraction of rules from privacy texts; first, we will report some progress on analysing cloud service-level agreements and associated privacy rules using natural-language processing tools.

III. Platforms and tools for Natural-language analysis of Cloud SLAs and Policy rules

A number of libraries and tools exist that facilitate the parsing and extraction of information from natural-language texts, provided one is willing to design a suitable grammar for the sentences of interest; our recent work has involved identifying common sentence structures in cloud service level agreements and also in privacy rulebooks used within enterprises. The essential characteristic of such documents is the prescriptive style in which they are written. For instance, one can extract simple rules from sentences that contain verbs issuing commands (*cf.* the sample sentence "Personal identifiers *must be* anonymised prior to export."). For these tasks we have been investigating the use of the Natural Language Toolkit (NLTK) [5], the use of definite-clause grammars in Prolog, and the GATE natural-language processing system [17]. The choice of these tools and techniques is justified by the amount of previous work that has been done for similar applications in security and privacy rule extraction in other fields, as surveyed in the next few sections.

[1] See `http://www.encore-project.info`.

[2] See `http://www.cs.ox.ac.uk/projects/TDoms/`.

IV. Knowledge Extraction from Texts and Learning

Antón and a number of different collaborators (see [1,7,8,9,21]) have used textual mining techniques to analyze privacy policies and a number of different privacy and privacy-related regulations. For example, in [1] the authors focus on privacy policies from financial institutions which claim to be compliant with the Gramm-Leach-Bliley Act (GLBA). Papers [1] and [7] refer to PGMT, which is a tool for representing and analyzing rules arising in privacy regulations as restricted natural-language statements. In [21] the authors discuss the extraction of structured rules from source texts using an NLP platform called *Cerno.*

In [9] Breaux, Anton and Vail use their approach of *semantic parameterization* to represent the US HIPAA (Health Insurance Portability and Accountability Act) Privacy Rule as a set of restricted natural-language statements, classified as rights, constraints or obligations. They identify standard phrases appearing in the legislative document, and note the frequency of their occurrence and the corresponding modality (right/obligation/interdiction etc.). They also discuss how to handle ambiguities. This work is extended further in [8], where the authors develop a detailed classification of constraints and introduce means of handling complex cross-references arising in the legal text of the HIPAA.

Delannoy et al. [19] combine a template-matching technique with machine learning in order to match rules from the Canadian 1991 tax guide with text describing case studies of particular individuals; this approach in principle allows one to see which tax rules apply in a given situation. The paper describes an architecture and tool called MaLTe, which is capable of learning how to apply rules to different input texts.
Delisle et al. [20] describe in detail a framework for extracting meaning from the structure of technical documents. Their approach is relevant to the analysis of prescriptive texts in that they assume that input documents are highly structured and somewhat predictable. The authors propose a number of techniques for identifying patterns in texts and converting sentences to Horn clauses. The Horn clauses represent knowledge about the domain in question; through the use of machine learning techniques, this knowledge is extended and refined as more documents are supplied.

Stamey and Rossi [36] use singular-value decomposition and latent semantic analysis techniques to analyze privacy policy texts. They identify commonly occurring topics and key terms and their relations. They are also able to detect similar word meanings; the strength of their approach is that they are able to pick out ambiguities in privacy policies and make them visible to the user. The tool *Hermes* developed by the authors allows automated analysis of an entire privacy policy text, outputting an overall ranking of the policy (when compared to a reference text).

We are also aware of much work on knowledge extraction from legislation [4,6,11,14,18]. Due to space limitations we will not expand on this further.

V. Semantic Models and Representations

Waterman [24] develops a simple table-based representation of particular laws. This author demonstrates a so-called 'intermediate isomorphic representation' of a rule from the US Privacy Act, and similarly for a rule from the Massachusetts Criminal Offender Records Law. The key idea here is to use a structured representation that can be mapped directly back to the original legal text and to corresponding computer code. The representation still uses natural language, but with additional logical structure. The additional structure helps to separate out actors, verbs, context and particular constraints that exist in the legal text (and which are often implied or included indirectly with the use of cross-references).

The framework proposed by Barth, Datta, Mitchell and Nissenbaum [2] comprises a formal model which is used to express and reason about norms of transmission of personal information. This work does not involve automatically analyzing text, but does provide a formalism for manually representing notions of privacy found in legislation – particularly in the texts of HIPAA, COPPA and GLBA. The formal model provides notations for defining sets of agents communicating via messages, with particular roles, in specified contexts; linear temporal logic, with a past operator, allows one to express properties that the agent behaviours should satisfy. Policy compliance is formally defined in terms of this model. Although the authors assume that their formalism is for a human user, we envisage the possibility that using natural-language analysis it should be possible to extract from texts some privacy rules expressed in this formalism.

May, Gunter and Lee [25] define a semantic model for expressing privacy properties, and apply it to the HIPAA Privacy Rule; it is based on a classical access control model used in operating system design. The authors translate the legal text into a structured format that uses the commands in the proposed access control model to express rules. The paper does not restrict itself to representation; once the legal rules have been formally expressed, the authors use a model checker to automatically reason about the consistency of the generated rules; they demonstrate subtle differences between the year 2000 and year 2003 versions of the HIPAA Privacy Rule.

It is clear that a uniform, consistent, formal representation of privacy knowledge and privacy rules in particular is useful for automated reasoning about privacy issues. We are keen to make use of existing formal representations of privacy rules when performing natural-language analysis of privacy-related texts, since the usefulness of such representations has already been demonstrated for complex texts, particularly American privacy legislation.

VI. Policy Enforcement and Compliance

To our best knowledge, we are not aware of any previous work that addresses the whole lifecycle of natural-language analysis of privacy texts with the goal of enforcing suitable rules, e.g. in an enterprise setting (although the EU CONSEQUENCE project does take an holistic approach it does not involve natural-language analysis). As stated in the Introduction, achieving compliance with privacy legislation and regulations is a central concern in enterprises, and means of automating compliance are highly desirable. Since new privacy rules are almost exclusively expressed using natural-language, means of automatically analyzing the appropriate texts and extracting rules from them necessary – the resulting rules can then be incorporated into existing enterprise rule-bases, such as those used in GRC platforms. We mention here some work on automated policy enforcement and compliance, which has so far been developed separately and independently of any consideration of automated knowledge and rule extraction.

The EnCoRe research project [37] is developing a platform for expressing and enforcing privacy preferences for personal data; recent case studies include a system for managing data held within an enterprise's HR systems, and health data stored about individuals and tissue samples in a biobank. Through the use of a suitable policy enforcement architecture, legal and regulatory privacy rules, along with individuals' privacy preferences, can be automatically enforced so that unauthorized and/or unsuitable access to data is prevented. In [15] we proposed a simple conceptual model for representing privacy rules, which can be directly mapped to technically enforceable access control policies (expressed e.g. using XACML).

In [31] Pearson et al. propose a tool for providing decision support with regards to privacy-sensitive projects that arise in an enterprise. Decision support systems are built on knowledge bases with rich sets of rules, and the process of translating legal texts, regulations and corporate guidelines into technically enforceable rules is complex and laborious. For this reason a conceptual model is a useful aid.

We believe there is scope for integration of several of the different approaches described so far into a natural-language processing pipeline, which can be integrated with technical enforcement mechanisms to achieve compliance for privacy: this starts with the initial task of analyzing natural-language privacy texts, to the extraction of formalized rules and their automatic enforcement. We are working on developing tools for automating privacy in cloud computing and, for this, natural-language analysis of provider SLAs, international laws and regulations will need to be combined with suitable enforcement methods such as distributed access control [3], sticky policies and policy-based obfuscation [27].

VII. Conclusions

We believe that it is beneficial and possible for cloud service providers to automate a number of tasks related to the requirement of accountability. We have identified some specific techniques, namely: natural-language analysis of law, regulation and corporate guidelines on security and privacy of customer data in order to generate technically enforceable policies; use of sticky policies to achieve a strong binding between data and the stipulations that apply to the use and dissemination of that data; and active monitoring of a cloud provider's infrastructure to detect potential compliance problems. More in-depth analyses of ways to achieve accountability in the cloud are available in some of our previous work (see also [22, 23, 29, 30, 32-33]).

We are actively working on the development of all these techniques which, combined with the deployment of technical policy enforcement mechanisms in a cloud provider's infrastructure, can help achieve accountability, which is a major concern in cloud computing today.

REFERENCES

[1] A. Antón, J. B. Earp, Q. He, W. Stufflebeam, D. Bolchini, C. Jensen. Financial Privacy Policies and the Need for Standardization. *IEEE Security and Privacy* **2**(2), pp. 36--45 (2004)

[2] A. Barth, A. Datta, J. C. Mitchell, H. Nissenbaum. Privacy and Contextual Integrity: Framework and Applications. In *Proceedings of IEEE Symposium on Security and Privacy* (2006)

[3] M. Becker, P. Sewell. Cassandra: Distributed Access Control Policies with Tunable Expressiveness. In Proceedings of 5th IEEE International Workshop on Policies for Distributed Systems and Networks (POLICY 2004), 7-9 June 2004, Yorktown Heights, NY, USA. 159--168, IEEE Computer Society (2004)

[4] V. Benjamins, P. Casanovas, J. Breuker, A. Gangemi (eds.). *Law and the Semantic Web*. Springer (2005)

[5] S. Bird, E. Loper. "NLTK: The Natural Language Toolkit". Proceedings of the ACL demonstration session. pp 214-217, Barcelona, Association for Computational Linguistics, July 2004.

[6] D. Bourcier. Legal Knowledge and Information Systems. IOS Press (2003)

[7] T. Breaux and A. I. Antón. Deriving Semantic Models from Privacy Policies. In Proceedings of the Sixth International Workshop on Policies for Distributed Systems and Networks (POLICY'05) (2005)

[8] T. Breaux, A. I. Antón. Analyzing Regulatory Rules for Privacy and Security Requirements. *IEEE Transactions on Software Engineering* **34**(1), 5--20 (2008)

[9] T. Breaux, M. W. Vail and A. I. Antón. Towards Regulatory Compliance: Extracting Rights and Obligations to Align Requirements with Regulations. In *Proceedings of 14th IEEE International Requirements Engineering Conference (RE'06)* (2006).

[10] J. Bret Michael, V. Ong, and N. C. Rowe. Natural-Language Processing Support for Developing Policy-Governed Software Systems. In *Proceedings of 39th International Conference and Exhibition on Technology of Object-Oriented Languages and Systems (TOOLS 39)*, 263--274 (2001)

[11] J. Breuker, P. Casanovas, M. C. A. Klein, E. Francesconi (eds.). *Law, Ontologies and the Semantic Web*. IOS Press (2009)

[12] A. Brodie, C. Karat, J. Karat. An Empirical Study of Natural Language Parsing of Privacy Policy Rules Using the SPARCLE Policy Workbench. In *Proceedings of Symposium on Usable Privacy and Security (SOUPS)* (2006)

[13] L. Bussard, M. Y. Becker. Can Access Control be Extended to Deal with Data Handling in Privacy Scenarios? In *Proceedings of W3C Workshop on Access Control Application Scenarios* (2009).

[14] P. Casanovas, G. Sartor, N. Casellas, R. Rubino (eds.). *Computable Models of the Law*. Springer (2008)

[15] M. Casassa Mont, S. Pearson, S. Creese, M. Goldsmith, N. Papanikolaou. A Conceptual Model for Privacy Policies with Consent and Revocation Requirements. In *Proceedings of PrimeLife/IFIP Summer School 2010: Privacy and Identity Management for Life*, Lecture Notes in Computer Science, Springer (2010)

[16] K. Chen, D. Wang. An aspect-oriented approach to privacy-aware access control. In Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007 (2007)

[17] H. Cunningham, et al. Text Processing with GATE (Version 6). University of Sheffield Department of Computer Science. 15 April 2011. ISBN 0956599311.

[18] J. Davies, M. Grobelnik, D. Mladenic (eds.). *Semantic Knowledge Management.* Springer (2009)

[19] J. Delannoy, C. Feng, S. Matwin, and S. Szpakowicz. Knowledge Extraction from Text: Machine Learning for Text-to-rule Translation. In *Proceedings of Machine Learning and Text Analysis Workshop (ECML-93)* (1993)

[20] S. Delisle, K. Barker, J. Delannoy, S. Matwin, S. Szpakowicz. From Text to Horn Clauses: Combining Linguistic Analysis and Machine Learning. In Proceedings of Canadian AI Conference (AI/GI/CV '94 (1994)

[21] N. Kiyavitskaya, N. Zeni, T. D. Breaux, A. I. Antón, J. R. Cordy, L. Mich, J. Mylopoulos. Extracting Rights and Obligations from Regulations: Toward a Tool-Supported Process. In *Proceedings of ASE'07* (2007)

[22] R. Ko, B. Lee and S. Pearson, "Towards achieving accountability, auditability and trust in cloud computing", A. Abraham et al. (Eds.), ACC 2011, Part IV, CCIS 193, pp. 432–444, Springer-Verlag Berlin Heidelberg, 2011.

[23] R. Ko, P. Jagadpramana, M. Mowbray, S. Pearson, M. Kirchberg, Q. Liang, B. Lee, "TrustCloud: A Framework for Accountability and Trust in Cloud Computing", 2nd IEEE Cloud Forum for Practitioners (ICFP), IEEE Computer Society, Washington DC, USA, 7-8 July 2011.

[24] K. Krasnow Waterman. Pre-processing Legal Text: Policy Parsing and Isomorphic Intermediate Representation. In *Proceedings of PRIVACY 2010 - Intelligent Information Privacy Management AAAI Spring Symposium*, Stanford Center for Computers and Law, Palo Alto, California, USA (2010)

[25] M. May, C. A. Gunter, I. Lee. Privacy APIs: Access Control Techniques to Analyze and Verify Legal Privacy Policies. In *Proceedings of Computer Security Foundations Workshop (CSFW'06)* (2006)

[26] B. Moulin and D. Rousseau. Automated Knowledge Acquisition from Regulatory Texts. *IEEE Expert* **7**(5), 27--35 (2002)

[27] M. Mowbray, S. Pearson and Y. Shen. Enhancing privacy in cloud computing via policy-based obfuscation. *Journal of Supercomputing*. DOI: 10.1007/s11227-010-0425-z.

[28] V. Ong. An Architecture and Prototype System for Automatically Processing Natural-Language Statements of Policy. Master's thesis, Naval Postgraduate School, Monterey, California (2001)

[29] N. Papanikolaou, S. Pearson and M. Casassa Mont, "Towards Natural-Language Understanding and Automated Enforcement of Privacy Rules and Regulations in the Cloud: Survey and Bibliography", Secure and Trust Computing, Data Management and Applications, Communications in Computer and Information Science, vol. 187, Springer Berlin Heidelburg, pp. 166-173, 2011.

[30] S. Pearson and M. Casassa Mont, "Sticky Policies: An Approach for Privacy Management across Multiple Parties", Special Issue on Security and Privacy in an Online World, IEEE Computer, pp. 693-702, July 2011.

[31] S. Pearson, P. Rao, T. Sander, A. Parry, A. Paull, S. Patruni, V. Dandamudi-Ratnakar, P. Sharma. Scalable, accountable privacy management for large organizations. In *Proceedings of 13th Enterprise Distributed Object Computing Conference Workshop (EDOCW 2009)*, 168--175 (2009)

[32] S. Pearson, "Toward Accountability in the Cloud", View from the Cloud, IEEE Internet Computing, IEEE Computer Society, July/August issue, vol. 15, no. 4, 2011.

[33] S. Pearson, M. Casassa Mont and G. Kounga, "Enhancing Accountability in the Cloud via Sticky Policies", Secure and Trust Computing, Data Management and Applications, Communications in Computer and Information Science, vol. 187, Springer Berlin Heidelburg, pp. 146-155, 2011.

[34] M. Peleg, D. Beimel, D. Dori, Y. Denekamp. Situation-Based Access Control: privacy management via modeling of patient data access scenarios. *Journal of Biomedical Informatics* (to appear).

[35] D. Pym, M. Sadler. Information Stewardship in Cloud Computing. In Proceedings of IJSSMET. 2010, 50-67.

[36] J. Stamey, R. A. Rossi. Automatically Identifying Relations in Privacy Policies. In *Proceedings of SIGDOC'09* (2009)

[37] C. Vanden Berghe, M. Schunter. Privacy Injector - Automated Privacy Enforcement through Aspects. In *Proceedings of 6th Workshop on Privacy Enhancing Technologies, 28-30 June 2006* (2006)