Data Flow-Oriented Process Mining to Support Security Audits

Thomas Stocker
Supervised by: Rafael Accorsi

University of Freiburg, Germany
stocker@iig.uni-freiburg.de

Abstract. The automated execution of dynamically-evolving business processes in service-oriented architectures requires audit methods to assert that they fulfill required security properties. Process mining techniques can provide models for the actual process behavior, but mostly disregard the dynamics of processes running in highly flexible environments and neglect the data flow perspective. This research plan is on novel data-oriented mining techniques to tackle these shortcomings in order to support effective security audits.

1 Problem Statement and Context

Business process management (BPM) allows the design, enactment, management and analysis of operational business processes [19]. In such "process-aware information systems", workflow adoption is expected to soar with the advent of cloud and service-oriented computing, providing a basis for structuring and integrating business processes [8]. Valuable synergy effects between BPM and SOA gave reasons for a new field of service-oriented process modeling tools and languages such as WS-BPEL. The deployment of configurable workflows "as-a-service" [1] that may evolve along time [20] allows enterprises to work more efficient and react more flexible on changing requirements such as customer demands or changes in the technological, business or legal context.

Although workflows are largely employed in mission-critical activities demanding strong security and privacy guarantees [7], there is today a lack of audit methods to assert that they fulfill the required properties [14]. In particular, computer assisted auditing techniques (CAAT) are missing that cope with the *security analysis* of *evolving* workflows [17]. As a consequence, a significant number of exploited process vulnerabilities and data leaks [12] goes undetected. Especially in the highly dynamic field of service-oriented computing the verifiability of security guarantees plays a key role in technology-adaption and trust-establishment. Besides direct illegal data flows, information leakage includes the use of covert channels that allow data usage and flow not intended by the system, i.e. by monitoring (temporally) system behavior [13]. Information flow analysis captures both types of data flow.

This research plan is on the development of advanced process mining techniques that allow detailed views on process enactments including data flows,

G. Pallis et al. (Eds.): ICSOC 2011, LNCS 7221, pp. 171–176, 2012.

thereby providing a basis for efficient security analysis with regard to information flow policies. Process mining is a method of distilling a structured process from a set of real executions [18]. The usual target meta-model for reconstruction is (some dialect of) Petri net, which provides an expressive formalism to capture the control flow and data flow of business processes [6]. Traditionally, workflow mining extracts a single, representative model that consolidates all the different executions happening in the log file. Recently, trace clustering techniques have been proposed as a preprocessing step for workflow mining [10,15,11]. The idea is to group traces according to different characteristics and, subsequently, apply workflow mining to a particular set of clusters.

The considered problem statement is twofold and follows directly from the identified shortcomings exhibited by current trace clustering and process mining approaches:

1. **Recovered processes neglect the history of changes.**
   Traces are mostly grouped according to their structural similarity, thereby neglecting the time aspect. In doing so, auditors cannot test for particular timeframes. Although trace clustering allows for the selective reconstruction of traces, it fails to mine the complete "history" (i.e. *evolution provenance*) of a business processes, identifying their diverse "tenancies" and how they differ one from the other.

2. **Process mining does not consider data flows.**
   Target meta-models for reconstruction typically consider the control flow of processes (i.e. the relation and sequence of process activities) and do not include data flows. As a result, reconstructed process models are not appropriate to serve as input for an information flow analysis. Security analysis is limited to control flow properties and cannot check data flow oriented properties.

2 Proposed Solution and Research Challenges

A first step on the way from process logs to meaningful security audits is the **development of time-based clustering algorithms** that are able to view the history of process structure changes during time. Reliable indicators for structural changes of processes have to be identified on the basis of process logs. The challenge here is that typically not every possible execution path of a workflow occurs with the same probability. In case of deviations from"typical" workflow behavior it is hard if not impossible to distinguish rare execution variants (traces) from traces related to a changed version of the workflow.

The essential part of this research plan is to **develop novel process mining techniques that incorporate data flow** aspects provided by process logs. This requires an analysis on which information can be extracted and which kind of meta-model is appropriate to describe data flows. In case of InDico [3] the meta-model is explicitly given and can be constructed by annotating reconstructed Petri

nets with data flow properties. Here the question is, if existing mining approaches can be extended to provide this functionality or if a complete redesign is required. On the other hand there may be other meta-models allowing different kinds of analysis. Which types of additional analysis techniques can be put on top of the proposed solution is another question picked up by this research plan.

Relying on process mining, the most relevant problem of this approach of finding a compromise between under- and overfitting a process log applies also for this research. A model is called to underfit a process log, if it allows too much behavior which is not reflected by the log. Algorithms showing this behavior have a tendency to over-generalize. Overfitting a log means to stay at a too specific level without any generalization. Such models typically establish sequential paths for each distinct trace and are merely another representation of the log.

2.1 Expected Impact

By introducing a novel trace clustering method, this research contributes to the fields *business provenance* that captures the lineage of business artifacts, including workflows [9]. Here, approaches to capture workflow evolution are missing. The proposed clustering technique *and* the subsequent mining should close this gap. It also contributes to *audit reduction* because the volume of audit records to facilitate manual review can be reduced. Auditors are able to select which among the various mined workflow specifications are to be tested for compliance with security policies.

Advancing process mining techniques to provide additional information about data flows forms a basis for applying manifold analysis techniques. Thus, this research plan also contributes to *security audits and CAAT*.

As an example the proposed techniques provide sufficient information for effective information flow analysis. Based upon InDico [2], the mined workflow specifications can be automatically tested for a multitude of security properties, including MAC-based, non-interference, and enterprise relevant properties, such as Chinese-wall and separation of duties. InDico aims to provide a well-founded, uniform approach and corresponding tool-support for the automated analysis of existing and/or mined workflow specifications for security properties. It defines a colored Petri net dialect, called IFnet, to model business process models and to formalize multi-level information flow properties [3].

Another possibility is to introduce a new form of analysis called contextualization. Combining the results of an information flow analysis showing covert channels, identified data flows can give evidence on their usage which enables much deeper checking of security requirements regarding information flow.

Allowing for a broader range of analysis techniques based on reconstructed process models, this research plan is an important contribution for the enhancement and automation of security audits in the BPM sector.

2.2 Preliminary Results

In order to find appropriate meta-models for efficient data- and information flow analysis, the starting point of this research was to analyze propagation graphs

capturing data flows within a process execution with extensional data flow properties, that denote what - instead of how - relevant industrial requirements are to be achieved [5]. Providing a sufficient basis for data flow analysis, propagation graphs do not reflect the control flow of process executions. While this first result was important to learn about the requirements of forensic analysis of business processes, it showed that alternative meta-models are needed to facilitate information flow analysis requiring both data flow and control flow.

A preliminary solution for the time-based clustering approach was already developed and presented at the *Workflow Security Audit and Certification* workshop at the BPM'11 conference [16]. It uses distances of process activities in the sense of intermediate activities to determine cluster borders. As long as the structure of a process does not change these distances move in fixed intervals. Changing operations cause boundary variations for at least one activity pair. Sequentially processing traces according to their timestamp, the algorithm uses samples to determine the typical behavior in terms of boundaries for the minimum and maximum observed value of activity-pair distances. Typical workflow behavior is determined on the basis of a parameter w (windowsize), that specifies the minimum number of traces used as "training" data and also defines the minimum cluster size. Using this clustering technique allows auditors to reconstruct the process history, so that they can appreciate the different ways in which a process evolves over time.

The appropriateness of *conformance checking* (process mining techniques providing proofs for the adherence of logged process traces to specific execution constraints) for security in the sense of CAATs was evaluated within a case study that will be presented at the Enterprise Engineering track of the ACM SAC conference in 2012.

3 Research Plan

The envisioned research is scheduled for a period of three years and divided into two packages picking up the aspects of section 2. Each package includes a phase where the applicability of the developed approaches is evaluated. Depending on the information about practical business process management and auditing that can be gained from expert interviews and cooperations with industry partners, evaluation will be either based on real process data (process models and logs) or on synthetic data having realistic characteristics.

1. Time-Oriented Clustering
   - Discovering suitable differentiation criteria within log files that support the detection of process changes
   - Development of new clustering algorithms that show process evolutions
   - Integration of developed algorithms in the security workflow analysis toolkit (SWAT) [4] developed at the University of Freiburg.
   - Conducting case-studies to evaluate approaches

2. Data Flow-Oriented Process Mining
   - Identification of data flow properties within log files
   - Reconstruction of IFnet-models
   - Integration of developed algorithms in SWAT.
   - Conducting case-studies to evaluate approaches

References

1. Accorsi, R.: Business process as a service: Chances for remote auditing. In: IEEE Computer Software and Applications Conference (2011)
2. Accorsi, R., Wonnemann, C.: Strong non-leak guarantees for workflow models. In: ACM Symposium on Applied Computing, pp. 308–314. ACM (2011)
3. Accorsi, R., Wonnemann, C.: InDico: Information Flow Analysis of Business Processes for Confidentiality Requirements. In: Cuellar, J., Lopez, J., Barthe, G., Pretschner, A. (eds.) STM 2010. LNCS, vol. 6710, pp. 194–209. Springer, Heidelberg (2011)
4. Accorsi, R., Wonnemann, C., Dochow, S.: SWAT: A security analysis toolkit for reliably process-aware information systems. In: Workshop on Security Aspects of Process-aware Information. IEEE
5. Accorsi, R., Wonnemann, C., Stocker, T.: Towards forensic data flow analysis of business process logs. In: Proceedings the IEEE Conference on Incident Management and Forensics. IEEE Computer Society (2011)
6. Adam, N., Atluri, V., Huang, W.: Modeling and analysis of workflows using petri nets. Intelligent Information Systems 10(2), 131–158 (1998)
7. Atluri, V., Warner, J.: Security for workflow systems. In: Handbook of Database Security, pp. 213–230 (2008)
8. Cummins, F.: BPM meets SOA. In: Handbook on Business Process Management 1. International Handbooks on Information Systems, pp. 461–479 (2010)
9. Curbera, F., Doganata, Y., Martens, A., Mukhi, N.K., Slominski, A.: Business Provenance – A Technology to Increase Traceability of End-to-End Operations. In: Meersman, R., Tari, Z. (eds.) OTM 2008, Part I. LNCS, vol. 5331, pp. 100–119. Springer, Heidelberg (2008)
10. de Medeiros, A.K.A., Guzzo, A., Greco, G., van der Aalst, W.M.P., Weijters, A.J.M.M., van Dongen, B.F., Saccà, D.: Process Mining Based on Clustering: A Quest for Precision. In: ter Hofstede, A.H.M., Benatallah, B., Paik, H.-Y. (eds.) BPM Workshops 2007. LNCS, vol. 4928, pp. 17–29. Springer, Heidelberg (2008)
11. Greco, G., Guzzo, A., Pontieri, L., Saccà, D.: Discovering expressive process models by clustering log traces. IEEE Transactions on Knowledge and Data Engineering 18(8), 1010–1027 (2006)
12. Lowis, L., Accorsi, R.: Finding vulnerabilities in SOA-based business processes. IEEE Transactions on Service Computing (2011) (to appear)
13. McHugh, J.: Handbook for the Computer Security Certification of Trusted Systems. Naval Research Laboratory (1995)
14. Sayana, A.: Using CAATs to support IS audit. Information Systems Control Journal, 1 (2003)
15. Song, M., Günther, C.W., van der Aalst, W.M.P.: Trace Clustering in Process Mining. In: Ardagna, D., Mecella, M., Yang, J. (eds.) BPM 2008 Workshops. LNBIP, vol. 17, pp. 109–120. Springer, Heidelberg (2009)

16. Stocker, T.: Time-Based Trace Clustering for Evolution-Aware Security Audits. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) BPM Workshops 2011, Part II. LNBIP, vol. 100, pp. 471–476. Springer, Heidelberg (2012)
17. Teeter, R., an Miklos Vasarhelyi, M.: Remote auditing: A research framework. Journal of Emerging Technology in Accounting (to appear)
18. van der Aalst, W., Weijters, T., Maruster, L.: Workflow mining: discovering process models from event logs. IEEE Transactions on Knowledge and Data Engineering 16(9), 1128–1142 (2004)
19. van der Aalst, W.M.P., ter Hofstede, A.H.M., Weske, M.: Business Process Management: A Survey. In: van der Aalst, W.M.P., ter Hofstede, A.H.M., Weske, M. (eds.) BPM 2003. LNCS, vol. 2678, pp. 1–12. Springer, Heidelberg (2003)
20. Wei, Y., Blake, M.: Service-oriented computing and cloud computing: Challenges and opportunities. IEEE Internet Computing 14, 72–75 (2010)