Semantic Query Federation for Scalable Security Log Analysis

Kabul Kurniawan(✉)

Multimedia and Information System Group, University of Vienna,
Wahringerstrasse 29, 1190 Vienna, Austria
kabulk87@univie.ac.at

Abstract. The digitalization of business processes increasingly exposes organizations to sophisticated cyber-security threats. To contain attacks and minimize their impact, it is essential to detect them early. To this end, it is necessary to analyze a wide range of log files that potentially provide clues about malicious activity. However, these logs are typically voluminous, heterogeneous, difficult to interpret, and stored in disparate locations, which makes it difficult to analyze them. Current approaches to analyze security logs mainly focus on regular expressions and statistical indicators and do not directly provide actionable insight to security analysts. To address these limitations, we propose a distributed approach that enables semantic querying of dispersed log sources in large-scale infrastructures. To automatically integrate and reason about security log information, we will leverage linked data technologies and state-of-the-art federated query processing systems. In this proposal, we discuss the research problem, methodology, approach and evaluation plan for scalable federated semantic security log analysis.

Keywords: Security log analysis · Semantic query federation
Linked data · Semantic reasoning

1 Introduction

Today, most organizations depend strongly upon information technology (IT) systems. While these IT systems provide a lot of benefits, they also have their drawbacks in the form of increasing non-trivial cyber-security threats. These threats manifest as specialized, sophisticated and non-trivial cyber attacks across network and region, which multiply the complexity and difficulty of the effort to keep the system safe. Organizations are now threatened by serious losses from these threats, such as business process disruptions, sensitive data thefts, and reputational damages [1].

IT Systems produce security logs that contain important information within the systems. Security log data can be utilized for supporting security analysis in order to address these threats. By extracting and mining those data, users can reveal events occurred on a certain system [2]. Log data are typically written in various data structures and formats (e.g. plain-text, XML, CSV etc.),

A. Gangemi et al. (Eds.): ESWC 2018 Satellite Events, LNCS 11155, pp. 294–303, 2018.
https://doi.org/10.1007/978-3-319-98192-5_48

coming from heterogeneous data sources (e.g. operating system logs, application logs, database logs, router logs, switch logs, firewall logs, etc.) and result in an enormous size (e.g. gigabyte, terabyte) of log data, which makes it difficult for users to comprehend. In addition, Cyber-attacks launched by adversaries, typically leave digital traces spread across machines in an organization's IT system, makes comprehensive manual security log analysis infeasible [3].

Current approaches in security log analysis mainly focus on system anomaly detection through log parsing and log mining [4]. These approaches typically encounters challenges with incomplete information of logs, which limits the extent of analysis [5]. These challenges emerge due to the difficulty of getting the relevant information from incomplete log sources, as an event occurred in a machine is typically distributed into different log sources in a networked IT systems. Another challenge is on the semantic information from security logs, which is not yet addressed in the current approaches. Without a correct understanding of information contained in logs, it is difficult to infer the causally related events from security logs.

To address these challenges, we propose an innovative approach leveraging linked data technology. Linked data is part of the semantic web technology stack [6], which provides a method of publishing structured data so that it can be interlinked and become more useful through semantic queries [7]. On the security domain, linked data can be applied to provide a conceptual model that can semantically lift heterogeneous security log data, integrate them, and support reasoning to infer implicit causally-related events.

The remainder of this paper is structured as follows: In Sect. 2 we will describe the state of the art and related work. The problem statement and contributions follow in Sect. 3. The methodology and approach to this problem are described in Sect. 4. The preliminary results are shown in Sect. 5 and the evaluation plan that is outlined in Sect. 6. Section 7 concludes the paper.

2 State of the Art

As we plan to leverage linked data technology in our research, in this section we provide the state-of-the-art from two different research domains, cyber-security and semantic web. For the cyber-security domain, we present the state-of-the-art on security log analysis research while on the semantic web research, we provide the state-of-the-art of semantic query federation.

Research on security log analysis has been conducted by several researchers for many years. Some of them have been launched as commercial products (e.g. ArcSight, SplungES, QRadar, LogRhythm, McAfee ESM, etc.). They are mostly implemented based on the Security Information and Event Management (SIEM) approach. SIEM is a combination approach of security information management (SIM) and security event management (SEM) [8]. The goal of this approach is to aggregate relevant information from the extracted log data from multiple different sources in order to provide system administrators access to logging information in a convenient way. SIEM employs statistical correlation engines to generate relationships between event log entries. The main difference between SIEM and our proposed approach is that we provide semantic information and background knowledge of extracted log data and events so that it has capability to infer implicit semantic relationships between those events in which this capability is not provided yet by SIEM systems.

Intrusion detection systems (IDS) are a mechanism to monitor networks or systems for malicious activities. There are two main detection approaches [9] (1) signature-based-detection (2) anomaly-based detection. The signature-based detection is used to detect known attacks by comparing event against a database of signature of malicious activities while the anomaly-based detection is used to detect not only an unknown pattern but also previously unseen pattern using statistical techniques. Machine learning techniques have recently been successfully applied to this context [10]. In our proposed approach, we also aim to identify malicious activities with the main focus on the context-rich high-level understanding of complete attacks.

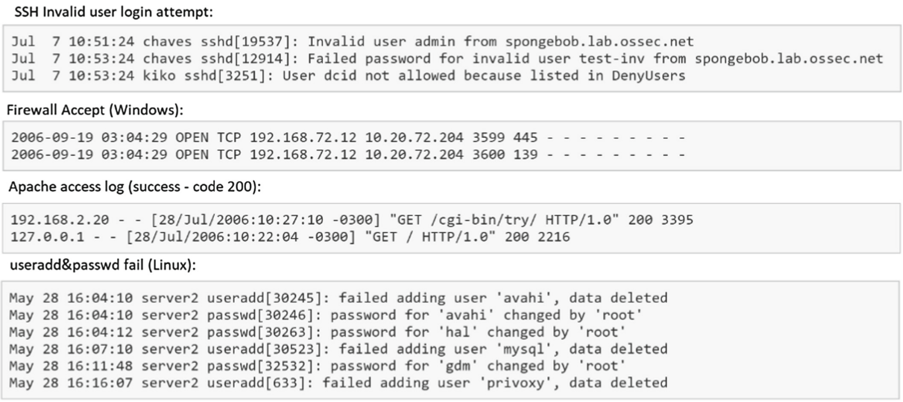
Another conceptual approach to network security assessment has also been recently proposed in [11]. It performs network information acquisition by collecting attributes of network including topology, service, vulnerabilities and configuration. The results of the network information acquisition are used as a basic security ontology to generate attack graphs iteratively and to infer potential attack using a reasoning engine. Compare to our proposed approach, we plan to use empirical research rather than just a conceptual research. We will use security log data as the main data sources and build background knowledge to infer causally-related events.

On the semantic web domain, we particularly focus on the state-of-the-art of semantic query federation approach. Semantic query federation is an approach to query linked data from multiple distributed datasets [19]. There are several existing approaches in terms of semantic query federation. Link Traversal [12] is a semantic query federation approach to discover potentially relevant data during the query execution. It provides high fresh data since the data is directly accessed from a data source. The query execution is initially from one single triple pattern as starting point. FedX [13] is an optimization SPARQL query processing on multiple distributed RDF datasets which are known and accessible via SPARQL Endpoint. SPLENDID [14] provides transparent query federation over distributed SPARQL endpoints. In order to achieve a good query execution performance, data source selection and query optimization are based on basic statistical information which is obtained from VOID descriptions. ANAPSID [15] employs VoiD (Vocabulary of Interlinked Datasets) as data catalogue that is loaded when the system is started and submit ASK SPARQL query to each dataset verification. ANAPSID provides hash join and bind join to merge result locally. Triple pattern fragment (TPF) [16] is a linked data interface which use triple pattern, metadata and controller to query linked dataset. TPF performs with an average speed in running time and top rank in precision-recall (compared with another federated system). TPF offers low-cost and scalable query processing of multiple distributed interlinked datasets.

3 Problem Statement and Contribution

The effort to keep systems safe from cyber-security threats requires comprehensive security analyses to precisely understand malicious events that occurred within them. Logs produced by the systems can be used to support security analyses as they record important system's information. System's logs are composed of log entries in which contain information related to events occurred in a systems or networks [2]. For instance, log messages describe user's activities when they attempt to gain access of a certain system via a networks (e.g. SSH access, FTP access, web portal access etc.). This activity will then be stored as a set of information including date, time-stamp, username, type of access, message etc.

As depicted in Fig. 1, there is a wide variety of structural schemas among log messages. For instance, the 'date' written in the SSH log message is different from the one which is written in the Firewall log (and two others). As illustrated by this example, different systems typically generate different structure of log messages. We can also find other differences between log messages such as the structure order and the attributes' name.



SSH Invalid user login attempt:

```
Jul  7 10:51:24 chaves sshd[19537]: Invalid user admin from spongebob.lab.ossec.net
Jul  7 10:53:24 chaves sshd[12914]: Failed password for invalid user test-inv from spongebob.lab.ossec.net
Jul  7 10:53:24 kiko sshd[3251]: User dcid not allowed because listed in DenyUsers
```

Firewall Accept (Windows):

```
2006-09-19 03:04:29 OPEN TCP 192.168.72.12 10.20.72.204 3599 445 - - - - - - - - -
2006-09-19 03:04:29 OPEN TCP 192.168.72.12 10.20.72.204 3600 139 - - - - - - - - -
```

Apache access log (success - code 200):

```
192.168.2.20 - - [28/Jul/2006:10:27:10 -0300] "GET /cgi-bin/try/ HTTP/1.0" 200 3395
127.0.0.1 - - [28/Jul/2006:10:22:04 -0300] "GET / HTTP/1.0" 200 2216
```

useradd&passwd fail (Linux):

```
May 28 16:04:10 server2 useradd[30245]: failed adding user 'avahi', data deleted
May 28 16:04:10 server2 passwd[30246]: password for 'avahi' changed by 'root'
May 28 16:04:12 server2 passwd[30263]: password for 'hal' changed by 'root'
May 28 16:07:10 server2 useradd[30523]: failed adding user 'mysql', data deleted
May 28 16:11:48 server2 passwd[32532]: password for 'gdm' changed by 'root'
May 28 16:16:07 server2 useradd[633]: failed adding user 'privoxy', data deleted
```

Fig. 1. Security log messages generated from heterogeneous applications and systems

Figure 2 provides an example case about data theft. There is a user who gets access to a certain system on a local computer using either legitimate or illegitimate credentials. This event will be recorded in a pertinent log (e.g. Win Even Log). After the user gets access to a certain server e.g. (file server, web server, database server etc.), the user may download a credential file or dump a database from the database server. These events will then be recorded to a relevant log (e.g. Webserver log, Share-point logs, Sys-logs on endpoint hosts, the file system operation auditing for downloaded files or database backup, DB audit logs, Firewall logs etc.). Subsequently, the user may attach any removable storage (e.g. USB flash-drive) to copy or move the downloaded files to her or his own storage. These events (attach and copy activity) will then be recorded in a system log (e.g. Win Event log). To this end, this example shows that a single case may generate several different logs and it may separately distributed to different systems and locations, depending on the activity of the user.

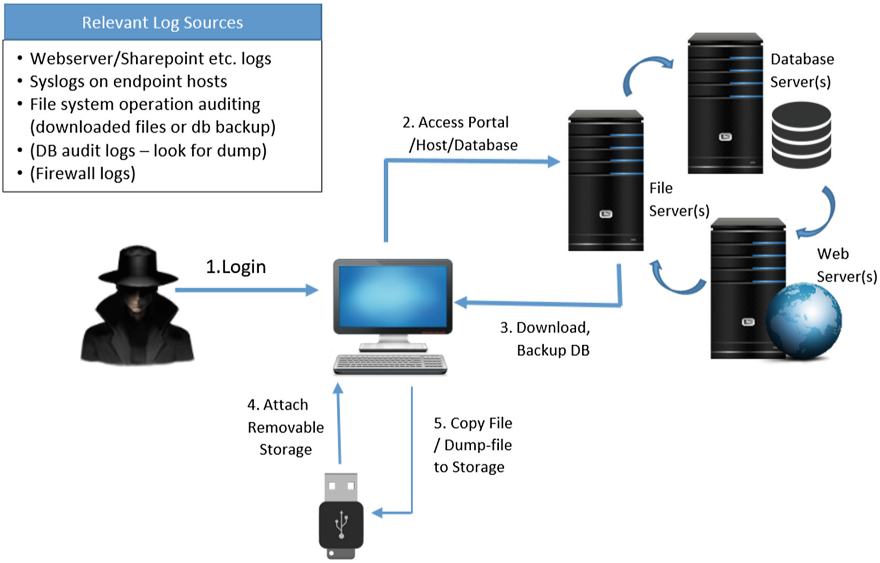

Relevant Log Sources
• Webserver/Sharepoint etc. logs
• Syslogs on endpoint hosts
• File system operation auditing (downloaded files or db backup)
• (DB audit logs – look for dump)
• (Firewall logs)
1.Login
2. Access Portal /Host/Database
3. Download, Backup DB
4. Attach Removable Storage
5. Copy File / Dump-file to Storage
Database Server(s)
File Server(s)
Web Server(s)

Fig. 2. An example of data theft case

Security logs are important sources for security analysts who tackle such security issue. However, log data typically have characteristics which can be problems in security log analysis. These characteristics include but are not limited to (1) heterogeneity (e.g. different terms of terminology, structure, format etc.) (2) separated (coming from different log sources e.g.operating system, application, databases, routers, switches, firewall etc.) (3) enormous sizes (systems may generated vast amount of log data: gigabyte, terabyte etc.) [2].

Thus, in order to address these challenges, we propose an innovative approach of security log analysis by leveraging the potential and the capability of linked data technology. We define a number of research questions based on the consideration of several hypotheses.

RQ1: *How to semantically integrate heterogeneous security log information?*

Although traditional security log analysis approaches and other academic research [17] have tackled the problem of "normalizing" log messages into a common format as mentioned in Sect. 2, they do not address the semantic aspect of log processing. They lack constructing formal conceptualization of security log data. Conceptualization of a domain knowledge can be used to support unambiguous and useful interlinking between log data so that it can be understood by a machine. A uniform conceptual model can semantically lift heterogeneous security log data from diverse sources. Therefore, our first research question based on the hypothesis:

Hypothesis 1: *Security log data can be structured and enhanced by semantics, to support unambiguous and useful interlinking between logs in a knowledge graph. A uniform conceptual model can semantically lift heterogeneous security log data from diverse sources.*

Distributed enormous sizes of security log data separated from different applications, system and host remains a problem. Traditional security log analyses typically tackle this problem through centralized data integration (e.g. log server, SIEM system, etc.). Centralized data analyses such as ETL (extract-transform-load) [18] style is a typical method for traditional data analytics, in which it doesn't solve the problems of highly verbose, redundant and incoherent and poorly structured information. As systems typically generate high-frequency, fine-grained and vast amounts of log data, analyzing data by means of a centralized method is not ideal for real-time processes or on-demand access, where fast response is required.

RQ2: *How to support analysis of scalable and separated security log information?*

Consider another example case about network interruption. When an interruption happens in a working network of a large system, it then will trigger to put different log messages to various related system logs with the same meaning. Hence, without a automated scalable integration, security analyst may loss of keeping track of log data, particularly when the interruption happens suddenly or in an emergency situation.

We expect that semantic query federation, through decentralized semantic data integration, can retrieve meaningful results from large-scale, dispersed security logs. We expect that it can overcome scalability issues [18] and flexibly combine different datasets to improve attack detection and causal analysis. Thus, we consider that RQ2 is fit with the Hypothesis 2.

Hypothesis 2: *Multiple distributed security log datasets can be retrieved* in a scalable manner *by means of semantic query federation.*

The growing amounts of log data available for security analyses inhibits a timely detection and response to attacks. Current security analysis processes typically rely on human intelligence rather than systems to perform better task inference. Although human experts typically are better to perform inference tasks, they are easily over-burdened by the vast amounts of data. Therefore, security analysts often find it difficult to identify the potential impact of a security incident. Current security log analyses systems provide insufficient inference capabilities as they typically do not provide semantic reasoning.

> **RQ3:** *How to represent and infer causally-related events from security log information?*

Semantic reasoning, in the security domain, can be applied to enable identification of potential attacks by exploiting property chains, transitive and reflexive properties of interlinked security log data. Moreover, semantic reasoning can also be processed in a streaming way by means of an RDF stream reasoner [20] which will allow us to perform continuous queries over incoming log event streams. Stream reasoning will hence allow us to detect potential attacks and suspicious behavior in real-time. Therefore, RQ3 is fit with Hypothesis 3.

> **Hypothesis 3:** *Semantic reasoning can be used to infer causally-related events from security log data. Furthermore, by means of an RDF stream reasoning will allow us to identify potential attacks and suspicious behavior in real-time.*

Our research will propose an innovative security log analysis approach which will contribute results at the intersection between semantic web and cybersecurity research. By providing theory, models and techniques to this approach we expect that the research will have a high-impact on cybersecurity research domain. We expect to obtain these following contributions: (1) conceptualization of security domain (e.g. system, log and event vocabulary, (2) semantic modeling approach of infrastructure and attack patterns (e.g. background knowledge), (3) a framework for data acquisition, integration and semantic reasoning of large-scale and disparate security log information. Our proposed approach will allow experts without particular skills in semantic web and query language to easily analyze large-scale disparate log information and to improve the identification of potential attacks and suspicious behavior in real-time. Therefore, it can improve situational security awareness.

4 Research Methodology and Approach

Based on the research questions we have defined, we decide to apply iterative research methodology Action Research (AR) described by Checkland and Holwell [19]. This method allows us to start research with a literature review, analyze examples of real-world security issues and evaluate existing technologies (e.g. approaches, frameworks, tools, etc.) which are used to tackle these issues in order to analyze their advantages and drawbacks. By this evaluation, we can consider whether we can adapt them in our research or not. We define our research method as three aspects: conceptual model, prototyping and evaluation. Both conceptual model and prototyping are discussed in this section, while evaluation aspect is discussed in Sect. 6.

For the conceptual model, we will conceptualize the architecture of our semantic log processing framework that includes test specification and metrics that can be used to validate our developed framework. We will develop ontologies that represent concepts of system infrastructure, log events, attack patterns and background knowledge. We will also reuse existing upper ontologies to facilitate semantic interoperability and cross-domain knowledge sharing.

As a prototype of our proposed framework, we will instantiate our developed ontologies (e.g. system, log events, patterns) and include several steps such as log extraction, event extraction and integration, and semantic log analysis. Log extraction process will cover log sources acquisition, extraction and conversion from raw data to a certain RDF serialization (e.g. XML/RDF, JSON-LD). By leveraging background knowledge, We will extract explicit events to discover new patterns as they appear in log messages and learn a new type of previously unseen log entries. Event integration process will cover the integration of related events from multiple different sources so that we will have a complete event pattern. Regarding the semantic log analysis, we will implement semantic query federation over distributed event datasets and stream reasoning to infer and discover potential attacks and suspicious behavior in real-time.

Furthermore, this research is related to a research project called SEPSES (Semantic Processing of Security Event Streams) which also serves as a source for ideas of how to approach problems arising on semantic processing of security log data. This project also serves as result comparison to evaluate the proposed approach.

5 Preliminary Result

As explained in Sect. 2, we have surveyed the state-of-the-art research from two different domains: security domain and semantic web domain. On the security domain, we found several existing approaches, both semantic and non - semantic, to analyze the security logs. We have investigated the gaps in security log analyses and formulated a number of research questions. On the semantic web domain, we also conducted surveys of the current semantic query federation approaches.

As a first step, we have already started evaluating several log parsing tools and libraries (e.g. Plaso, Splunk, Logstash) to acquire and parse different log sources from different machines (e.g. Syslog, Authlog, Apachelog etc.). From there, we got several terms (e.g. host, message, timestamps, etc.) and focused to find the most important terms which can generalize the informations of log sources. Then, by those terms we started to define our log vocabularies.

We realize that there are a lot of different type of log sources which might come from different platforms and machines. Therefore, we defined our log vocabulary modularly. It means that we have one log vocabulary as a core and on top of that we have another specific log vocabulary which fit with a certain type of log. We also reused several existing vocabularies that are relevant to our vocabulary concept by attaching them in the log extraction process and we got results as RDF-based log entries on JSON-LD format [21]. We also have already submitted our first paper on log extraction to the upcoming 2018 SEMANTICS Conference.

6 Evaluation Plan

We will continuously evaluate the results of each part of our developed framework. The evaluation will be conducted to measure and to check whether the results have met our research goals. We will assess the ability of our developed framework to evaluate performance characteristics such as throughput and latency. The research will be started with a simple scenario (e.g. login scenario) and implement elaborate data generator that simulate real-world event data (e.g. Syslog, Apachelog, Sys-Log etc.). We will setup a system with various log sources in a virtual environment. Regarding to the semantic stream processing, we will conduct the evaluation using live data that generated by several tester's actions. We will also conduct evaluation towards the addition of new scenarios in order to measure the scalability performance. Based on the evaluation results, we will able to draw conclusions about accuracy, latency and completeness of detection and the overall scalability of our research approach.

7 Conclusion

In this proposal, we provide a research roadmap for security log analysis based on federated linked data querying technology. We outline the problem of technical, syntactical and semantical heterogeneity, physical and logical separation of log data, and the enormous size of security log data that hinders efficient and effective analysis of security log information. Based on our analysis of the state-of-the-art, we formulate three specific research questions, for which we provide our initial hypotheses in this proposal. We describe our research methodology and approach that will guide our research. To conclude, we aim to contribute both to the cyber-security and semantic web domains by developing a novel method that improves the current state of the art in security log analyses.

Acknowledgement. This work was supported by the Ministry of Education and Culture, Indonesia. Furthermore, support for project SEPSES by the Austrian Science Fund (FWF) and netidee SCIENCE: P 30437-N31 is gratefully acknowledged. I want to thank my supervisors, Prof. A Min Tjoa, Prof. Gerald Quichmayr, Dr. Elmar Kiesling and my colleague Fajar Ekaputra, Peb Aryan and Niina Novak for their helpful discussion, comments and feedback.

References

1. FT Services: Cybercrime survey report insight and perspective (2017)
2. Calvanese, D., Montali, M., Syamsiyah, A., Van Der Aalst, W.M.P.: Ontology-driven extraction of event logs from relational databases **256**, 140–153 (2016)
3. Kent, K., Souppaya, M.: Guide to computer security log management. National Institute of Standards and Technology, pp. 1–72 (2006)
4. He, P., Zhu, J., He, S., Li, J., Lyu, M.R.: An evaluation study on log parsing and its use in log mining. In: Proceedings - 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, DSN 2016, pp. 654–661 (2016)

5. Xu, W.: Advances and challenges in log analysis. Commun. ACM **55**(2), 55–61 (2012)
6. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Sci. Am. **284**, 34–43 (2001)
7. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. Int. J. Semant. Web Inf. Syst. **5**(3), 1–22 (2009)
8. Miller, D.R., Harris, S., Harper, A., VanDyke, S., Blask, C.: Security Information and Event Management. McGraw-Hill Osborne Media (2010)
9. Axelsson, S.: Intrusion detection systems: a survey and taxonomy. Department of Computer Engineering (2009)
10. Gander, M., Felderer, M., Katt, B., Tolbaru, A., Breu, R., Moschitti, A.: Anomaly detection in the cloud: detecting security incidents via machine learning. In: Moschitti, A., Plank, B. (eds.) EternalS 2012. CCIS, vol. 379, pp. 103–116. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-45260-4_8
11. Wu, S., Zhang, Y., Cao, W.: Network security assessment using a semantic reasoning and graph based approach. Comput. Electr. Eng. **64**, 96–109 (2017)
12. Hartig, O.: Zero-knowledge query planning for an iterator implementation of link traversal based query execution. In: Antoniou, G., et al. (eds.) ESWC 2011. LNCS, vol. 6643, pp. 154–169. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21034-1_11
13. Schwarte, A., Haase, P., Hose, K., Schenkel, R., Schmidt, M.: FedX: optimization techniques for federated query processing on linked data. In: Aroyo, L., et al. (eds.) ISWC 2011, Part I. LNCS, vol. 7031, pp. 601–616. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-25073-6_38
14. Gorlitz, O., Staab, S.: SPLENDID: SPARQL endpoint federation exploiting VOID descriptions. In: Proceedings of the 2nd International Workshop on Consuming Linked Data, Bonn, Germany (2011)
15. Acosta, M., Vidal, M.-E., Lampo, T., Castillo, J., Ruckhaus, E.: ANAPSID: an adaptive query processing engine for SPARQL endpoints. In: Aroyo, L., et al. (eds.) ISWC 2011. LNCS, vol. 7031, pp. 18–34. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-25073-6_2
16. Verborgh, R., et al.: Triple pattern fragments: a low-cost knowledge graph interface for the web. J. Web Semant. **37–38**, 184–206 (2016)
17. Azodi, A., Jaeger, D., Cheng, F., Meinel, C.: Pushing the limits in event normalisation to improve attack detection in IDS/SIEM systems. In: Proceedings of the 2013 International Conference on Advanced Cloud and Big Data, pp. 69–76. IEEE (2013)
18. Kimball, R., Caserta, J: The Data Warehouse ETL Toolkit. Wiley Publishing, Inc., Indianapolis (2004)
19. Della Valle, E., Ceri, S., van Harmelen, F., Fensel, D.: It's a streaming world! Reasoning upon rapidly changing information. IEEE Intell. Syst. **24**(6), 83–89 (2009)
20. Checkland, P., Holwell, S.: Action research: its nature and validity. Syst. Pract. Action Res. **11**(1), 9–21 (1989)
21. Sporny, M., et al.: A JSON-based serialization for linked data (2014)