A Grammatical Inference Approach to Language-Based Anomaly Detection in XML

Harald Lampesberger
Christian Doppler Laboratory for Client-Centric Cloud Computing, Softwarepark 21, 4232 Hagenberg, Austria
Email: h.lampesberger@cdcc.faw.jku.at

Abstract—False-positives are a problem in anomaly-based intrusion detection systems. To counter this issue, we discuss anomaly detection for the eXtensible Markup Language (XML) in a language-theoretic view. We argue that many XML-based attacks target the syntactic level, i.e. the tree structure or element content, and syntax validation of XML documents reduces the attack surface. XML offers so-called schemas for validation, but in real world, schemas are often unavailable, ignored or too general. In this work-in-progress paper we describe a grammatical inference approach to learn an automaton from example XML documents for detecting documents with anomalous syntax.

We discuss properties and expressiveness of XML to understand limits of learnability. Our contributions are an XML Schema compatible lexical datatype system to abstract content in XML and an algorithm to learn visibly pushdown automata (VPA) directly from a set of examples. The proposed algorithm does not require the tree representation of XML, so it can process large documents or streams. The resulting deterministic VPA then allows stream validation of documents to recognize deviations in the underlying tree structure or datatypes.

Keywords—intrusion detection; anomaly detection; XML; grammatical inference

I. INTRODUCTION

Detecting attacks against software is the research field of intrusion detection systems (IDS). We distinguish IDS techniques into misuse- and anomaly-based detection for hosts and networks: A misuse-based IDS matches signatures in a stream of events or network traffic. Contrary, an anomaly-based IDS isolates events or network packets that deviate from normal behavior. While signatures represent known patterns of misbehavior, normality for anomaly detection is usually approximated from observations by machine learning or stochastic methods.

In theory, anomaly detection has the advantage of recognizing yet unknown (zero-day) or targeted attacks that are specifically designed to evade signatures. Bilge and Dumitras [1] show that zero-day attacks are actually frequent and targeted attacks like Stuxnet [2] will likely reoccur in the future. Anomaly detection seems like a perfect solution but suffers from severe practical problems.

False-positives and the high costs associated with them are one major problem [3]. We don’t know beforehand how often attacks occur, so the ratio of normal to abnormal events can be heavily skewed: Even a system with low false-positive rate could generate an unacceptable number of false-positives. Sommer and Paxson [4] identify further issues why anomaly detection is not adopted outside academia: It is hard to understand semantics of a detected anomaly and the notion of normality is unstable, especially in networks. Commercial antivirus and network IDS software still relies on signature-based techniques and anomaly detection only plays a supporting role in products that offer behavioral analysis.

The goal of this paper is a more promising anomaly detection technique for the eXtensible Markup Language (XML). XML is a platform-independent language for semi-structured data and a pillar of today’s Web. Reducing the attack surface therefore makes sense. For our approach we resort to formal language theory and grammatical inference to understand language-theoretic properties and learnability of XML. We believe, a detection technique can only guarantee low false-positive and high detection rates if it respects these properties.

A. Problem Definition

We consider anomaly detection similar to grammatical inference: Learning a representation of a language, e.g. a grammar or automaton, from the presentation of a language, e.g. from examples, counter-examples or an oracle. Grammatical inference assumes that there is some hidden target representation to be discovered, where language class and type of presentation influence successfullness of learning [5, pp. 141–172]. A learning algorithm is said to converge if the hidden representation is uncovered.

We define our problem as follows: Given a set of example XML documents, a learner returns an automaton that allows validation of syntactic structure and datatypes to decide normality of future documents.

While XML is exchanged as document, the underlying logical model is a tree. Processing the tree as Document Object Model (DOM) [6] requires all the information in memory and this becomes harder with increasing size. We require both automaton and learner to operate in a streaming fashion, where memory and time for processing is limited. The Simple API for XML (SAX) [7] is our streaming interface to documents.

We approach the problem by first discussing expressiveness of XML. For that, we introduce a formal abstraction of practical schema languages and show that language representation through visibly pushdown automata (VPA) is equivalently expressive. VPA are an executable model capable of stream processing documents and satisfy our requirement. We then characterize an XML language class that can be efficiently learned from a set of example...
documents. Content in XML is from an unknown language class in general. We therefore introduce a datatype system for abstracting content of possibly infinite nature into a finite set of datatypes. The contributions are an XML Schema compatible lexical datatype system and a state-merging algorithm for learning VPA. An inferred VPA can validate future documents to recognize anomalous syntax.

The paper is structured as follows: In the remaining introduction we discuss vulnerabilities, XML-based attacks and introduce our learning setting. In Section II we define notations, schema languages and analyze XML expressiveness and its limits for stream validation. VPA are introduced in Section III. Section IV presents our datatype system and learning algorithm. Related work is listed in Section V and Section VI concludes this paper.

B. A Language-Theoretic View on Security

Sassaman et al. [8] analyze the software vulnerability problem using formal language theory. Modularization and composition is an important process in software engineering but implicitly requires interfaces and protocols between components. A protocol basically specifies the syntax and semantics of a formal language for encoding information, e.g. a file format or network message.

When two components $R$ and $S$ interact, the sender $S$ encodes information w.r.t. the protocol as transportable object, e.g. a network message or file. The receiver $R$ decodes (parses) this object according to the protocol and $R$’s internal state is updated in the process. Unfortunately, protocols in the real world are often ambiguous, underspecified or implementations have errors [8]. Sender $S$ might be able to craft a special object such that $R$ moves into an unexpected or insecure state upon parsing. This object is then called exploit because it bends or breaks the original intention of the protocol; We say $S$ abuses a vulnerability in the protocol to attack $R$.

An unambiguous and precise protocol specification is required to resolve vulnerabilities such that the receiving component can reject malformed entities [8]. This is exactly the membership decision problem in formal languages and it may be intractable or undecidable depending on the language class. Another difficulty is that protocols are often layered such that several languages are embedded within each other, e.g. TCP/IP or content in an XML document.

Today’s IDS are typically engineered around a specific language class, where computational complexity is tractable. Nevertheless, their goal is to detect exploits in a possibly larger language class or across several layers of embedded languages. False-positives and false-negatives are a direct consequence of mismatching language classes. For example, misuse-based IDS are often restricted to the class of regular word languages ($\mathit{REG}$). If the class of the observed protocol is greater than $\mathit{REG}$ and there is a vulnerability, there might be infinite variants of exploits that evade signatures over $\mathit{REG}$. Understanding the language-theoretic problems is therefore important.

C. Why Secure XML Processing Matters

XML takes the role of the protocol in Web browsers, mobile applications and Web services. The logical tree structure allows high expressiveness but correct processing becomes more complex and vulnerabilities arise. DOM parsers are vulnerable to Denial-of-Service (DoS) attacks that exhaust time and memory, for example by overlong element names or oversized payload. A coercive parsing attack causes DoS by nesting a vast amount of tags [9].

If an XML parser respects the Document Type Definition (DTD) in the preamble of a document, several DoS attacks based on entity expansion become a threat. Furthermore, the XML parser could expose confidential information if external entity references enable local file import [9].

XML injection is a large class of XML attacks, where the attacker controls parts of a document. Figure 1 describes a fictional transaction document, where a monetary amount is given and the user provides a credit card number. In Figure 1c, the attacker manipulates the transaction value in the DOM tree when a DOM parser is in place [9]. XML injection affects SAX parsers too if the parser state is not propagated correctly. Cross-Site Scripting in the Web is also a form of injection, where a script or iframe is embedded. Classic attacks like SQL-, command- or XPATH-injection are also a threat if the application that utilizes the XML parser is vulnerable.

Note that all the presented example attacks change the expected syntax of a document. Unexpected tree structure or wrong datatypes could lead to harmful interpretation in the XML processing component. Falkenberg et al. [9] and Jensen et al. [10] recommend strict validation of XML documents to mitigate attacks but validation requires a language representation, i.e. a schema.

Unfortunately, validation is not common. Only 8.9% of XML documents in the Web refer validate to a schema [11]. Also, Web paradigms like Asynchronous JavaScript and XML (AJAX) [12] do not enforce schemas or validation, so developers are misled to ad-hoc design. This motivates learning a language representation from effectively communicated XML for later validation.
D. Learning in the Limit

We consider Gold’s learning in the limit from positive examples [13] as our grammatical inference setting. The target class, a language class \( \mathcal{L} \) expressible by a class of language describing devices \( \mathcal{A} \), is identifiable in the limit if there exists a learner \( I \) with the following properties: Learner \( I \) receives as input enumerated examples \( E(1), E(2), \ldots \) of some language \( L \in \mathcal{L} \), where \( E : \mathbb{N} \rightarrow L \) is an enumeration of \( L \), and examples may be in arbitrary order with possible repetitions. With every input, \( I \) returns the current hypothesis \( A_i \in \mathcal{A} \), e.g. a grammar or automaton, and there is a point of convergence \( N(E) \): For all \( j \geq N(E) \), \( A_j = A_{N(E)} \) and the language of \( A_{N(E)} \) is \( L \). We call \( I \) a learner for target class \( \mathcal{L} \) if there is convergence for all \( L \in \mathcal{L} \). A sample set \( S_+ \subseteq L \) is called characteristic if learning converges when \( S_+ \) is enumerated to \( I \) [14].

Unfortunately, grammatical inference is hard and even the class \( \mathcal{REG} \) is not learnable in the limit from positive examples only [13]. Learning from XML documents is even harder because it is a context-free word language. Ignorance of learnability properties reflects in bad practical performance of anomaly-based IDS. We therefore approach the problem more formally and present a learner for a restricted class of XML in Section IV.

II. XML

The logical structure of XML is a tree, where \( \Sigma \) always denotes the alphabet of element names. We encode attributes as elements with a leading `@`-character and namespaces as part of the element name. An encoding example is in Figure 2. We disregard identifiers and references because they change the logical structure.

The structure without element content or attribute values is characterized by \( \Sigma \)-trees [15]. The inductive definition of \( \mathcal{T}_\Sigma \), the set of all \( \Sigma \)-trees, is as follows: (1) every \( e \in \Sigma \) is a \( \Sigma \)-tree; (2) if \( e \in \Sigma \) and \( t_1, \ldots, t_n \in \mathcal{T}_\Sigma \), then \( Dom(t) = \{e\} \cup \{i. u \mid i \in \{1, \ldots, n\}, u \in Dom(t_i)\} \). A tree \( t \) is the root of the tree and node \( v, j \) is the \( j \)-th child of node \( v \). The label of \( v \) in \( t \) is \( lab(t, v) \). A tree language \( L \) over \( \Sigma \) is a set of trees such that \( L \subseteq \mathcal{T}_\Sigma \).

A document is a \( \Sigma \)-tree encoded with tags. For notational convenience, we strip angled brackets such that the set of open-tags is \( \Sigma \) and the set of close-tags becomes \( \Sigma = \{\tau \mid c \in \Sigma\} \). We use variables \( c_t, c_1, c_2, \ldots, c_n \in \Sigma \) for open-tags and \( \tau_t, \tau_1, \tau_2, \ldots, \tau_n \in \Sigma \) for the according close-tags. XML documents without content are words over \( (\Sigma \cup \Sigma) \cup \{\varepsilon\} \) and well-matched if they obey the grammar \( W := W W | c W \tau \mid \varepsilon \).

A tree is translated into a document by pre-order traversal (document order) and we denote the function \( doc \) as the bijection between trees and documents.

A. Schemas and Types

A schema is a tree grammar that restricts expressible XML over elements \( \Sigma \) and implicitly gives meaning to structure. DTDs are the simplest form of schemas:

**Definition 1 (DTD [16]):** A DTD is a triple \( (\Sigma, d, s_d) \), where production rules \( d : \Sigma \rightarrow \mathcal{REG}(\Sigma) \) map element names to regular expressions over \( \Sigma \) and \( s_d \) is the distinguished start element. The right-hand side of production rules is called content model and \( L(d) \) is the set of trees that satisfy \( d \).

The expressible language class \( DTD \) is rather limited and practical schema languages like XML Schema (XSD) [17] or Relax NG [18] offer types to increase expressivity. Types are from a finite set, each type is associated with a unique element name and the start element has exactly one type [16]. Variables \( m, m_0, n \) in this paper always denote types. As a formal abstraction of practical schema languages we recall the definition of extended DTD (EDTD):

**Definition 2 (EDTD [16]):** An EDTD \( D \) is a tuple \( D = (\Sigma, M, d, m_0, \mu) \), where \( M \) is a set of types, \( \mu : M \rightarrow \Sigma \) is a surjection from types onto element names and \( (M, m_0) \) is a DTD over types. A tree \( t \) satisfies \( D \) if \( t = \mu(t') \) for some \( t' \in L(d) \), where \( \mu \) ranges over trees. Tree \( t' \) is called witness for \( t \), \( L(D) \) denotes the set of trees and \( L^w(D) \) denotes the documents that satisfy \( D \). The language class \( \mathcal{EDTD} \) expressible by EDTDs is equivalent to the regular tree languages [15].

B. Stream Validation and Expressiveness

The stream validation or type-checking problem is to decide whether a document is in the language of a given schema within a single pass. Typing a document \( w \) is to assign every position \( i \) (every tag) some type. A document is valid w.r.t. a schema if such an assignment is possible for all positions. Note that function \( \mu \) is surjective and a position can have multiple types in general.

Martens et al. [16] and Murata et al. [19] discuss ambiguity and determinism of schemas: A schema is ambiguous if there is a document in the language with multiple types at some position. A schema is deterministic if for all described documents at all positions the choice is limited to a single type. In other words, a schema is deterministic if every type assignment is clear when an open-tag is read. Note that ambiguity always implies nondeterminism. Martens et al. [16] introduce the 1-pass pre-order typed (1PPT) property for EDTDs that are deterministic and therefore allow efficient stream validation.

Note that determinism is also important for efficient processing. This is one factor why content models in practical schema languages like DTD and XSD have restrictions to enforce determinism [17]. We direct the reader to Martens et al. [16] for a thorough analysis of expressiveness of schemas.

We have \( DTD \subseteq \mathcal{EDTD}^d \subseteq \mathcal{EDTD}^c \subseteq \mathcal{EDTD} \), where \( \mathcal{EDTD}^d \) is the class of schemas that satisfy the restrictions of XSD and \( \mathcal{EDTD}^c \) is the class of deterministic schemas, whereas the 1PPT property holds [16].
C. Datatypes and Mixed Content

XML documents carry data as element contents or attribute values and the tag encoding of documents guarantees that tags and data are not confused. We denote data in a document as words over alphabet \( U \), i.e. Unicode, and variables \( r, s \in U^* \) always represent data. But data could be from any language class, for example natural language or program code. As an abstraction of data, we introduce datatypes:

**Definition 3:** A so-called lexical datatype system is a tuple \( (\Delta, U, \phi) \), where \( \Delta \) is a finite set of datatypes and \( \phi : \Delta \rightarrow \mathcal{P}(U^*) \) is a surjection that assigns every datatype its lexical space as some language over \( U \). Because it is surjective, a datum \( r \) may have several matching datatypes. Variables \( a, b \in \Delta \) always denote datatypes in this paper.

A datatype system has functions \( \text{types}: U^* \rightarrow \mathcal{P}(\Delta) \) and \( \text{firstType}: U^* \rightarrow \Delta \). While \( \text{types}(r) \) returns all matching datatypes for some \( r \), \( \text{firstType}(r) \) chooses one matching datatype to reduce arbitrary data to datatypes.

With respect to the tree structure, content \( r \) between tags \( cr\tau \) is encoded into a single child node \( v,j \) representing the datatype, where \( lab^f(v) = c \) and \( lab^f(v,j) = r \). We call such a node data node for short. Data nodes are always tree leafs and we now refine EDTDs with datatypes:

**Definition 4:** A datatype extended DTD (\( \Delta \)-EDTD) is a tuple \( D = (\Sigma, M, d, m_0, \mu, \Delta, U, \phi) \), where elements \( (\Sigma, M, d, m_0, \mu) \) form an EDTD and \( (\Delta, U, \phi) \) is a datatype system. As an extension to EDTDs, production rules \( d : M \rightarrow \mathcal{R}(\mathcal{E}(M \cup \Delta)) \) assign every type a content model as regular expression over both types and datatypes. Production rules only allow expressions, where a datatype is followed by either a type or \( \epsilon \), but never a subsequent datatype. Lexical spaces \( \phi: \Delta \rightarrow \mathcal{R}(\mathcal{E}(U)) \) are restricted to regular expressions over \( U \).

A datatyped tree \( t \) over \( (\Sigma \cup \Delta) \) satisfies \( D \) if \( t' = \mu(t'') \) for some \( t'' \in L(d) \), where \( \mu \) applies only to elements. We denote \( L_\Delta(D) \) as the set of datatyped trees that satisfy \( D \). A tree with data nodes \( t \) satisfies \( D \) if there is a datatyped tree \( t' \in L_\Delta(D) \) such that \( lab^f(u) \in \phi(lab^f(u)) \) holds for all data nodes \( u \) and \( lab^f(v) = lab^f(v) \) holds for all nodes \( v \) with \( lab^f(v) \in \Sigma \). Language \( L(D) \) denotes the set of trees with data nodes that satisfy \( D \).

Accordingly, we define the word languages generated by \( \Delta \)-EDTD \( D \). Suppose that bijection \( \text{doc} \) transforms trees with data nodes and datatyped trees into documents like in Figure 2. Then \( L^\text{doc}_\Delta(D) = \{ \text{doc}(t) \mid t \in L_\Delta(D) \} \) is the datatyped word language and the document (word) language generated by \( D \) is \( L^\text{doc}_\text{w}(D) = \{ \text{doc}(t) \mid t \in L(D) \} \).

\( \Delta \)-EDTDs allow so-called mixed content, where between two tags both data and other nested tags are allowed. Mixed content typically appears in markup languages, e.g. the XML Hypertext Markup Language (XHTML). Regarding structural expressiveness, the language classes of EDTDs also translate to our definition of \( \Delta \)-EDTDs. We now have an abstraction of schema languages that captures attributes, datatypes and mixed-content on a syntactic level.

III. VISIBLY PUSHDOWN AUTOMATA FOR XML

The well-matched tags in documents induce a visible nesting relation. In fact, XML is a visibly pushdown language (VPL) [20] and Kumar et al. [21] show that every EDTD-definable document language is a VPL. This property holds for our definition of \( \Delta \)-EDTDs because tags are still well-matched and we encode attributes as nested elements. VPLs are accepted by visibly pushdown automata (VPA), a restricted form of pushdown automata, where the input symbol determines the stack action.

**Definition 5 (VPA [20]):** \( A = (\Sigma, Q, q_0, Q^F, \Gamma, \delta) \) is a VPA, where \( \Sigma = (\Sigma_\text{call}, \Sigma_\text{int}, \Sigma_\text{ret}) \) is the pushdown alphabet made of three distinct alphabets, \( Q \) is the set of states, \( q_0 \in Q \) is the start state, \( Q^F \subseteq Q \) are the final states, \( \Gamma \) is the stack alphabet and the transition relation is \( \delta = \delta_\text{call} \cup \delta_\text{int} \cup \delta_\text{ret} \), where \( \delta_\text{call} \subseteq (Q \times \Sigma_\text{call} \times Q \times \Gamma) \), \( \delta_\text{int} \subseteq (Q \times \Sigma_\text{int} \times Q \), and \( \delta_\text{ret} \subseteq (Q \times \Gamma \times \Sigma_\text{ret} \times Q) \).

A transition \((q,c,q',\gamma) \in \delta_\text{call},\) denoted as \( q \xrightarrow{c\gamma} q' \), is a call-transition from state \( q \) to \( q' \) that pushes \( \gamma \) on the stack when symbol \( c \in \Sigma_{\text{call}} \) is read. A transition \((q,\gamma,\tau,q') \in \delta_\text{ret},\) written as \( q \xrightarrow{\gamma\tau} q' \), is a return-transition from state \( q \) to \( q' \) that pops \( \gamma \) from the stack when symbol \( \tau \in \Sigma_{\text{ret}} \) is read. An internal transition \((q,a,q') \in \delta_\text{int},\) denoted as \( q \xrightarrow{a} q' \), moves from state \( q \) to \( q' \) at input \( a \in \Sigma_{\text{int}} \) without changing the stack. We direct the reader to Alur and Madhusudan [20] for the semantics of VPA.

Contrary to traditional pushdown automata, VPA can be determinized and are closed under complement, intersection, union, concatenation and Kleene-star. Also language equivalence, emptiness, universality and inclusion are decidable.
Next we will show the equivalence of $\Delta$-EDTDs and XML VPA (XVPA) [21]. XVPA are a special form of modular VPA that go back to program modeling. In a modular VPA, states are partitioned into modules and the stack alphabet is exactly the set of states. When a module calls another one, the current state is saved on the stack and popped for returning. With respect to XML, modules are exactly the types. Call, return and internal transitions of the VPA are the open-tag, close-tag and character events of the SAX interface to documents.

We assume the following about SAX: There exists a global datatype system $(\Delta, U, \phi)$ and every datum $r$ between two tags or attribute value is reduced to one character event. For stream validation, the SAX interface reports only the first matching datatype $\text{firstType}(r)$ to the XVPA instead of $r$ for efficiency. So, the XVPA processes datatyped documents and the internal alphabet over datatypes is guaranteed to be finite.

**Definition 6 (XVPA [21]):** An XVPA $A$ is a tuple $A = (\Sigma, \Delta, M, \mu, \{Q_m, c_m, X_m, \delta_m\}_{m \in M}, m_0, F)$, where $\Sigma, \Delta, M$ and $\mu$ have the same meaning as in $\Delta$-EDTDs, $m_0$ is the distinguished start type and $F = X_{m_0}$ are final exit states. Every type $m \in M$ characterizes a module, where

- $Q_m$ is the finite set of module states,
- $c_m \subseteq Q_m$ is a single entry state of the module,
- $X_m \subseteq Q_m$ is the exit of module $m$ (exit states),
- Transitions $\delta_m = \delta^\text{call}_m \cup \delta^\text{ret}_m \cup \delta^\text{int}_m$, where
  - $\delta^\text{call}_m \subseteq \{q_m \xrightarrow[c_m]{e_m} q_n | n \in \mu^{-1}(e)\}$,
  - $\delta^\text{ret}_m \subseteq \{q_m \xrightarrow[r_m]{q_n} q_n | q_n \in X_m \land n \in \mu^{-1}(e)\}$
  and is deterministic, i.e. $q_n = q'_n$ whenever $q_m \xrightarrow[r_m]{q_n}$ and $q_m \xrightarrow[r_m]{q'}$, and
  - $\delta^\text{int}_m \subseteq \{q \xrightarrow[a_m]{q'} | q, q' \in Q_m \land a \in \Delta\}$.

Return transitions are always deterministic by definition. The XVPA is deterministic if also the call transitions are deterministic. The semantics of an XVPA are given by its corresponding VPA $A' = (\Sigma, Q, q_0, q_f, Q, \delta)$, where $\Sigma = (\Sigma, \Delta, \delta)$, $q_0$ and $q_f$ are start and accepting state, $Q = \{q_0, q_f\} \cup \bigcup_{m \in M} Q_m$ and transition function $\delta$ is defined as

$$
\delta = \bigcup_{m \in M} \{q_m \xrightarrow[\mu(m)/q_0]{e_m} q_n \} \cup \{q \xrightarrow[\mu(m)/q_0]{q_n} q_f | q \in F\}.
$$

The language $L_A(m)$ of module $m$ is a datatyped word language and accepted words are of form $\mu(m)w\mu(m)$. The accepted language $L(A) = L_A(m_0)$ of XVPA $A$ is the datatyped word language $L(A')$ of its corresponding VPA.

The set $X_m$ are exit states, where at least one return transition originates from. In a valid XVPA, the single-exit property [21] must hold: If there is some return transition $q_m \xrightarrow[r_m]{q_n}$ from module $m$ to $n$, then there must be return transitions $q'_m \xrightarrow[r_m]{q_n}$ for all exit states $q'_m \in X_m$. The single-exit property guarantees that $L_A(m)$ is always the same, independent from the calling state or module.

**Theorem 1:** Given a datatype system $(\Delta, U, \phi)$, every $\Delta$-EDTD $D$ has a corresponding XVPA $A$ such that their datatyped word languages are equal $L(A) = L^w_D(D)$. Also, for every XVPA $A$ there is an equivalent $\Delta$-EDTD $D$ such that $L^w_D(D) = L(A)$.

The proof is skipped, it refines Kumar et al. [21] with datatypes. Intuitively, every type $m$ has an intermediate DFA $D_m$ in the translation between XVPA modules and regular expressions $d(m)$ in $\Delta$-EDTD production rules. The time complexity of the automaton for processing a document is linear in the length of the document and the required space is bounded by the nesting depth.

**IV. INFERENCE FROM STREAMING XML**

The goal is to learn an XVPA from a set of example documents $S_+$.

Inference boils down to (1) defining a datatype system, (2) characterizing types and states and (3) learning the language over types and datatypes from examples in $S_+$.

**A. XML Schema Compatible Lexical Datatype System**

XVPA in our definition use datatypes as finite alphabet for internal transitions. In Section II we introduce the notion of datatype system but for inference we need a concrete instance. XSD defines a rich set of 47 atomic datatypes together with a hierarchy [22], where every datatype has a semantical value space and a lexical space that is characterized by a regular expression. Unfortunately, the lexical spaces of XSD datatypes heavily overlap, for example the word ‘0’ is in the lexical space of datatypes boolean, Integer, or string to name a few. A learner only experiences the lexical space and this leads
to the problem of choosing the correct datatype for a set of words.

Let \((\Delta, U, \phi)\) be our datatype system, where \(U\) is the Unicode alphabet. Based on XSD datatypes we define \(\Delta\) as a poset of 44 datatypes and it is shown in Figure 3. The partial order is the subset relation \(\subseteq\) over individual lexical spaces, i.e.

\[
a, b \in \Delta: a \leq b \iff \phi(a) \subseteq \phi(b) ,
\]

surjection \(\phi\) maps datatypes to the lexical space definitions of XSD [22] respectively. The following adoptions to datatypes and lexical spaces are made:

- Datatype anyURI has an unrestricted lexical space in the XSD standard. In our definition, a datatypes has anyURI if it is a RFC 2396 Unified Resource Identifier with a defined scheme and path.
- The exponents of datatype \(\Delta\) are double are unrestricted.
- Datatypes float, IDREF, IDREFS, ENTITY, ENTITIES, ID, NOTATION and NMTOKENS are dropped because their lexical spaces are indistinguishable from others.
- We add boolean0, boolean1, booleanNum, evenLen-Integer and base64BinaryLF to resolve some of these ambiguities.

If some content \(r\) matches datatype \(a\) then it also matches datatypes \(b_1, b_2, \ldots, b_n\) if \(a \leq b_i\) for \(1 \leq i \leq n\). So, the best characterization of \(r\) is its minimal datatype. We refine functions types and firstType to reflect the partial order: Function \(\text{types}(r)\) returns all matching minimal datatypes and \(\text{firstType}(r)\) chooses one matching minimal datatype. For inference, we define the inverse closure \(cl^{-1}: \mathcal{P}(\Delta) \rightarrow \mathcal{P}(\Delta)\) that returns all datatypes that are smaller or equivalent than a given set of datatypes.

B. Characterizing Types and States

Martens et al. [16] characterize types in XSD based on ancestors. Given document \(w\) and position \(i\), then the ancestor string \(\text{anc-str}(w, i) = c_1c_2\ldots c_j\) is the string of unmatched open-tags in the document prefix \(w_{1,i} = c_1w_1c_2w_2\ldots c_jw_j\). A schema has ancestor-based types if there exists a hypothetical function \(f: \Sigma^* \rightarrow M\) that assigns every open-tag at position \(i\) in document \(w\) a single type \(f(\text{anc-str}(w_{1,i}, i))\). This restriction is exactly the Element Declarations Consistent (EDC) rule of XSD [17]. Identifying types is then defining relation \(\sim_M\) that partitions \(\Sigma^*\) into equivalence classes of ancestor strings. Note that we restrict our learning algorithm automatically to a subset of language class \(\Delta - \mathcal{EDT^D}^{\text{at}}\) by assuming that types are ancestor-based.

Regarding content models we know that the full class of regular languages is not learnable from positive examples [13]. Bex et al. [23] show that the majority of regular expressions in real world schemas are in fact simple such that every type occurs at most \(k\) times in an expression \((k\text{-ORE})\). The language of a \(k\text{-ORE}\) is a \((k + 1)\text{-testable regular language}, where grammatical inference from positive examples is feasible [24].

1: function \(\text{DrVPPA}^{\text{at}}(S_{\tau})\)
2: \(
\text{global } (\Delta, U, \phi) \quad \triangleright \text{datatype system}
\)
3: \(\Sigma \leftarrow Q_{\sigma} \quad \leftarrow \delta_{\text{out}} \quad \leftarrow \delta_{\text{ret}} \quad \leftarrow \delta_{\text{int}} \quad \leftarrow \emptyset \quad \triangleright \text{initialization}
\)
4: \(t: Q \times Q \rightarrow \mathcal{P}(\Delta) \quad \triangleright \text{empty dictionary}
\)
5: \(q_0 \leftarrow (\epsilon, \epsilon) \quad \triangleright \text{init}
\)
6: \(Q \leftarrow \{q_0\} \quad \triangleright \text{init}
\)
7: for all \(w \in S_{\tau}\) do \(\triangleright \text{iterate over documents}
\)
8: \(\text{stack} \leftarrow [\_] \quad \triangleright \text{init}
\)
9: \(q \leftarrow q_0 \quad \triangleright \text{init}
\)
10: for all \((\text{event}, \text{data}) \in \text{SAXEvents}[w]\) do
11: if \(\text{startElement}(\text{event})\) then \(\triangleright \text{open-tag}
\)
12: \(\Sigma \leftarrow \Sigma \cup \{\text{data}\} \quad \triangleright \text{open-tag}
\)
13: \(p \leftarrow \text{pop}(\text{stack}) \quad \triangleright \text{close-tag}
\)
14: \(q' \leftarrow (\pi_1(q) \cdot \text{data}, \epsilon) \quad \triangleright \text{close-tag}
\)
15: \(\delta_{\text{ret}} \leftarrow \delta_{\text{ret}} \cup \{(q \text{ data/q } q')\} \quad \triangleright \text{close-tag}
\)
16: else if \(\text{endElement}(\text{event})\) then \(\triangleright \text{close-tag}
\)
17: assert\(\text{data} = \pi_{-1}(\pi_1(q))\) \(\triangleright \text{match}\)
18: \(p \leftarrow \text{pop}(\text{stack}) \quad \triangleright \text{open-tag}
\)
19: \(q' \leftarrow (\pi_1(q), \pi_2(p) \cdot \text{data}) \quad \triangleright \text{open-tag}
\)
20: \(\delta_{\text{int}} \leftarrow \delta_{\text{int}} \cup \{q \text{ data/q } q'\} \quad \triangleright \text{context}
\)
21: else if \(\text{character}(\text{event})\) then \(\triangleright \text{content}
\)
22: \(q' \leftarrow (\pi_1(q), \pi_2(q) \cdot \epsilon) \quad \triangleright \text{content}
\)
23: \(\delta_{\text{ret}} \leftarrow \delta_{\text{ret}} \cup \{(q \text{ data/q } q')\} \quad \triangleright \text{context}
\)
24: end if
25: \(Q \leftarrow Q \cup \{q'\} \quad \triangleright \text{context}
\)
26: \(q \leftarrow q' \quad \triangleright \text{context}
\)
27: end for
28: \(Q_{F} \leftarrow Q_{F} \cup \{q\} \quad \triangleright \text{context}
\)
29: end for
30: \(\delta_{\text{int}} = \{q \rightarrow q' \mid a \in \epsilon \rightarrow \ell((t(q, q'))\} \quad \triangleright \text{int. transitions}
\)
31: \(\text{return } \{(\Sigma, \Delta, \Sigma), Q, q_0, Q_{F}, \delta_{\text{int}} \cup \delta_{\text{ret}}\} \quad \triangleright \text{context}
\)
32: end function

Figure 4. Visibly Pushdown Prefix Acceptor for class \(\Delta - \mathcal{EDT^D}^{\text{at}}\).

Our learning strategy is state-merging: We first construct a specific VPA that represents exactly \(S_{\tau}\) and then generalize by merging similar states. We denote pairs \((x, y) \subseteq (\Sigma^* \times (\Sigma \cup \{\$\})^*))\) as VPA states, where \(x\) is an ancestor string and \(y\) is a left sibling string \(lsib-str(w, i)\). Symbol \(\Sigma \not\in \Sigma\) denotes a placeholder for XML content and \(lsib-str(w, i) = d_1c_1d_2c_2\ldots d_nc_n\). Here \(c\) is the rightmost unmatched open-tag in the document prefix \(w_{1,i} = ucd_1c_1v_1\overline{v_1}d_2c_2v_2\overline{v_2}\ldots d_{n-1}v_{n-1}c_{n-1}\overline{c_{n-1}}d_n\), \(d_1, d_2, \ldots, d_n \in \{\$, \epsilon\}\) are optional placeholders and the well-matched substrings \(c_i v_i \overline{v_i}\) for \(1 \leq j < n\) represent sibling nodes in the tree w.r.t. position \(i\). As an example, suppose \(w\) is the document in Figure 2a and position \(i\) is just before tag \(</\text{review}>\) then \(lsib-str(w, i) = $ \cdot \epsilon \cdot $ \).

C. The Learning Algorithm

Intuitively, the inference algorithm (1) constructs a so-called visibly pushdown prefix acceptor (VPPA) from the sample set, (2) merges similar states, (3) partitions states into modules, (4) adds missing return transitions to satisfy the single-exit property of XVPA and (5) minimizes the XVPA by merging equivalent modules. Figure 5 gives the full algorithm.

With \(\pi_1, \pi_2, \ldots, \pi_n\) we denote projections of the first, second and \(n\)-th element and \(\pi_{-1}\) is the last element of a tuple or word A VPPA is a deterministic VPA that represents exactly the examples from \(S_{\tau}\) and construction requires only a single pass. The idea of a VPPA is that
1: function INFERDtXVPA\(\delta\)(k, l, S\(_+\))
2: global \((\Delta, U, \phi)\) \(\triangleright \) datatype system
3: \((\Sigma, \ldots, Q, Q_0, P, F', \delta) \rightarrow \) DTVP\(\delta\)(S\(_+\))
4: while \(\exists q, q_2 \in Q; q_1 \prec_{k,l} q_2\) do \(\triangleright \) state merging
5: mergeStates\((f_{k,l}, q_1, q_2)\)
6: end while
7: \(M \leftarrow \{\pi_1(q)\} \text{ for all } q \in Q \land \pi_1(q) \neq \epsilon\)
8: \(m_0 \leftarrow \pi_1(\delta(q_0, c, q_0)) \triangleright \) module called by \(q_0\)
9: for all \(m \in M\) do \(\triangleright \) XVPA conversion
10: \(c_m \leftarrow (m, \epsilon)\)
11: \(Q_m \leftarrow \{q \in Q \mid \pi_1(q) = m\}\)
12: \(\delta_m \leftarrow \{\delta \mid \pi_1(\delta) \in Q_m\}\)
13: \(X_m \leftarrow \{q_m \mid c, p_m; q_m : (q_n, \pi_{p_m} \rightarrow q_n) \in \delta_m\}\)
14: \(\delta_m \leftarrow \delta_m \cup \{q_m \mid \pi_{p_m} \rightarrow q_n \mid \text{ for all } q_m \in X_m\}
15: \text{if } \exists q_m' : (q_m', \pi_{p_m} \rightarrow q_n) \in \delta_m\}
16: end for
17: while \(\exists m,n \in M; m \sim_{\Delta} n\) do \(\triangleright \) minimization
18: mergeModules\((m,n)\)
19: end while
20: \(\mu \leftarrow \{m \rightarrow c \mid m \in M \land \exists q : (q \rightarrow_c e_m) \in \delta^{cat}\}\)
21: end function

Figure 5. The learning algorithm returns an XVPA with datatypes.

D. Example and Discussion

Figure 6 gives a toy example, where the sample set holds a single document. The SAX interface abstracts the contents 10.0 and TEXT into simplified datatypes decimal\(_\Delta\) and stringa\(_\Delta\) respectively. Note that the state (ab, \(\epsilon\)) is visited twice during the VPPA construction because both open-tags \(b\) in context of element \(a\) have the same ancestor-string \(ab\). During state merging, the states (a, ab) and (a, abb) are collapsed into the single state (a, b). In the example, the parameter \(l = 2\) leads to two different types \(a\) and aa in the final XVPA. While both corresponding modules are called by the same tag \(a\), they have completely different content models.

The parameters \(k\) and \(l\) constrain locality of a state. The language class \(\Delta^{\text{EDT}}D_{k,l}^{\Delta} \subseteq \Delta^{\text{EDT}}D_{k,l}^{\text{str}}\) is learnable if \(k\) and \(l\) are bound and \(S_+\) is characteristic such that every valid transition in the XVPA appears at least once in the set. Unfortunately, we do not know whether a sample set is characteristic. But we can guarantee that the quality of the learned automaton stays the same or improves with every example in the sample set if the hidden target is in language class \(\Delta^{\text{EDT}}D_{k,l}^{\text{str}}\).

If \(l = 1\) then types are exactly element names and the algorithm learns a proper subset of \(\Delta^{\text{DTD}}\). A parameter \(k = 1\) limits the left-sibling string of a state to element names or the \(\$\) symbol, so inferred XVPA modules become equivalent to Single Occurrence Automata [26] in terms of expressiveness. In the case that \(k\) and \(l\) are chosen too small, the resulting automaton over-generalizes the language. Contrary, increasing the parameters requires much larger characteristic sets for convergence.

V. RELATED WORK

XML stream validation is first discussed by Segoufin and Vianu [27]. Kumar et al. [21] introduce VPA as an executable model for XML that captures the entire class of regular tree languages. Schewe et al. [28] extend VPA for approximate XML validation and Picalausa et al. [29] present an XML Schema framework using VPA.

For a survey of grammatical inference we direct the reader to the book of de la Higuera [5]. Fernau [25] introduces function distinguishable languages and we apply this concept in Section IV for state merging. Kumar et al. [30] mention that query learning VPA with counterexamples is possible but our setting is different.

Several results on DTD inference from XML have been published [26], [31]–[33], but we aim for the strictly larger class of XSDs. Mlýnková [34] presents a survey of XSD inference. The general idea is to start with an extended context-free grammar as schema abstraction, inferred from examples, and merge non-terminals [35]. Hegwald et al. [36] and Chidlovskii [37] also handle datatypes in their

\(D_m\) and \(D_n\) as constructed in the proof of Theorem 1 are equivalent. If \(m \sim_{\Delta} n\) we redirect all calls and returns from \(n\) to \(m\) and remove \(n\). Finally, \(\mu\) maps all types to the elements they are called by. Note that learning the VPPA and state merging can be combined into one efficient step.
presented methods. Our approach is similar to Bex et al. [38]. Their algorithms use tree automata for learning l-local Single Occurrence XSDs in a probabilistic setting but without datatypes.

In the field of information retrieval, Kosala et al. [39] and Raeymaekers et al. [40] give algorithms to infer HTML wrappers as tree automata. Regarding intrusion detection, Rieck et al. [41] introduce approximate tree kernels as a similarity measure for trees and use them for anomaly detection in HTML.

To our knowledge the presented approach is the first that directly learns an automaton model with both streaming and datatypes in mind. A hard problem in learning schemas is to find nice regular expressions for content models. We focus on learning an automaton representation and intentionally leave conversion to regular expressions open, as many of the noted references propose heuristics or solutions.

VI. CONCLUSION AND FUTURE WORK

We approached the problem of anomaly detection in XML more formally and introduced $\Delta$-EDTDTs as abstraction of practical schema languages with datatypes. We showed that XVPA are an equivalent model capable of stream validation and contributed a lexical datatype system and an algorithm for learning an XVPA from a set of documents. The algorithm converges for target class $\Delta$-EDTDT$^k_{\ell,r}$ given the sample set is characteristic. A learned automaton could theoretically be converted into an XSD schema.

The presented work is still in an early stage. We already have a working prototype which is our baseline for further research and the next step is a thorough evaluation with XML-based attacks. First experiments with the prototype indicate that abstraction by the lexical datatype system using XSD datatypes is too coarse in some cases. We will therefore look into approximations of specific datatypes during learning. Other improvements are to extend the learnable language class and redefine the algorithms for incremental learning. Also, we do not know if some sample set is characteristic and leads to convergence. A refinement to a probabilistic learning setting could enhance applicability when sample sets are incomplete or noisy.

Finally, it is of great interest how our approach to XML inference and stream validation translates to other prominent semi-structured languages like JSON or HTML. An application in mind is a client-side component that learns how Web applications and services communicate with a Web client and detects syntactical deviations, for example caused by Cross-Site Scripting attacks.

ACKNOWLEDGMENT

We thank Philipp Winter for the helpful feedback and suggestions. This research has been supported by the Christian Doppler Society.

REFERENCES


