This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

Discovering the Top-k Unexplained Sequences in Time-Stamped Observation Data

Massimiliano Albanese, Cristian Molinaro, Fabio Persia, Antonio Picariello, V. S. Subrahmanian

Abstract—There are numerous applications where we wish to discover unexpected activities in a sequence of time-stamped observation data—for instance, we may want to detect inexplicable events in transactions at a web site or in video of an airport tarmac. In this paper, we start with a known set $\mathcal{A}$ of activities (both innocuous and dangerous) that we wish to monitor. However, in addition, we wish to identify "unexplained" subsequences in an observation sequence that are poorly explained (e.g., because they may contain occurrences of activities that have never been seen or anticipated before, i.e. they are not in $\mathcal{A}$). We formally define the probability that a sequence of observations is unexplained (*totally* or *partially*) w.r.t. $\mathcal{A}$. We develop efficient algorithms to identify the top-k *Totally* and *Partially Unexplained Sequences* w.r.t. $\mathcal{A}$. These algorithms leverage theorems that enable us to speed up the search for totally/partially unexplained sequences. We describe experiments using real-world video and cyber security datasets showing that our approach works well in practice in terms of both running time and accuracy.

Index Terms—I.2.4 Knowledge Representation Formalisms and Methods < I.2 Artificial Intelligence < I Computing Methodologies, I.2.4.d Knowledge base management < I.2.4 Knowledge Representation Formalisms and Methods < I.2 Artificial Intelligence < I Computing Methodologies

1 Introduction

Identifying unexpected activities is an important problem in a wide variety of applications such as video surveillance, cyber security, fault detection in safety critical systems, and fraud detection.

For instance, airport baggage areas are continuously monitored for suspicious activities by video surveillance. In crime-ridden neighborhoods, police often monitor streets and parking lots using video surveillance. In Israel, highways are monitored for suspicious activities by a central authority. However, all these applications search for *known* activities—activities that have been identified in advance as being either innocuous or dangerous. For instance, in the highway application, security officers may look both for normal behavior (e.g. driving along the highway in a certain speed range unless traffic is slow) as well as "suspicious" behavior (e.g. stopping the car near a bridge, taking a package out and leaving it on the side of the road before driving away).

In cyber security, intrusion detection can monitor network traffic for suspicious behavior and trigger security alerts. Alert correlation methods aggregate alerts into multi-step attack scenarios. However, both techniques rely on models encoding a priori knowledge of either normal or malicious behavior. They cannot deal with events such as "zero day" attacks that have never been seen before. In practice, all these methods are incapable of quantifying how well available models explain a sequence of events observed in an observation stream.

Figure 1 shows how our framework would work in practice. We start with a set of activity models $\mathcal{A}$ for both "good" and "bad" activities. Good activities are activities that are considered appropriate (e.g., certain permitted behaviors in an airport secure baggage zone) while bad activities are ones known to be inappropriate (e.g., a baggage handler opening a suitcase, taking items out, and putting them in a different bag). Techniques already exist to find occurrences of activities in time-stamped observation data (e.g., a video, a sequence of transactions at a website, etc.) with each occurrence having an associated probability.

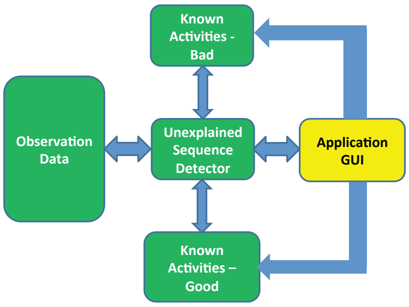

Fig. 1: Overall working of unexplained sequences

- M. Albanese is with the Department of Applied Information Technology, George Mason University, Nguyen Engineering Building, Fairfax, VA 22030. E-mail: malbanes@gmu.edu
- C. Molinaro and V. S. Subrahmanian are with the Department of Computer Science and UMIACS, University of Maryland, A.V. Williams Building, College Park, MD 20742. E-mail: {molinaro, vs}@umiacs.umd.edu
- F. Persia and A. Picariello, are with Dipartimento di Informatica e Sistemistica, Università di Napoli "Federico II", Via Claudio 21, 80125 Napoli, Italy. E-mail: {fabio.persia, picus}@unina.it

Digital Object Indentifier 10.1109/TKDE.2013.33 1041-4347/13/$31.00 © 2013 IEEE

In this paper, our goal is to find an *unexplained sequence detector*, i.e. to identify subsequences of the observation data, called *unexplained sequences*, that known models are not able to "explain" with a certain confidence. In other words, what is happening in unexplained sequences is not well captured by the available activity models in $\mathcal{A}$. Once such subsequences have been identified, they can be further analyzed, e.g., to learn new activity models from them. Or, as shown in Figure 1, each unexplained sequence can be shown to a domain expert (e.g., airport security or cyber security expert) who can then add these observed sequences or generalizations thereof to the currently known list of good or bad activities.

Unexplained sequences allow an application to identify activities never seen or imagined before by experts, and to add them to an increasing body of such knowledge. For instance, a new type of terrorist attack at an airport or a zero-day attack on a computer system, may involve sequences of actions (observations) not seen before—and hence not captured by past activity models (i.e., those in $\mathcal{A}$). In this paper, we primarily focus on the unexplained sequence detector component of Figure 1.

We achieve this via a possible-worlds based model and define the probability that a sequence of observations is *totally* (or *partially*) unexplained. Users can then look for all observation sequences that are totally (or partially) unexplained with a probability exceeding a threshold that they specify. We show important properties of our mathematical model that can be leveraged to speed up the search for unexplained sequences. We define algorithms to find top-k totally and partially unexplained sequences. We develop a prototype implementation and report on experiments using two video data sets and a cyber security dataset showing that the algorithms work well in practice, both from an efficiency perspective and an accuracy perspective.

The paper starts (Section 2) with an overview of related work. Section 3 provides basic definitions of stochastic activities slightly extending [1]. Section 4 defines the probability that a sequence is totally (or partially) unexplained. We also define the problem of finding the top-k (totally or partially) unexplained sequences and classes. Section 5 derives theorems that enable fast search for totally and partially unexplained sequences. Section 6 presents algorithms for solving the problems introduced in Section 4. Section 7 describes our experiments.

2 Related Work

We are not aware of domain-independent prior work on discovering unexplained sequences. However, specific work in the domains of video and cyber-security have focused on anomalous activity detection.

2.1 Video Analysis

A Priori Definitions. Several researchers have studied how to search for specifically defined patterns of normal/abnormal activities [2]. [3] studies how HMMs can be used to recognize complex activites, while [4] and [5] use coupled HMMs. [6] uses Dynamic Bayesian Networks (DBNs) to capture causal relationships between observations and hidden states. [1] developed a stochastic automaton based language to detect activities in video, while [7] presented an HMM-based algorithm. *In contrast, this paper starts with a set $\mathcal{A}$ of activity models (corresponding to innocuous/dangerous activities) and finds observation sequences that are not sufficiently explained by the models in $\mathcal{A}$.* Such unexplained sequences reflect activity occurrences that differ from the application's expectations.

Learning and then detecting abnormality. Several researchers first learn normal activity models and then detect abnormal/unusual events. [8] suggests a semi-supervised approach to detect abnormal events that are rare, unexpected, and relevant. We do not require "unexplained" events to either be rare or relevant. [9] uses HMMs to detect rare events, while [10] defines an anomaly as an atypical behavior pattern that is not represented by sufficient samples in a training dataset and satisfies an abnormal pattern. [11] defines abnormality as unseen or rarely occurring events—an initial video is used to learn normal behaviors. [12] shows how to detect users with abnormal activities from sensors attached to human bodies. An abnormal activity is defined as "an event that occurs rarely and has not been expected in advance". The same notion of abnormal activity is considered in [13] and [14]. [15] learns patterns of activities over time in an unsupervised way. [16] detects individual anomalies in crowd scenes—an anomaly is defined as a rare or infrequent behavior compared to all other behaviors. Common activities are accepted as normal and infrequent activity patterns are flagged as abnormal. All these approaches first learn normal activity models and then detect abnormal/unusual events. These papers differ from ours as they consider rare events to be abnormal. In contrast, we consider activities to be unexplained even if they are not rare and the available models are not able to capture them. For example, if a new way to break into cars has occurred many times (and we do not have a model for it), then we want to flag sequences where those activities occur as "unexplained" even if they are not rare. In addition, if a model exists for a rare activity, we would flag it as "explained", while many of these frameworks would not.

Similarity-based abnormality. [17] proposes an unsupervised technique in which no explicit models of normal activities are built. Each event in the video is compared with all other observed events to determine how many similar events exist. Unusual events are events for which there are no similar events in the video. Hence, this work also considers unusual activity as a rare event and a large number of observations is required to verify if an activity is unusual. [18] uses a similar approach: a scene is considered anomalous when the maximum similarity between the scene and all previously viewed scenes is below a threshold. In [19], frequently occurring patterns are normal and patterns that are dissimilar from most patterns are anomalous. [20] learns trajectory prototypes and detects anomalous behaviors when visual trajectories deviate from the learned representations of typical behaviors. An unsupervised approach, where an abnormal trajectory refers to something that has never (or rarely) seen, has been proposed in [21]. A normal trajectory is intended to be one similar enough to one or more trajectories that the system already knows. In [3], activities performed by a group of moving and interacting objects are modeled as shapes and abnormal activities are defined as a change in the shape activity model. In the context of elder care, [22] proposes an approach that first analyzes and designs features, and then detects abnormal activities using a method based on the designed features and Support Vector Data Description. [23] proposes a methodology to characterize novel scenes over long time periods without a priori knowledge. A hierarchical modeling process, characterizing an activity at multiple levels of resolution, is developed to classify and predict future activities and detect abnormal behavior.

Other relevant work. In [24], unusual events are detected by monitoring the scene with *monitors* which extract local low-level observations from the video stream. The monitor computes the likelihood of a new observation with respect to the probability distribution of prior observations. If the likelihood falls below a threshold, then the monitor outputs an alert. The local alerts issued by the monitors are then combined. [25] automatically learns high frequency events (taking spatio-temporal aspects into account) and declares them normal—events deviating from these rules are anomalies. [26] learns storylines from weakly labeled videos. A storyline includes the actions that occur in a video and their causal relationships. AND-OR graphs are used to represent storyline models.

The notion of unexplained sequences used in this paper has been proposed in [27].

2.2 Cyber Security

Intrusion detection systems (IDSs) monitor network traffic for suspicious behavior and trigger alerts [28], [29], [30]. Alert correlation methods aggregate such alerts into multi-step attacks [31], [32], [33], [34], [35], [36].

Intrusion detection. Intrusion detection techniques can be broadly classified into *signature-based* [30] and *profile-based* (or *anomaly-based*) [29] methods. A signature refers to a set of conditions that characterize intrusion activities w.r.t. packet headers and payload content. Historically, signature-based methods have been used extensively to detect malicious activities. On the other hand, in profile-based methods, a known deviation from the norm is considered anomalous (e.g. HTTP traffic on a non-standard port).

In contrast, in this paper, we consider the case where we have a set $\mathcal{A}$ of known activities (both innocuous and dangerous)—and we are looking for observation sequences that cannot be explained by either (if they were, they would constitute patterns that were known a priori). These need to be flagged as they might represent “zero day” attacks—attacks that were never seen before and vary significantly from past known access patterns.

Correlation techniques. The goal of correlation is to find causal relationships between alerts in order to reconstruct attacks from isolated alerts. The main role of correlation is to provide a higher level view of the actual attacks [33], [35], [34], [37], [38], [39].

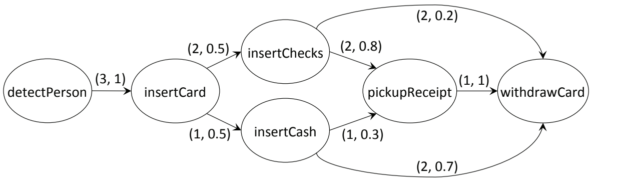

Fig. 2: Example of stochastic activity: ATM deposit

IDSs and correlation techniques rely on models encoding a priori knowledge of either normal or malicious behavior, and cannot appropriately deal with events that are not explained by the underlying models.

The framework and algorithms for identifying unexplained sequences presented in this paper are domain independent and may be applied to any domain including both activity detection in video and in cyber-security.

3 Basic Activity Model

This section extends the stochastic activity model of [1] by adding a function δ which expresses a constraint on the maximum “temporal distance” between two actions in an activity (though we make no claims of novelty for this).

We assume the existence of a finite set $\mathcal{S}$ of *action symbols*, corresponding to observable atomic actions. For instance, in the video domain, action symbols might be recognized by sophisticated image processing algorithms, while in the cyber-security domain, they may simply be read from a log file. Though our unexplained sequence detection framework is domain-independent, in some domains such as video surveillance, the problem of recognizing low-level actions in video can be a big challenge.

Definition 3.1 (Stochastic activity): A *stochastic activity* is a labeled directed graph $A = (V, E, \delta, \rho)$ where

- V is a finite set of nodes labeled with action symbols from $\mathcal{S}$;
- $E \subseteq V \times V$ is a set of edges;
- $\delta : E \to \mathbb{N}^+$ associates, with each edge $\langle v_i, v_j\rangle$, an upper bound on the time that can elapse between v_i and v_j;
- ρ is a function that associates, with each node $v \in V$ having out-degree 1 or more, a probability distribution on $\{\langle v, v'\rangle \mid \langle v, v'\rangle \in E\}$, i.e., $\sum_{\langle v,v'\rangle \in E} \rho(\langle v, v'\rangle) = 1$;
- there exists at least one *start node* in the activity definition, i.e. $\{v \in V \mid \nexists\ v' \in V\ s.t.\ \langle v', v\rangle \in E\} \neq \emptyset$;
- there exists at least one *end node* in the activity definition, i.e. $\{v \in V \mid \nexists\ v' \in V\ s.t.\ \langle v, v'\rangle \in E\} \neq \emptyset$.

Figure 2 shows a stochastic activity of deposits at an Automatic Teller Machine (ATM). Each edge e is labeled with $(\delta(e), \rho(e))$. For an edge $e = \langle s_1, s_2\rangle$, $\delta(e)$ specifies the maximum time between when s_1 is observed and when s_2 is observed. ρ specifies the probability of going from one node to another (the probability distribution