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.

associated with a node gives the transition probability from that node). For instance, the two edges starting at node insertCard mean that there is a 50% probability of going to node insertChecks and a 50% probability of going to node insertCash from node insertCard. In addition, insertChecks and insertCash must follow insertCard within 2 and 1 time units, respectively.

In general, each node of a stochastic activity definition is something that can be detected by application code. For instance, if we are tracking activities in video, each node in a stochastic activity would be something that can be detected by an image processing program, e.g. "Detect Person" in Figure 2 may be identified as holding only if a probabilistic face recognition program returns a probability over some threshold that a given frame (or block of frames) contains a face in it. Likewise, in a cybersecurity application, a node in an activity such as "Attempted login" may only be identified as occurring if a log file archives a login attempt. For the sake of simplicity, we use "high level" descriptions of nodes in our examples, as opposed to low level descriptions (e.g., the color histogram of a given image shows over 70% of the colors are a certain shade).

An instance of a stochastic activity A is a path in A from a start node to an end node.

Definition 3.2 (Stochastic activity instance): An *instance* of a stochastic activity (V, E, δ, ρ) is a sequence $\langle s_1, \ldots, s_m \rangle$ of nodes in V such that

- $\langle s_i, s_{i+1} \rangle \in E$ for $1 \leq i < m$;
- $\{s \mid \langle s, s_1 \rangle \in E\} = \emptyset$, i.e., s_1 is a start node; and
- $\{s \mid \langle s_m, s \rangle \in E\} = \emptyset$, i.e., s_m is an end node.

The probability of the instance is $\prod_{i=1}^{m-1} \rho(\langle s_i, s_{i+1} \rangle)$.

In Figure 2, ⟨detectPerson, insertCard, insertCash, withdrawCard⟩ is an instance with probability 0.35. Throughout this paper, we assume an arbitrary but fixed set $\mathcal{A}$ of stochastic activities.

The preceding definitions do not take observation sequences into account. In order to define when activity occurrences are detected in a sequence of time-stamped observation data, we first need to formally define an observation sequence. An *observation sequence* is a finite sequence of *observation IDs*. An observation ID (OID) f has an associated timestamp, denoted $f.ts$, and an associated set of action symbols, denoted $f.obs$. Without loss of generality, we assume timestamps to be positive integers. For instance, if our observation sequence is a video, then the OIDs may be frame IDs with $f.ts$ being the timestamp associated with frame f and $f.obs$ being the actions detected in frame f. On the other hand, if our observation sequence is a sequence of transactions at a website, the OIDs are transaction IDs, $f.ts$ is the timestamp associated with transaction f, and $f.obs$ are the actions associated with transaction f.

Example 3.1 (Video example): An observation sequence might be a video $v = \langle f_1, f_2, f_3, f_4, f_5 \rangle$, where the f_i's are frame IDs, $f_i.ts = i$ for $1 \leq i \leq 5$, $f_1.obs = \{\text{detectPerson}\}$, $f_2.obs = \{\text{insertCard}\}$, $f_3.obs = \{\text{insertCash}\}$, $f_4.obs = \{\text{withdrawCash}\}$, $f_5.obs = \{\text{withdrawCard}\}$. Notice that withdrawCash in frame f_4 does not appear in the stochastic activity of Figure 2. In general, action symbols may be detected in a frame even if they do not appear in the definition of a stochastic activity because it is irrelevant for that activity.

Throughout the paper, we use the following terminology and notation for (general) sequences. Suppose $S_1 = \langle a_1, \ldots, a_n \rangle$ and $S_2 = \langle b_1, \ldots, b_m \rangle$ are two sequences. S_2 is a *subsequence* of S_1 iff there exist $1 \leq j_1 < j_2 < \ldots < j_m \leq n$ s.t. $b_i = a_{j_i}$ for $1 \leq i \leq m$. If $j_i = j_{i+1} - 1$ for $1 \leq i < m$, then S_2 is a *contiguous* subsequence of S_1. We write $S_1 \cap S_2 \neq \emptyset$ iff S_1 and S_2 have a common element and write $e \in S_1$ iff e is an element appearing in S_1. The *concatenation* of S_1 and S_2, i.e., the sequence $\langle a_1, \ldots, a_n, b_1, \ldots, b_m \rangle$, is denoted by $S_1 \cdot S_2$. Finally, $|S_1|$ denotes the number of elements in S_1.

We now define an occurrence of a stochastic activity in an observation sequence.

Definition 3.3 (Activity occurrence): Let v be an observation sequence and $A = (V, E, \delta, \rho)$ a stochastic activity. An *occurrence* o of A in v is a sequence $\langle (f_1, s_1), \ldots, (f_m, s_m) \rangle$ such that

- $\langle f_1, \ldots, f_m \rangle$ is a subsequence of v,
- $\langle s_1, \ldots, s_m \rangle$ is an instance of A,
- $s_i \in f_i.obs$, for $1 \leq i \leq m$, and [1]
- $f_{i+1}.ts - f_i.ts \leq \delta(\langle s_i, s_{i+1} \rangle)$, for $1 \leq i < m$.

The probability of o, denoted $p(o)$, is the probability of the instance $\langle s_1, \ldots, s_m \rangle$.

When concurrently monitoring multiple activities, shorter activity instances generally tend to have higher probability. To remedy this, we normalize occurrence probabilities by introducing the relative probability $p^*(o)$ of an occurrence o of activity A as $p^*(o) = \frac{p(o)}{p_{max}}$, where p_{max} is the highest probability of any instance of A.

Example 3.2 (Video example): Consider the video of Example 3.1. An occurrence of the activity of Figure 2 is $o = \langle (f_1, \text{detectPerson}), (f_2, \text{insertCard}), (f_3, \text{insertCash}), (f_5, \text{withdrawCard}) \rangle$, and $p^*(o) = 0.875$. Notice that if the edge going from insertCash to withdrawCard was labeled with 1 by δ, then o would not have been an activity occurrence because withdrawCard was required to follow insertCash within at most 1 time unit, whereas it occurs after 2 time units in the video.

We use $\mathcal{O}(v)$ to denote the set of all activity occurrences in v. Whenever v is clear from the context, we write $\mathcal{O}$ instead of $\mathcal{O}(v)$.

The next section describes our framework for discovering unexplained sequences in an application-independent manner. It is worth noting that the actual input of the framework consists of an observation sequence and a set of activity occurrences (each with a probability). Though our framework is domain-independent, there can be challenges in providing the observations associated with OIDs in some domains (e.g. video surveillance, where identifying the low level actions in a video frame can be highly non-trivial).

1. With a slight abuse of notation, we use s_i to refer to both node s_i and the action symbol labeling it.

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.

4 Unexplained Sequence Probability Model

This section defines the probability that an observation sequence is unexplained by $\mathcal{A}$. We note that the occurrence of an activity in an observation sequence can involve conflicts. For instance, consider the activity occurrence o in Example 3.2 and suppose there is a second activity occurrence o' such that $(f_1, \mathsf{detectPerson}) \in o'$. In this case, there is an implicit conflict because $(f_1, \mathsf{detectPerson})$ belongs to both occurrences, but in fact, detectPerson can only belong to one activity occurrence, i.e. though o and o' may both have a non-zero probability, the probability that these two activity occurrences coexist is 0. Formally, we say two activity occurrences o, o' *conflict*, denoted $o \nsim o'$, iff $o \cap o' \neq \emptyset$. We now use this to define possible worlds.

Definition 4.1 (Possible world): Let $\mathcal{O}$ be the set of all activity occurrences in an observation sequence v. A *possible world* for v is a subset w of $\mathcal{O}$ s.t. $\nexists o_i, o_j \in w, o_i \nsim o_j$.

Thus, a possible world is a set of activity occurrences which do not conflict with one another, i.e., an action symbol in an OID cannot belong to two distinct activity occurrences in the same world. We use $\mathcal{W}(v)$ to denote the set of all possible worlds for an observation sequence v; whenever v is clear from the context, we simply write $\mathcal{W}$.

Example 4.1 (Video example): Consider a video with two conflicting occurrences o_1, o_2. There are 3 possible worlds: $w_0 = \emptyset$, $w_1 = \{o_1\}$, and $w_2 = \{o_2\}$. Note that $\{o_1, o_2\}$ is not a world as $o_1 \nsim o_2$. Each world represents a way of explaining what is observed. The first world corresponds to the case where nothing is explained, the second and third worlds correspond to the scenarios where we use one of the two possible occurrences to explain the observed action symbols.

Note that any subset of $\mathcal{O}$ not containing conflicting occurrences is a legitimate possible world—possible worlds are not required to be maximal w.r.t. $\subseteq$. In the above example, the empty set is a possible world even though there are two other possible worlds $w_1 = \{o_1\}$ and $w_2 = \{o_2\}$ which are supersets of it. The reason is that o_1 and o_2 are uncertain, so the scenario where neither o_1 nor o_2 occurs is a legitimate one. We illustrate this below.

Example 4.2 (Video example): Suppose we have a video where a single occurrence o has $p^*(o) = 0.6$. In this case, it is natural to say that there are two possible worlds $w_0 = \emptyset$ and $w_1 = \{o\}$ and expect the probabilities of w_0 and w_1 to be 0.4 and 0.6, respectively. By restricting ourselves to maximal possible worlds only, we would have only one possible world, w_1, whose probability is 1, which is wrong.

It is worth noting that the problem of finding possible worlds corresponds to the problem of finding the independent sets of a graph: occurrences are vertices, conflicts are edges, possible worlds are independent sets. Thus, algorithms to find maximal independent sets can be directly applied to compute possible worlds—all possible worlds can be simply obtained by taking all subsets of the maximal independent sets. An efficient algorithm for generating all the maximal independent sets has been proposed in [40], processing time and memory space are bounded by $O(nm\mu)$ and $O(n+m)$, respectively, where n, m, and μ are the numbers of vertices (occurrences in our case), edges (conflicts in our case), and maximal independent sets (possible worlds in our case) of a graph.

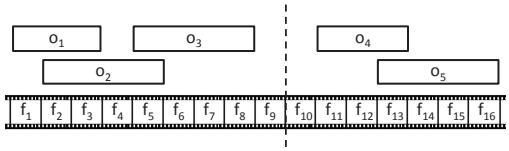

Fig. 3: Conflict-Based Partitioning of a video

We use $\overset{*}{\nsim}$ to denote the transitive closure of $\nsim$. Clearly, $\overset{*}{\nsim}$ is an equivalence relation and determines a partition of $\mathcal{O}$ into equivalence classes $\mathcal{O}_1, \ldots, \mathcal{O}_m$. Here the basic idea is to partition the observation sequence into subsequences containing occurrences that conflict directly or in a "transitive" way (we will formally define this with the notion of a *Conflict-Based Partitioning* in Definition 4.2 and illustrate it in Example 4.4). Equivalence classes that temporally overlap are collapsed into a single one.[2]

Example 4.3 (Video example): Suppose we have a video $v = \langle f_1, \ldots, f_{16} \rangle$ s.t. five occurrences o_1, o_2, o_3, o_4, o_5 are detected as depicted in Figure 3, that is, $o_1 \nsim o_2$, $o_2 \nsim o_3$, and $o_4 \nsim o_5$. There are two equivalence classes determined by $\overset{*}{\nsim}$, namely $\mathcal{O}_1 = \{o_1, o_2, o_3\}$ and $\mathcal{O}_2 = \{o_4, o_5\}$.

The equivalence classes determined by $\overset{*}{\nsim}$ lead to a conflict-based partitioning of an observation sequence.

Definition 4.2 (Conflict-Based Partitioning): Let v be an observation sequence and $\mathcal{O}_1, \ldots, \mathcal{O}_m$ the equivalence classes determined by $\overset{*}{\nsim}$. A *Conflict-Based Partitioning* (CBP) of v is a sequence $\langle v_1, \ldots, v_m \rangle$ such that:

- $v_1 \cdot \ldots \cdot v_m = v$, and
- $\mathcal{O}(v_i) = \mathcal{O}_i$, for $1 \leq i \leq m$.

The v_i's are called *segments*.

Example 4.4 (Video example): A CBP of the video in Example 4.3 is $\langle v_1, v_2 \rangle$, where $v_1 = \langle f_1, \ldots, f_9 \rangle$ and $v_2 = \langle f_{10}, \ldots, f_{16} \rangle$. Another partitioning of the same video is the one where $v_1 = \langle f_1, \ldots, f_{10} \rangle$ and $v_2 = \langle f_{11}, \ldots, f_{16} \rangle$.

Thus, activity occurrences determine a set of possible worlds (different ways of explaining an observation sequence). We wish to find a probability distribution over all possible worlds that (i) is consistent with the relative probabilities of the occurrences, and (ii) takes conflicts into account. We assume the user specifies a function $Weight : \mathcal{A} \rightarrow \mathbb{R}^+$ which assigns a weight to each activity and prioritizes the importance of the activity.[3] The weight

2. Two equivalence classes $\mathcal{O}_i$ and $\mathcal{O}_j$ temporally overlap iff $[min(\mathcal{O}_i), max(\mathcal{O}_i)] \cap [min(\mathcal{O}_j), max(\mathcal{O}_j)] \neq 0$, where $min(\mathcal{O}_i) = \min\{f.ts \mid \exists o \in \mathcal{O}_i, (f,s) \in o\}$, $max(\mathcal{O}_i) = \max\{f.ts \mid \exists o \in \mathcal{O}_i, (f,s) \in o\}$, and $min(\mathcal{O}_j)$, $max(\mathcal{O}_j)$ are analogously defined.

3. For instance, highly threatening activities may be assigned a high weight.

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.

of an occurrence o of activity A is the weight of A. We use $C(o)$ to denote the set of occurrences conflicting with o, i.e., $C(o) = \{o' \mid o' \in \mathcal{O} \wedge o' \nsim o\}$. Note that $o \in C(o)$; and $C(o) = \{o\}$ when o does not conflict with any other occurrence. Finally, we assume that activity occurrences belonging to different segments are independent events. Suppose p_w denotes the (unknown) probability of world w. As we know the probability of occurrences, and as each occurrence occurs in certain worlds, we can induce a set of nonlinear constraints that will subsequently be used to learn the values of the p_w's.

Definition 4.3: Let v be an observation sequence and $\mathcal{O}_1, \ldots, \mathcal{O}_m$ the equivalence classes determined by $\overset{*}{\nsim}$. We define the non-linear constraints $NLC(v)$ as follows:

$$\begin{cases} p_w \geq 0, \quad \forall w \in \mathcal{W} \\ \sum_{w \in \mathcal{W}} p_w = 1 \\ \sum_{w \in \mathcal{W} \, s.t. \, o \in w} p_w = p^*(o) \cdot \frac{Weight(o)}{\sum_{o_j \in C(o)} Weight(o_j)}, \forall o \in \mathcal{O} \\ p_w = \prod_{k=1}^{m} \sum_{w' \in \mathcal{W} \, s.t. \, w' \cap \mathcal{O}_k = w \cap \mathcal{O}_k} p_{w'} \qquad \forall w \in \mathcal{W} \end{cases}$$

The first two types of constraints enforce a probability distribution over the set of possible worlds. The third type of constraint ensures that the probability of occurrence o—which is the sum of the probabilities of the worlds containing o—is equal to its relative probability $p^*(o)$ weighted by $\frac{Weight(o)}{\sum_{o_j \in C(o)} Weight(o_j)}$. Note that: (i) the value on the right-hand side of the third type of constraint decreases as the amount of conflict increases, (ii) if an occurrence o is not conflicting with any other occurrence, then its probability $\sum_{w \in \mathcal{W} \, s.t. \, o \in w} p_w$ is equal to $p^*(o)$, i.e., the probability returned by the stochastic automaton. The last kind of constraint reflects independence between segments. In general $NLC(v)$ might admit multiple solutions.

Example 4.5 (Video example): Consider a single-segment video consisting of frames $f_1, \ldots, f_9$ (cf. Figure 3). Suppose o_1, o_2, o_3 have been detected with relative probabilities 0.3, 0.6, and 0.5, respectively. Suppose the weights of o_1, o_2, o_3 are 1, 2, 3, respectively. Five worlds are possible: $w_0 = \emptyset$, $w_1 = \{o_1\}$, $w_2 = \{o_2\}$, $w_3 = \{o_3\}$, and $w_4 = \{o_1, o_3\}$. Then, $NLC(v)$ is:[4]

$$\begin{array}{l} p_i \geq 0 \qquad 0 \leq i \leq 4 \\ p_0 + p_1 + p_2 + p_3 + p_4 = 1 \\ p_1 + p_4 = 0.3 \cdot \frac{1}{3} \\ p_2 = 0.6 \cdot \frac{1}{3} \\ p_3 + p_4 = 0.5 \cdot \frac{3}{5} \end{array}$$

which has multiple solutions. One solution is $p_0 = 0.4$, $p_1 = 0.1$, $p_2 = 0.2$, $p_3 = 0.3$, $p_4 = 0$. Another solution is $p_0 = 0.5$, $p_1 = 0$, $p_2 = 0.2$, $p_3 = 0.2$, $p_4 = 0.1$.

In the rest of the paper, we assume that $NLC(v)$ is solvable.[5] We say that a sequence $S = \langle (f_1, s_1), \ldots, (f_n, s_n) \rangle$ *occurs in* an observation sequence v iff $\langle f_1, \ldots, f_n \rangle$ is a contiguous subsequence of v and $s_i \in f_i.obs$ for $1 \leq i \leq n$. We give two semantics for S to be unexplained in a world $w \in \mathcal{W}$. Intuitively, S is *totally* (resp. *partially*) unexplained in w iff w does not explain every (resp. at least one) symbol of S. More formally:

1) S is *totally unexplained* in w, denoted $w \nvDash_T S$, iff $\forall (f_i, s_i) \in S, \nexists o \in w, (f_i, s_i) \in o$;
2) S is *partially unexplained* in w, denoted $w \nvDash_P S$, iff $\exists (f_i, s_i) \in S, \nexists o \in w, (f_i, s_i) \in o$.

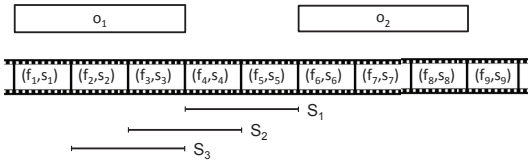

Fig. 4: Totally and partially unexplained sequences

Example 4.6 (Video example): Suppose we have a video $v = \langle f_1, \ldots, f_9 \rangle$ such that $f_i.obs = \{s_i\}$, $1 \leq i \leq 9$, and two occurrences o_1 and o_2 are detected (cf. Figure 4). The four possible worlds are: $w_0 = \emptyset$, $w_1 = \{o_1\}$, $w_2 = \{o_2\}$, $w_3 = \{o_1, o_2\}$. Let $S_1 = \langle (f_4, s_4), (f_5, s_5) \rangle$, $S_2 = \langle (f_3, s_3), (f_4, s_4) \rangle$, $S_3 = \langle (f_2, s_2), (f_3, s_3) \rangle$ be sequences occurring in v. S_1 is totally (and partially) unexplained in every world. S_2 is totally unexplained in w_0 and w_2 but not in w_1 and w_3; moreover, S_2 is partially unexplained in every world. S_3 is totally and partially unexplained in w_0 and w_2 but not in w_1 and w_3.

We now define the probability of a sequence in an observation sequence being totally/partially unexplained.

Definition 4.4: Let S be a sequence occurring in an observation sequence v. The probability interval that S is totally unexplained in v is $\mathcal{I}_T(S) = [l, u]$, where:

$$\begin{array}{c} l = \textbf{minimize} \sum_{w \in \mathcal{W} \; s.t. \; w \nvDash_T S} p_w \\ \textbf{subject to } NLC(v) \\ u = \textbf{maximize} \sum_{w \in \mathcal{W} \; s.t. \; w \nvDash_T S} p_w \\ \textbf{subject to } NLC(v) \end{array}$$

The probability interval that S is partially unexplained in v is $\mathcal{I}_P(S) = [l', u']$, where l', u' are derived in exactly the same way as l, u above by replacing the $\nvDash_T$ symbols in the above optimization problems by $\nvDash_P$.

Thus, the probability that a sequence S occurring in v is totally (resp. partially) unexplained w.r.t. a solution of $NLC(v)$ is the sum of the probabilities of the worlds in which S is totally (resp. partially) unexplained. As $NLC(v)$ may have multiple solutions, we find the tightest interval $[l, u]$ (resp. $[l', u']$) containing this probability for any solution. Different criteria can be used to infer a value from an interval $[l, u]$, e.g. the MIN l, the MAX u, the average (i.e., $(l + u)/2$), etc. The only requirement is that this value has to be in $[l, u]$. We henceforth assume that such criterion has been chosen—$\mathcal{P}_T(S)$ (resp. $\mathcal{P}_P(S)$) denotes the probability that S is totally (resp. partially) unexplained.

The following proposition says that the probability that a sequence is totally (resp. partially) unexplained is no higher

4. For brevity, we do not explicitly list the independence constraints.

5. This can be easily checked via both a non-linear constraint solver, as well as methods developed in the next section.