Computational Intelligence, Volume 16, Number 3, 2000

TEMPORAL REASONING AND BAYESIAN NETWORKS

AHMED Y. TAWFIK
University of Prince Edward Island, Charlottetown, PE, C1A 4P3 Canada

ERIC M. NEUFELD
University of Saskatchewan, Saskatoon, SK, S7N 5A9 Canada

This work examines important issues in probabilistic temporal representation and reasoning using Bayesian networks (also known as belief networks). The representation proposed here utilizes temporal (or dynamic) probabilities to represent facts, events, and the effects of events. The architecture of a belief network may change with time to indicate a different causal context. Probability variations with time capture temporal properties such as persistence and causation. They also capture event interaction, and when the interaction between events follows known models such as the competing risks model, the additive model, or the dominating event model, the net effect of many interacting events on the temporal probabilities can be calculated efficiently. This representation of reasoning also exploits the notion of temporal degeneration of relevance due to information obsolescence to improve the efficiency.

Key words: temporal representation and reasoning, uncertain reasoning, Bayesian (belief) networks, models of interaction.

1. INTRODUCTION

Recent efforts to introduce temporality into Bayesian networks have resulted in a variety of networks intended primarily for applications such as planning, diagnosis, forecasting, and scheduling. Dynamic nets (Dagum, Galper, and Horvitz 1992) use an instantiation of the network for each time point, with the different instantiations linked by edges representing persistence and causation. Temporal Bayes nets (Dean and Kanazawa 1989) use survival functions to represent persistence. The underlying time model is discrete and each time point corresponds to a copy of the network. Arcs linking two copies propagate the effects of previous states and observations. Networks of dates (Berzuini 1990) represent a departure from the multiple instantiations approach because each temporal duration is represented by a node. Berzuini associates a probability density with each temporal random variable to represent continuous time. Time nets (Kanazawa 1992) define a network model that uses continuous time and extends the "networks of dates" by introducing a representation for facts (or fluents). The dHugin time-sliced Bayesian nets (Kjaerulff 1995) are based on the multiple-instantiation approach (each time slice corresponds to a copy of the network) similar to temporal Bayes nets and dynamic nets but the reasoning is based on a dynamic version of Hugin and a smoothing operator is used to approximate the effect of temporally distant occurrences. The kappa calculus approximation of probability functions and the representation of persistence through suppressors are two features introduced in action networks (Darwiche and Goldszmidt 1994). Action networks use different instantiations of the original network for different time points. A time-sliced Bayes nets generation algorithm (Ngo, Haddawy, and Helwig 1995) optimizes the network to answer a query efficiently.

Despite these efforts, there is no consensus on several issues such as when to duplicate the network, how to represent instantaneous effects, what conclusions can be made regarding the time interval between two instantiations, and how to represent continuous

Address correspondence to A. Y. Tawfik at the University of Prince Edward Island, Charlottetown, PE, C1A 4P3 Canada; e-mail: atawfik@upei.ca

time. A host of issues have not been considered yet within this framework. This work tries to answer some questions in uncertain temporal reasoning using a probabilistic representation. This representation allows the use of discrete time, continuous time, hybrid time, and other forms such as counterbased timing. Moreover, it is not based on multiple instantiations of a probabilistic network, thus avoiding the instantiation-related problems. This treatment considers a single belief network. The architecture of the network is only modified when the causal context changes. Changes in conditional probabilities over time reflect changes in causal influences. For example, recent information has a stronger effect on current beliefs. We define a probabilistic relevance criterion we call *extraneousness* that allows us to consider only a limited time interval when making inference at a particular instant.

The extraneousness criterion extends the notion of independence as it also weeds out weak dependencies. Exploiting this notion yields smaller, more manageable, problems. In addition to extending the notion of probabilistic independence, we also present a set of models of event interaction that help reduce the complexity of knowledge acquisition and inference. Those models are the temporal equivalent to static models of interaction (e.g., noisy-or).

The remainder of this section introduces terminology and an example used throughout the paper. Section 2 shows how to use dynamic probabilities to represent and reason about fluents. Section 3 examines the degeneration of information relevance under uncertainty. Section 4 addresses the representation of events and reasoning about them. Section 5 provides a causal characterization for common models of event interactions. Section 6 is devoted to some issues in reasoning about the past and the future. Sections 7 and 8 conclude the paper with a discussion of related work and issues.

Throughout this paper, we use the network shown in Figure 1 (from Charniak 1991), which represents the statements: *"When the family goes out they turn on the outdoor light and put the dog in the backyard. The dog's barking is heard when it is out in the backyard."* This network has four binary random variables: *family-out, dog-out, light-on, and hear-bark*, which henceforth are sometimes abbreviated *fo*, *do*, *lo*, and *hb*.

1.1. Terminology

An *event* here is an occurrence having subsequent effects. An event may be an action or an observed change in one or more states. An event can have null duration or it can happen continuously over an interval of time, producing effects during some subintervals. "It started raining," "It has been raining for five minutes now,"

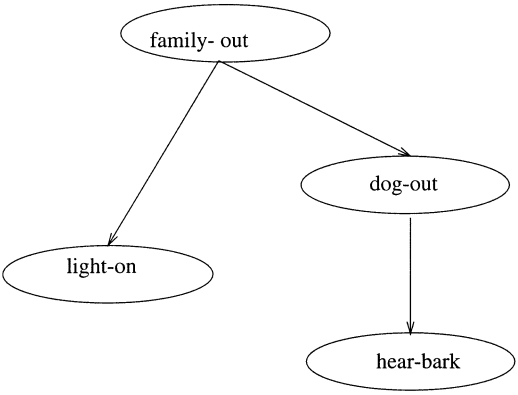


FIGURE 1. Bayesian network.

or “I opened my umbrella” are events. An *event type* is a class of events with similar effects such as “starting to rain.” An *event token* is a particular instance of the event class, such as “It started to rain today at 10:05 a.m.” The time of occurrence of an event can be deterministic (if known or directly observed) or probabilistic (if expressed as a probability distribution along the time line). “It should have started raining between 10:00 a.m. and 10:15 a.m.” is represented by a distribution. The total probability associated with an event token is at most one. A total probability less than one indicates that we are not certain whether the event did (or will) actually happen.

A *fluent* is a proposition that changes its truth value. Each fluent is either true or false at any instant. The truth value of a fluent over an interval is represented by a continuous probability distribution. The truth value over subintervals of this interval can be deduced from the distribution. “The house is red” is a fluent.

Dynamic probabilities are functions providing a mapping from the set of fluents at time t to real numbers in the unit interval. $P_t(X)$ and $P_X(t)$ are used to indicate the probability of fluent X at time t.

2. REPRESENTATION OF FLUENTS

Probabilities can have discrete or continuous distributions. Using discrete probability distributions to represent discrete time and continuous distributions to represent continuous time adds more uniformity and flexibility to the knowledge representation. The choice of a method for associating probabilities with time affects the realizability of this objective. For example, approaches that instantiate a network for different time points such as dynamic nets, temporal Bayes nets, and dHugin time-slice nets, cannot be used for continuous time. The networks of dates represent time intervals as network variables. These networks can therefore represent both continuous and discrete times.[1] An extension of this model (Kanazawa 1992) allows a fact to hold over a duration defined by two end points but the resulting time nets do not provide a uniform interface for continuous and discrete times.

Consider the problem of adding the following statements to the *family-out* example introduced earlier:

1. If the family is out between 6:00 a.m. and 6:00 p.m. they do not turn the light on.
2. If the family is not out during that period they open some windows.

Representing *before-six* (bs) as a random variable, continuous or discrete, is therefore a possible solution. Unfortunately, this complicates the network by increasing the number of nodes. A further complication is that the probabilities of *window-open* (wo) and light-on (lo) depend on the joint probability of bs and fo. The representation may get even more cumbersome if we try to represent the following additional statements:

3. Usually the family is out from 9:00 a.m. until 5:00 p.m.
4. Sometimes they come home for lunch between 12:00 noon and 1:00 p.m.
5. When they come home for lunch they do not bring the dog in.
6. They visit friends between 7:00 p.m. and 11:00 p.m.

Now, three more temporal variables are needed. It is necessary to define the appropriate joint probabilities. Different properties of temporal ordering (e.g., that five o’clock comes before six o’clock, that it cannot be five and six o’clock at the same time, . . . , etc.)

[1] Berzuini (1990) does not mention this, however it seems possible if continuous time is represented by continuous random variables and discrete time is represented by discrete ones.

will have to be encoded in the network. Alternatively, we can use the probabilistic function of time to represent the above statements.

Observation 1. Dynamic probabilities can represent fluents.

Rationale. Since a fluent has a single truth value (true or false) at any time, then a function of time can represent the truth value of a fluent. Under uncertainty, the truth value is given as a probability instead of true or false. The probability has a single value at a time and can therefore be represented as a function of time.

Probabilistic reasoning using multiple instantiations of a static network expresses the probabilities at time t in terms of probabilities at time $t-\Delta$ as recurrence relations. The dynamic probabilities are a more explicit, nonrecurrent representation of the same patterns. Figure 2 illustrates the variation of probabilities over a single day. The period "a day" is a *complete cycle* after which probabilities follow the same pattern repeatedly. Cyclic probability patterns capture the cyclic property of time useful in many applications. If the problem does not exhibit this cyclic property, probabilities are expressed over a *window of interest*.

Some economy is provided because not every probability is time dependent. For example, it is reasonable to assume $P(do|hb)$, $P(do|\neg hb)$, and $P(lo|\neg fo)$ are time independent.

Given this representation, reasoning to answer questions of the form *"What is happening at time t?"* is straightforward, but the question *"When does x happen?"* is more difficult. However, in most applications a direct probabilistic answer for the first question at different time points approximates the probability distribution for the answer to the second, hence giving the required answer.

3. DEGENERATION OF RELEVANCE

Depending on the dynamic nature of a system, two time intervals (or points) are either related or extraneous. In general, the relevance of the knowledge K at time t_j to the fluent f at time t_i degenerates as the duration separating them gets longer.[2] If the maximum effect of k_j on f_i is less than a small value δ, then t_i and t_j are temporally extraneous with respect to f. Temporal extraneousness is a relaxation of probabilistic independence, indicating that the effect of K_j on f_i is negligible.

Returning to our example, suppose an observer goes past the house every few days, instantaneously looks at the lights and checks if the dog is barking. This observer can use the notion of temporal extraneousness to rely on current observations only and to reach a conclusion about *family-out*. It is reasonable to consider, in this case, that each observation is extraneous to the others given that these observations are separated by a sufficiently long interval. On the other hand, if the observer stays for a few hours to continuously observe the light and listen to the dog, then it should be possible to relate what happens at one instant with previous and future instants.

In this section, we examine extraneousness more formally, then we turn our attention to related times and introduce a representation of events compatible with the representation of fluents developed so far.

[2] The notation X_i means X is true at time t_i.

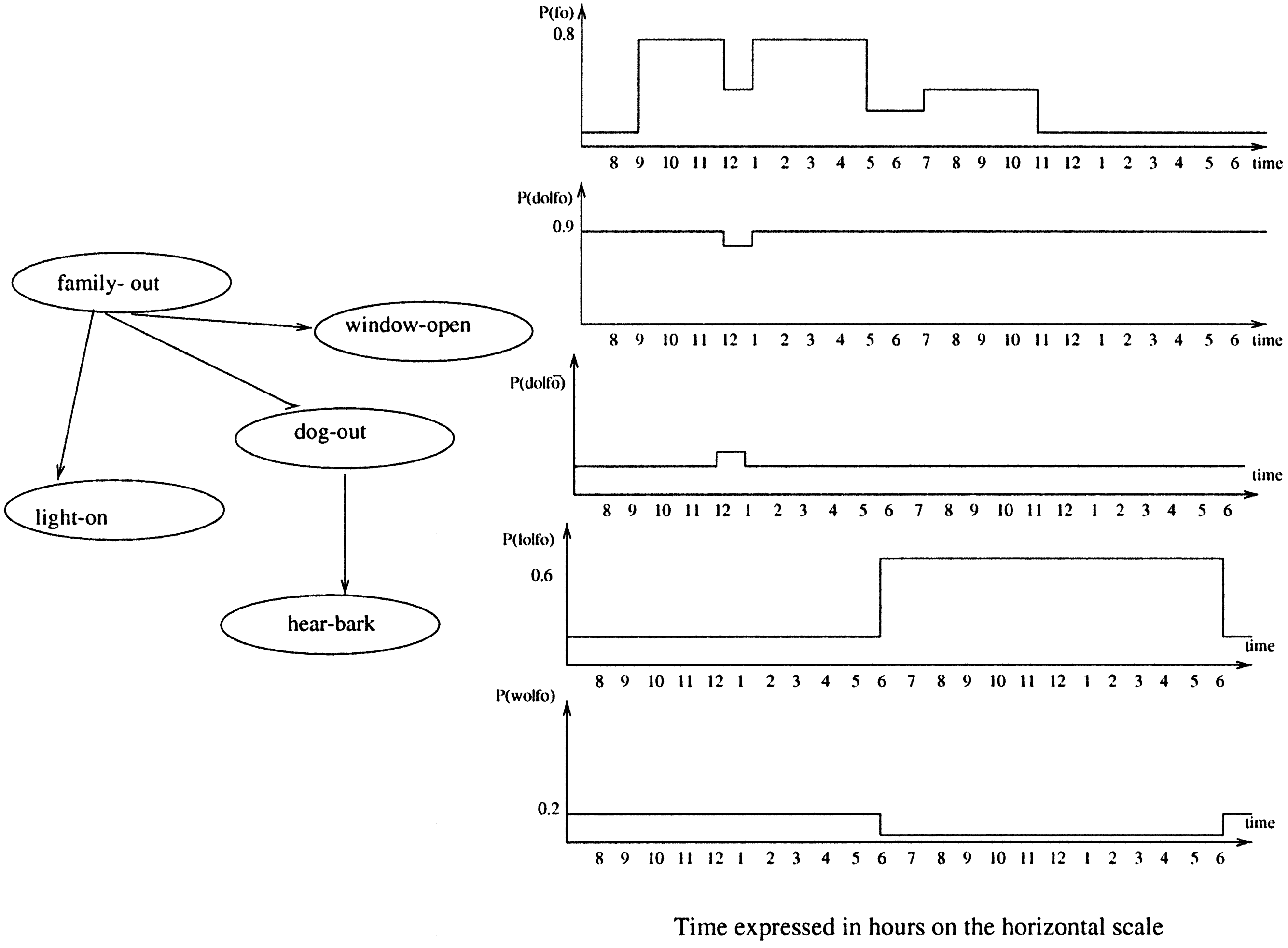


Figure 2. Probabilities as functions of time.

3.1. Extraneous Time Instants or Periods

The extraneousness assumption implies observations made at t_i do not affect conclusions at t_j in any significant way. An observer who both sees the light and hears the dog barking can use the probabilities in Figure 2 to evaluate the probability of *family-out*. Any new observations and conclusions are almost independent of previous ones because changes could have happened between the two time points. To an instantaneous observer, temporal reasoning is therefore a set of nontemporal Bayes nets except that the probabilities must correspond to the time point under consideration. Reasoning is not any harder than in the nontemporal case.

The following theorem shows that there exists a duration T such that the probability of a fluent f at a time $t > t_0 + T$ changes by a small factor δ depending on the truth (or falsity) of f at t_0. Moreover, for any choice of an arbitrarily small δ a different duration T can be found.

Without loss of generality, we prove the property assuming a discrete-time Markov chain and a single fluent. Possible generalizations are discussed afterward.

Theorem 1. Consider a fluent f represented by a Markov process with states f_i and $\bar{f}_i$ and transition probabilities $P(\bar{f}_{i+1}|f_i) = p_1$, $P(f_{i+1}|\bar{f}_i) = p_2$, $P(f_{i+1}|f_i) = 1 - p_1$, and $P(\bar{f}_{i+1}|\bar{f}_i) = 1 - p_2$ such that $0 < p_1, p_2 < 1$. If the system is in state f_i, then the fluent is true at time i. Let the probability that the system described by this Markov process be in state f_t at time t be $P(f_t)$. The claim here is that for any $\delta \ll 1$ there exists T such that

$$\forall t \geq T, |P(f_t|f_0) - P(f_t|\bar{f}_0)| < \delta.$$

Proof.

$$P(f_t|f_0) = P(f_t|f_{t-1})P(f_{t-1}|f_0) + P(f_t|\bar{f}_{t-1})P(\bar{f}_{t-1}|f_0)$$

$$P(f_t|f_0) = (1 - p_1)P(f_{t-1}|f_0) + p_2 P(\bar{f}_{t-1}|f_0).$$

But $P(\bar{f}_{t-1}|f_0) = 1 - P(f_{t-1}|f_0)$.

$$P(f_t|f_0) = (1 - p_1)P(f_{t-1}|f_0) + p_2(1 - P(f_{t-1}|f_0))$$

$$P(f_t|f_0) = (1 - p_1 - p_2)P(f_{t-1}|f_0) + p_2.$$

This is a recurrence relation that can be solved using the iteration method to get

$$\begin{aligned} P(f_t|f_0) = {} & (1 - p_1 - p_2)^{t-1}P(f_1|f_0) + p_2(1 - p_1 - p_2)^{t-2} \\ & + p_2(1 - p_1 - p_2)^{t-3} + \cdots + p_2(1 - p_1 - p_2) + p_2. \end{aligned}$$

By substitution for $P(f_1|f_0)$ and summing the geometric series in the above expression, we get

$$P(f_t|f_0) = p_1(1 - p_1 - p_2)^{t-1}\left[\frac{1}{p_1 + p_2} - 1\right] + \frac{p_2}{p_1 + p_2}.$$

Similarly,

$$\begin{aligned} P(f_t|\bar{f}_0) = {} & (1 - p_1 - p_2)^{t-1}P(f_1|\bar{f}_0) + p_2(1 - p_1 - p_2)^{t-2} \\ & + p_2(1 - p_1 - p_2)^{t-3} + \cdots + p_2(1 - p_1 - p_2) + p_2 \end{aligned}$$

$$P(f_t|\bar{f}_0) = p_2(1 - p_1 - p_2)^{t-1}\left[1 - \frac{1}{p_1 + p_2}\right] + \frac{p_2}{p_1 + p_2}.$$

Therefore,

$$|P(f_t|f_0) - P(f_t|\bar{f}_0)| = |(1 - p_1 - p_2)^t|.$$

Depending on the values of p_1, p_2, and δ, the duration T that makes the difference less than δ can be determined as follows:

$$T = \frac{\log(\delta)}{\log|1 - p_1 - p_2|}. \qquad \square$$

This theorem quantifies the intuitive notion of information obsolescence. Information about a given time point may not help in reasoning about another time point if the interval between the two points is long enough. The information is outdated by the dynamic nature of the system and the lack of knowledge about the developments occurring between the two points. Ignoring such outdated information affects the prediction within a small range δ.

The probabilities p_1 and p_2 determine the rate of change in the system as well as the duration T beyond which we can consider the observations extraneous.

In the special case $p_1 = 0$ and $p_2 = 0$ the system maintains an initial truth value. In this case no finite T can be found because the denominator $\log(1 - p_1 - p_2)$ becomes zero. If one of the states is an absorption state (i.e., if either $p_1 = 0$ or $p_2 = 0$) the system would remain in this state once it reaches it and its behavior is fully determined henceforth. The other extreme case occurs when $p_1 = 1$ and $p_2 = 1$, the system oscillates and is always predictable.

Figure 3 shows the change of the time T for different values of the sum of transition probabilities $p_1 + p_2$ for $\delta = 0.01$. It is worth noting that this curve is symmetric around the point $p_1 + p_2 = 1$. At this point the difference $P(f_T|f_0) - P(f_T|\bar{f}_0)$ is equal to zero for any t and the truth of f at any time is independent from its previous states.

3.2. Generalizing the Theorem

The proof of the theorem assumes a stationary process (i.e., time invariant transition probabilities), but the same analysis applies to time varying chains. For example, an upper bound on T may be obtained by considering the lowest possible value for $p_1 + p_2$ if this sum is less than unity, or the highest possible value for the sum greater than unity.

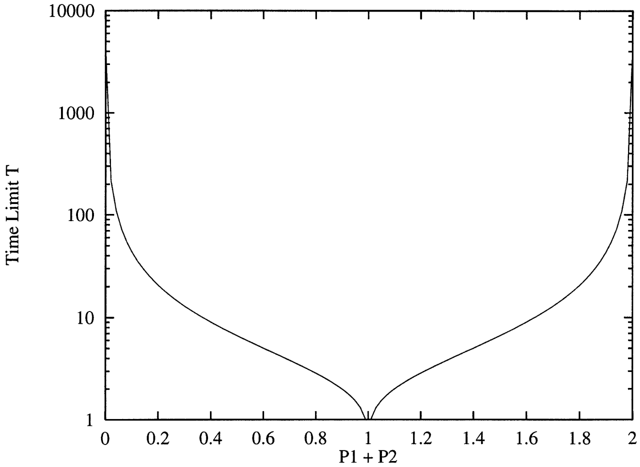


FIGURE 3. Extraneousness time versus sum of probabilities.

It is possible to generalize the proof to nonbinary variables; for a variable with n truth values, the proof proceeds in a similar fashion except for the number of probabilities, as n^2 transition probabilities are needed. The degree of extraneousness can be defined in the following definition:

Definition 1. The degree of extraneousness of Θ_j with respect to fluent f_l is δ if, for all possible initial states of Θ_j, the maximum change in the probability $P(f_l|\Theta_j)$ is less than δ.

This definition avoids the disjunctive factors problem (Hitchcock 1993) that occurs whenever a factor is compared to a disjunction of other factors. Here, for example, the probability $P(f_l|\Theta_j)$ would have to be compared to the probability $P(f_l|\neg\Theta_j)$, but $\neg\Theta_j$ is a disjunction of $n - 1$ possible states at time j. The probability of $P(f_l|\neg\Theta_j) = \sum_{S_i \neq \Theta} P(S_{i_j})P(f_l|S_{i_j}) / \sum_{S_i \neq \Theta} S_{i_j}$. The definition avoids the problem by comparing pairs of assignments instead of a singleton and a disjunction.

The convergence property of Markov chains is closely related to the above theorem. Kemeny and Snell (1976) studied the convergence of regular chains and showed that dependence on initial state decays. The upper bound on the difference in probabilities due to initial state is $(1 - 2\epsilon)^n$, where ϵ is the smallest transition probability, and n is the time. Recent work on convergence time shows that the convergence time T and its distribution can be bounded by $K \cdot \lambda^t$, where λ is the absolute value of the second highest among the eigenvalues of the transition matrix. In many cases, the convergence of Markov chains exhibits a cutoff behavior (Diaconis 1996); after an initial period of seemingly little change, the probabilities converge quickly to the steady state.

3.3. Extraneousness in Belief Network

The above proof applies directly to a time-sliced belief network. In fact, we could have done the proof directly assuming a time-sliced belief network. By definition the probabilities are given by $p_1 = P(\bar{f}_{t+\Delta}|f_t)$ and $p_2 = P(f_{t+\Delta}|\bar{f}_t)$. For example, Figure 4 shows a time-sliced Bayesian network that can be used to calculate p_1 and p_2 for a fluent f. This fluent is affected by two possible events, e_1 and e_2, where e_1 causes f to

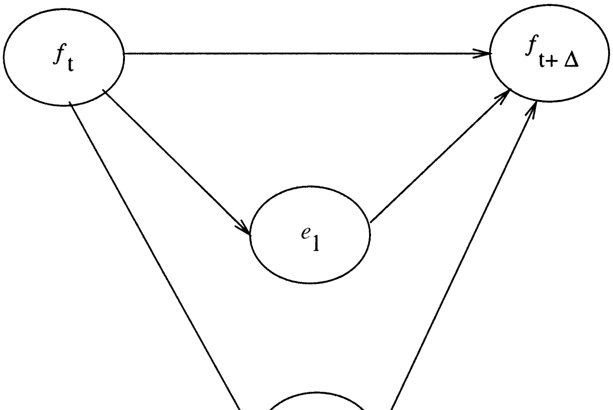


FIGURE 4. Generic time-sliced Bayesian network.

be true and e_2 turns it false. In this case,

$$p_1 = 1 - P(f_{t+\Delta}|f_t)$$

$$p_1 = 1 - \sum_{i=1}^{4} P(f_{t+\Delta}|E_i, f_t)P(E_i|f_t),$$

where $E_1 \equiv (e_1 \wedge e_2)$, $E_2 \equiv (e_1 \wedge \bar{e}_2)$, $E_3 \equiv (\bar{e}_1 \wedge e_2)$, and $E_4 \equiv (\bar{e}_1 \wedge \bar{e}_2)$.

$$p_2 = P(f_{t+\Delta}|\bar{f}_t)$$

$$p_2 = \sum_{i=1}^{4} P(f_{t+\Delta}|E_i, \bar{f}_t)P(E_i|\bar{f}_t).$$

The four terms in the summation correspond to both events happening concurrently, a single event happening, or both not happening. The transition probabilities p_1 and p_2 are conditional probabilities. Therefore, the knowledge K available about different states during the interval $[t, t + \Delta]$ affects the probabilities. In general, $p_1 = P(\bar{f}_{t+\Delta}|f_t, K_{[t,t+\Delta]})$ and $p_2 = P(f_{t+\Delta}|\bar{f}_t, K_{[t,\, t+\Delta]})$.

Example 1. The dog is seen in the living room at 9:00 a.m. We are interested in evaluating the probability that the dog is in the room at 2:00 p.m. The dog may leave the room during any minute with probability $P(leave|inside) = 0.00579$ and it may enter the room during any minute with probability $P(enter|outside) = 0.00773$. The theorem shows that there exists a duration T such that the probability of fluent f at a time $t > t_0 + T$ changes by at most δ depending on the truth of f at t_0. Knowing the location of the dog at 9:00 a.m. affects the belief *dog is in the living room* at 2:00 p.m. by less than 0.02. It is therefore acceptable to answer the query based on the steady state probability[3] of the dog being in the living room at 2:00 p.m. without considering a 300 time-slice Bayesian net. This probability is 0.558. At steady state the probabilities do not change with time. To evaluate the steady state probabilities, we assume that

$$P(\text{DogInRoom}_{2:00\text{PM}}) \approx P(\text{DogInRoom}_{1:59\text{PM}}) \approx P(\text{DogInRoom}_{\infty})$$
$$P(\text{DogInRoom}_{\infty}) = P(\text{DogInRoom}_{\infty})P(\text{DogStayInRoom}|\text{DogInRoom})$$
$$+ P(\text{DogOut}_{\infty})P(\text{DogEnter}|\text{DogOut})$$

and

$$P(\text{DogOut}_{\infty}) = P(\text{DogInRoom}_{\infty})P(\text{DogLeave}|\text{DogInRoom})$$
$$+ P(\text{DogOut}_{\infty})(\text{DogStayOut}|\text{DogOut}).$$

The probabilities $P(\text{DogStayOut}|\text{DogInRoom}) = 1 - P(\text{DogLeaves}|\text{DogInRoom})$, and $P(\text{DogStayOut}|\text{DogOut}) = 1 - P(\text{DogEnter}|\text{DogOut})$ are known, therefore we can solve the linear system described above to get the probabilities $P(\text{DogInRoom}_{\infty}) = 0.558$ and $P(\text{DogOut}_{\infty}) = 0.441$. Exact calculations show that the probability $P(\text{DogInRoom}_{2:00\text{PM}}) = 0.567$ and $P(\text{DogOut}_{2:00\text{PM}}) = 0.432$.

[3] The term *steady state probability* denotes the value toward which the probability converges as time goes to infinity.

An empirical upper bound on relevance time in general Bayesian networks can be obtained by considering the largest difference between any pair of probabilities in the conditional probability table associated with a node. The relevance time is obtained by dividing the logarithm of the desired δ value by the logarithm of the obtained largest difference. This upper bound works in about 98% of the cases (Tawfik and Barrie 2000).

3.4. Dependent Time Instants or Periods

Two time points or intervals are dependent if information available at one point affects beliefs at the other point. The duration between two dependent times is shorter than the extraneousness time.

As an example of dependent times, consider an observer monitoring the status of the light and listening to the barking of the dog. If at time t_i the dog's barking is heard, the observer should conclude that the dog is out in the backyard at this instant and for some time thereafter. But after listening for some time and not hearing the dog, the observer should be less certain about whether the dog is out or not. This decay in certainty with time is also a function of time that relates probabilities at all instants with an *event*.

Observation 2. The knowledge required for probabilistic temporal reasoning consists of probabilistic knowledge about states, events, and a causal structure specifying the effects produced by the events.

Rationale. For the generic time-sliced Bayesian network as in Figure 4, e_1 and e_2 are two events affecting a fluent f that tends to persist if nothing happens with probability

$$P(f_{t+\Delta}) = \sum_{i=1}^{4} P(f_{t+\Delta}|E_i, f_t)P(E_i|f_t)P(f_t) + \sum_{i=1}^{4} P(f_{t+\Delta}|E_i, \neg f_t)P(E_i|\neg f_t)P(\neg f_t),$$

where $E_1 \equiv (e_1 \wedge e_2)$, $E_2 \equiv (e_1 \wedge \neg e_2)$, $E_3 \equiv (\neg e_1 \wedge e_2)$, and $E_4 \equiv (\neg e_1 \wedge \neg e_2)$.

To get the same expressions as those used for planning using temporal Bayes nets (Dean and Kanazawa 1989), assume the e_1 and e_2 are mutually exclusive (ME). The mutual exclusion assumption is equivalent to the single event STRIPS assumption.

$$P(f_{t+\Delta}) = \sum_{i=1}^{3} P(f_{t+\Delta}|f_t, E_{ME_i})P(f_t, E_{ME_i}) + \sum_{i=1}^{3} P(f_t|\neg f_t, E_{ME_i})P(\neg f_t, E_{ME_i}),$$

where $E_{ME_1} \equiv (\neg e_1 \wedge \neg e_2)$, $E_{ME_2} \equiv e_1$, and $E_{ME_3} \equiv e_2$.

$$P(f_{t+\Delta}|f_t, E_{ME_i})P(f_t, E_{ME_i}) = P(f_{t+\Delta}|f_t, E_{ME_i})P(E_{ME_i}|f_t)P(f_t).$$

Similarly,

$$P(f_{t+\Delta}|\neg f_t, E_{ME_i})P(\neg f_t, E_{ME_i}) = P(f_{t+\Delta}|\neg f_t, E_{ME_i})P(E_{ME_i}|\neg f_t)P(\neg f_t).$$

The purpose of the above observation and analysis is to emphasize that the three basic elements a probabilistic temporal representation has to express are the following:

- Probabilities of fluents at any time ($P(f_t)$).
- Probability of occurrence of events and their dependence on states ($P(E_i|f_t)$).
- The probabilistic effects of the events ($P(f_{t+\Delta}|E_i)$).

The structure of a time-sliced Bayesian network captures the notion of persistence and that of causation. These two notions must be represented here as well.

4. REPRESENTING EVENTS, EFFECTS AND INTERACTIONS

In this section, the temporal probabilistic formalism is further developed to represent the states of the world as a function of time, deduce the possible events from observations about the world, and predict the possible effects of the events and their interaction with each other and the rest of the states. To represent the states of the world as a function of time, we use the dynamic probabilities introduced earlier.

4.1. Representing Isolated Events

A set of characteristic functions represents the effect of each event token of a given event type on other variables. *A probability function* defines a relation between $P(x|e)$ and time t for all t where x is a random variable and e is an event type. Such functions can represent causation and persistence. For example, if the persistence of *dog-out* is given by $P_t(do)$ and represents the probability that the dog stays out, then, assuming that the rate of occurrence of the event *the dog enters* is constant, we get an exponentially decaying persistence function of the form $e^{-h_0 t}$, where h_0 is the rate of occurrence of the event. Using this probability function, and applying Bayes' rule, we find that the probability that the dog is outside given that we heard its barking is given by an exponentially decaying function as shown in Figure 5.

Because the same event may have different effects, it may also have different probability transfer functions. For example, the event of switching the light on has two functions: the first represents its effect on *fo* because somebody at home might have turned it on; and the second represents the persistence of *lo* because once turned on, we expect the light to stay on for some time. Figure 6 shows the network and the probability functions associated with a *light-turned-on* event (abbreviated as *lton*).

Turning the light off is not the complement of turning the light on. Both events affect *family-out*, but because the bulb may burn out, turning the light on may provide stronger evidence supporting that someone is at home than seeing the light go off. On the other hand, a light going off token should affect *light-on* by making it false. As time progresses, the observation that the light was turned on some time ago no longer contributes to the conclusion of *family-out*. In the absence of more recent information, this observation becomes extraneous after few hours.

4.2. Reasoning with Isolated Events

In general, the time of occurrence of an event is uncertain and is expressed as a probability distribution. The same may be true for effects. Let T_1 represent the time of occurrence of the event and let T_2 be the time it takes a subsequent effect to develop. The time T when the effect starts is

$$T = T_1 + T_2.$$

The sum of any two independent random variables produces a new random variable whose distribution is given by the convolution[4] of the distributions of its constituents. Let f_1, f_2, and f be the probability distributions for T_1, T_2, and T, respectively. Then

$$f = f_1 \otimes f_2.$$

For example, consider a simplified Bayesian network with two nodes: *dog-out* and *hear-bark*. If f_1 is the distribution of t_{hb} when the dog barked, and f_2 is the distribution of

[4] For continuous time, the convolution is evaluated with the integral $f(t) = \int_{-\infty}^{\infty} f_1(t-\tau) f_2(\tau) d\tau$. For discrete time $f(n) = \sum_{m=0}^{\infty} f_1(m) f_2(n-m)$ is used.

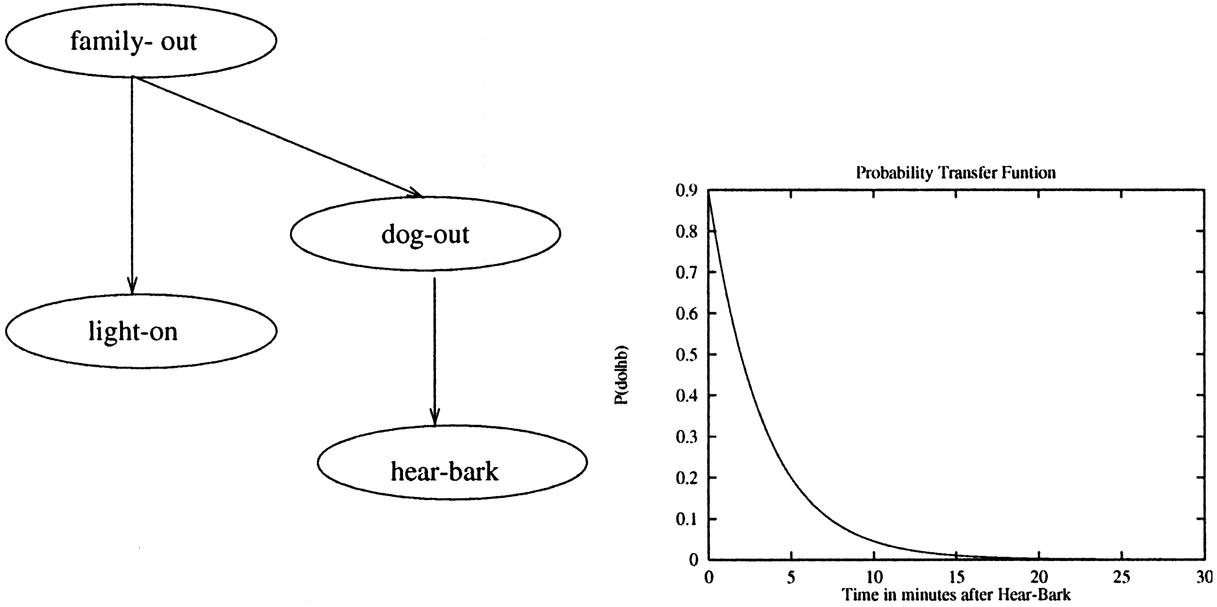


FIGURE 5. Posterior probability $P(do|hb)$ is also a function of time.

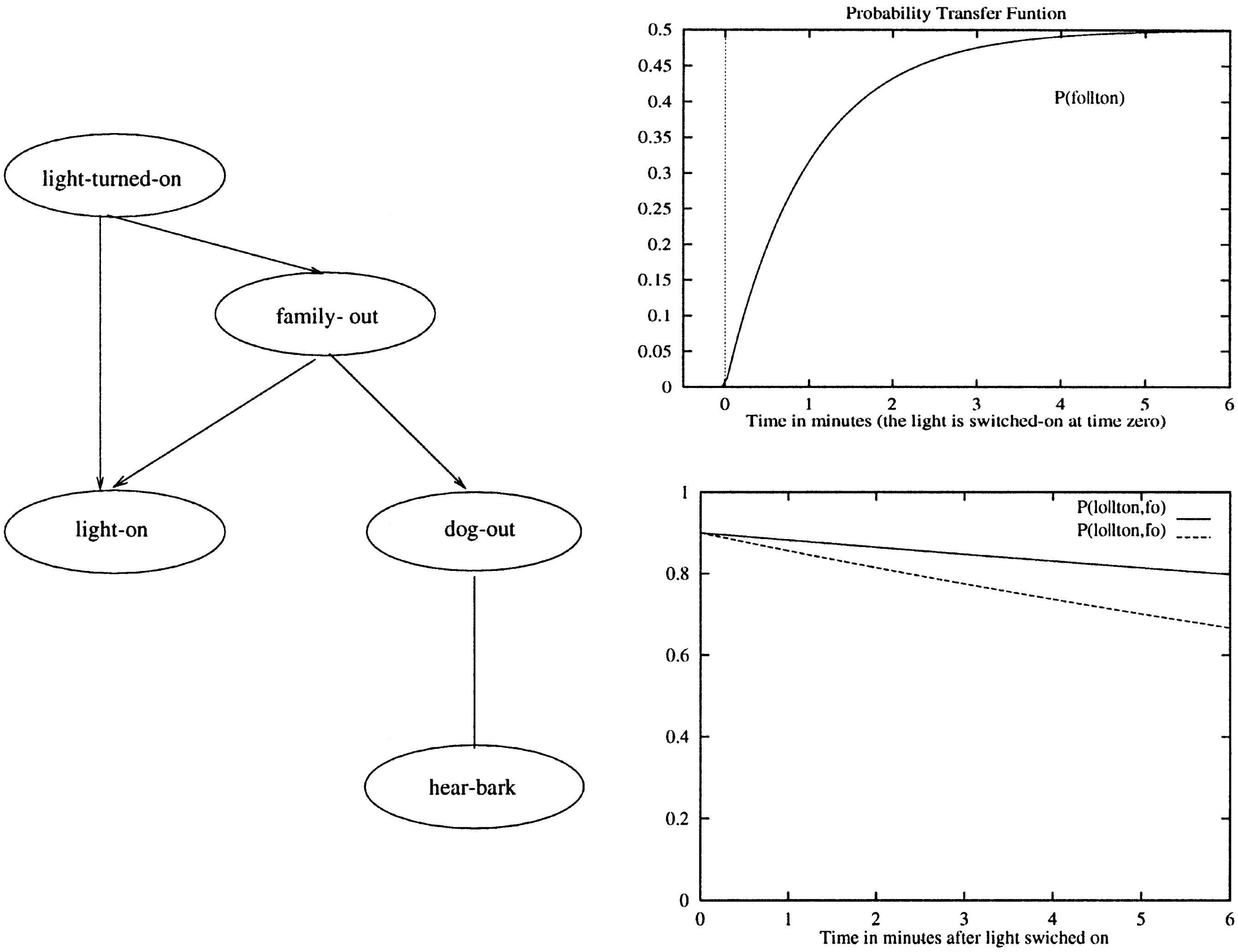


Figure 6. Light goes on.

dog-out given *hear-bark*, the convolution of f_1 and f_2 gives the probability of *dog-out* following a time-stamped *hear-bark* token. If the exact instant t_{hb} is known accurately, f_2 is simply then a unit impulse at time t_{hb}. The transfer functions used here represent the change in probabilities of the effects following the event. If we choose to represent the rate of occurrence of the effect, the probabilistic functions then correspond to what is usually referred to in survival analysis as *hazard functions*. The Appendix reviews the basic concepts of survival analysis as they relate to artificial intelligence.

The dog may be out and not barking (thus, *dog-out* may be true but without a corresponding *hear-bark* event). To allow for that, a probability $P_t(do|\bar{hb})$ is required; this is the probability that the dog is out and not barking. The probability that the dog is out at any time t is given by

$$P_t(do) = P_t(do, hb) + P_t(do, \bar{h}b).$$

To provide a failure interpretation for the above representation, consider failure to be ¬*dog-out*. The time to failure T is the duration the dog tends to spend outside after barking, the duration of persistence of *dog-out*. Here, the survival function represents persistence. To represent causation with survival functions, we treat effects as failures, and the failure time is the time between the cause and the effect. For example, the time taken by an infection to create a fever is the time to failure.

Events seldom produce the same effects independent of the rest of the environment. For example, the amount of time the dog spends outside depends on weather conditions. This dependence takes several forms. For n binary factors, $O(2^n)$ functions are needed to represent the interactions. Section 5 addresses interactions between events. Here, we consider interaction with background information. The background for an event is a set of variables that affect the outcome of the event under consideration, do not inhibit nor activate the event, are not affected directly or indirectly by this event, have a known temporal probabilistic profile or truth value, and are more stable than the event under consideration.

Representing these interactions requires an exponential number of conditional probabilities. To reduce the exponential complexity associated with this interaction, we propose the use of two statistical models that have been useful in applications in some physical sciences, social sciences, medicine, and engineering. The models provide a compact combination rule that replaces the exponential number of conditional temporal probability functions by expressing $P_t(Y|X_1, \ldots, X_m)$ as a function $\Phi(X_1, \ldots, X_m, t)$, where $X_1, \ldots, X_m$ are the background factors (or explanatory variables). The factors can be binary or real-valued.

A temporal distribution giving the probability that a random variable X_i takes a value x_i at time t ($P_t(X_i = x_i)$) is necessary to deal with the uncertainty about the value of X_i. For example, assume a binary outcome Y that depends on n factors, all having known values except for X_j, which takes one of m possible values at random. The probability $P_t(Y|X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n)$ can easily be evaluated when the temporal distribution of X_j is known, in which case

$$P_t(Y|X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n) = \sum_i^m \Phi(X_1, \ldots, X_j = x_i, \ldots, X_n, t) P_t(X_j = x_i),$$

where Φ is a function modeling the interdependencies between Y on the X's.

The two models proposed here are known in the statistics literature as the proportional hazard model and the accelerated time model. Although distinct, they can be seen as domain-dependent alternatives (Wei 1992).

4.2.1. Proportional Hazard Model. The proportional hazard model is a widely used survival model; it is a parametric model that allows environmental and background factors to be taken into consideration. As first proposed (Cox 1972), it assumes that the natural logarithm of the ratio of the conditional hazard function (in the presence of explanatory variables) to the hazard $h_0(t)$ (in their absence) is a linear-weighted sum of the risks, or

$$h(Y|X_1 \ldots X_m) = h_0(t)e^{\sum_{i=1}^{m} \beta_i X_i(t)}.$$

Each risk factor X_i multiplies the hazard by a constant $e^{\beta_i X_i}$ if it is present, hence the name *proportional hazard*.

As an example, consider that the dog is less likely to stay out when it is raining. The hazard function therefore becomes

$$h(do|X_r) = h_0(t)e^{\beta_r X_r},$$

where X_r represents the rain, β_r is a parameter that reflects the effect of the rain on the probability, and $h_0(t)$ is calculated in the absence of *rain*. This new function models the effect of *rain* on *dog out*. Now, we verify that *rain* is a background variable. There is a causal relation between *rain* and *dog-out*. *Rain* results in *dog-wet*, and *dog-out* is less likely given *dog-wet*. Second, *rain* does not inhibit nor activate *dog-out*. Third, it is not affected by *dog out*. Fourth, the variable *rain* is true when the dog is out. Fifth, *rain* tends to last longer than *dog-out*.

The model assumes a linear sum in the exponent $\sum_{i=1}^{m} \beta_i X_i(t)$, which is similar to the linearity assumption in linear regression. To overcome the limitations implied by the linearity assumption made by this model, nonlinear mapping functions may be used such that the hazard function remains a linear combination of the mapped factors.

A limitation of the proportional hazard model as described above is that it assumes time invariant effects of the factors, which is not always true. To represent time dependence, it is possible to use functions $\beta_i(t)$ instead of the parameters β_i in the model described above. Techniques that support such models have been proposed for continuous time (West 1991) and discrete time (Singer and Willett 1993).

4.2.2. Accelerated Time. Another way of dealing with the effects of background variables on lifetimes is to consider a different time scale t' and find a function $m(X_1, \ldots, X_n)$ such that a normal time unit (e.g., a second) in the absence of the time invariant background factors corresponds to $m(X_1, \ldots, X_n)$ units[5] owing to the effect of the factors. This is done by assuming that the time depends on the background variables (Kalbfleisch and Prentice 1980). Substituting t' for t in the original survival function produces the new function in the presence of the factors:

$$S(Y|X_1 \ldots X_m) = S_0(tm(X_1, \ldots, X_n)).$$

The hazard function is then

$$h(Y|X_1 \ldots X_m) = m(X_1, \ldots, X_n)h_0(tm(X_1, \ldots, X_n)).$$

For example, *rain* is a background factor affecting the probability of *dog-out*. Representing the effect of rain using this model results in the following hazard function:

$$h(do|X_r) = e^{\beta_r X_r}h_0(e^{\beta_r X_r}t).$$

When it is not raining, $X_r = 0$ and $h(do|X_r = 0) = h_0(t)$.

[5] $e^{\sum_i \beta_i X_i}$ is a commonly used function.

It is worth noting that the accelerated time model is more suitable when the background factors affect the time taken by the event to happen. The proportional hazard is more useful in expressing the change in the likelihood of the event because of the explanatory variables.

5. CAUSALITY AND MODELS OF INTERACTION

Reasoning just with isolated events restricts the possible application domains of a probabilistic temporal reasoning formalism. In most applications, repeated events of the same type or event tokens of different types interact. In nontemporal systems, causal interactions generally result in intractable knowledge acquisition and inference. Models such as the noisy-or model (Pearl 1988) or its generalization (Srinivas 1993) are useful in reducing this intractability. Each is designed to capture some important features of causal interactions. A similar approach is followed here. In the present treatment of causal interaction, no distinction is made between interacting recurring (repeated) events of one type and interactions between events of different types.

A hierarchical nontemporal classification of causal interaction (Heckerman and Breese 1994) distinguishes six classes of causal interactions. The most general class is that of general, multiple-cause interaction. The second most general class assumes independent causal inputs. The usefulness of this classification is rooted in its relation to some models of interactions. Here, we propose a temporal classification of causal interaction and show how the different classes are related to some interaction models.

Temporal event interactions can be divided into the following classes:

- Noninteracting events: Each event has a distinct causal structure so that the occurrence of one event does not affect the others' chance of happening nor its effects.
- Temporally independent events: The effect e produced by event c_j at time t is independent of other possible causes of e if none of them occurred concurrently with c_j (Heckerman 1993).
- Temporal necessity: The final effect is produced through two or more consecutive causal processes. Each causal process produces an event that determines the final outcome only partially by enabling some of the processes in the next level in the causal chain.
- Interaction through effects: Causes interact to produce a single outcome. This class includes three subclasses (Allison 1984) that roughly correspond to the following:
  - Competing causes: The success of a cause in producing the effect blocks the success of the other causes.
  - Asymmetric blocking: An event can block the other causes but is not affected by their outcome.
  - Collaborating causes: Events that help each other by raising (or lowering) the likelihood of a given outcome.

The above classification is not meant to cover all possible event interactions. The net effect of the interaction of two events is not always related to their individual effects. For example, pressing certain keys on a keyboard simultaneously can produce completely different effects from those produced if the same keys are pressed sequentially.

Models of interaction can be combined by model composition so that a larger model can be built by merging some of the simpler models. For example, two of three causes may collaborate with each other but compete against the third.

It is also possible to use model mixtures. Model mixtures are useful when some events randomly interact in more than one way. For example, two events may compete with probability p_c and collaborate with probability p_l. The mixture model is expressed as

$$P_t(Y) = p_c P_{t_c}(Y) + p_l P_{t_l}(Y),$$

where $P_{t_c}(Y)$ is the probability obtained from the competition between events and $P_{t_l}(Y)$ is the probability of Y when the events are collaborating.

Attaching causal semantics to statistical models makes it easier to integrate them into a knowledge-based system. Causal models assume that events trigger effects either immediately or after a time lag. The effects produced by an event may be affected by other events. Events that do not interact at all, at neither the causal side nor on the effect side, are completely independent or noninteracting. Two causal relationships, *RESULT* and *ENABLE*, suffice to describe the causal interactions underlying the models. *RESULT* establishes a cause-effect relationship between an event and a state. A state is said to *ENABLE* an event if it is a necessary condition for the event to take place (Pazzani 1990). The less formal notation "X is a *RESULT* of $C_1 \vee \cdots \vee C_n$" is used instead of "$RESULT(X, C_1 \vee \cdots \vee C_n)$" for improved readability whenever possible. Similarly, for *ENABLE* it is employed rather casually and "a state *DISABLEs* an event" if it inhibits it. The terms *more likely* and *less likely* represent the effect on the probability of an event.

The generalized noisy-or model (Srinivas 1993) can be used for independent events as well as temporally independent events (Heckerman 1993).

5.1. Events Occurring in Sequence

To represent temporal necessity, we introduce the sequential model. Frequently, an event E_2 can only occur after another event E_1. For example, *walking on the moon* cannot occur except after *arriving on the moon*. Let T_1 be the time E_1 occurs and T_2 be the time it takes E_2 to occur after E_1; the total time T required is therefore the sum of T_1 and T_2. Usually T_1 and T_2 are random variables with failure distributions[6] $F_1(t)$ and $F_2(t)$. The total time T has a distribution $f(t)$ given by their convolution:

$$F(t) = \int_0^{\infty} F_1(t) F_2(t-\tau)d\tau.$$

The underlying causal model assumes that event E_1 is initially enabled. The completion E_1 results in a state *ENABLing* E_2. This model generalizes to a sequence of n events. For example, an E_0 (*building the rocket*) may be necessary for E_1. In this case, $T = T_0 + T_1 + T_2$, where T_0 is the time required to achieve E_0. If a different requirement, E_{00} (*designing the telecommunications equipment*), is to be satisfied in parallel with E_0 then E_1 cannot start until both tasks have been completed. In this case, assuming that E_0 and E_{00} are independent, the distribution is given by $F_{00\wedge 0} = F_{00}(t)F_0(t)$. This type of interaction is usually useful in scheduling.

5.2. Competing Risks Model

As the name suggests, the competing risks model represents a world in which two or more potential causes race to achieve an outcome, but the success of one inhibits the others. Any one of $C_1, \ldots, C_n$ *RESULTs* in state S and state S *DISABLEs* $C_1, \ldots, C_n$ from succeeding. State S may be death or any state that cannot happen twice within

[6] Assuming a failure interpretation, T_1 is the time when the failure *arrive on the moon* occurs. This sounds awkward but any event (even a success) is a failure from a survival analysis point of view.

the given time frame, and $C_1, \ldots, C_n$ are potential causes for S. S is not necessarily a final state, but may be one that just briefly blocks the other competing causes. For example, consider the case of two infections with the same virus. The state *antibodies present in blood* blocks *second infection*. Competition is a relation between events, and in the statistical analysis of this model the nature of S does not affect the analysis.

The probability distribution for the result of the competition of two causes with failure densities f_1 and f_2 is given by

$$f(t) = f_1(t)S_2(t) + f_2(t)S_1(t),$$

where the survival functions S_1 and S_2 are defined as in the Appendix. Intuitively the equation means that the failure can occur because of the first hazard and surviving the other one or vice versa. Survival analysis provides a compact and efficient way to represent and evaluate the overall effect of the class of competing causes. This model can be used to model the airport pickup problem (Dean and Wellman 1991), in which John, who has arrived at the airport, tends to wait for some time. His tendency to wait decays with time as he gets bored. We are interested in the probability that John stays at the airport until we arrive to pick him up. Fred can meet John at the airport and give him a ride. Fred and boredom constitute two competing risks, each of which can cause the failure *John leaves the airport*. If we have a probability distribution for the time when Fred arrives at the airport, we can use the above formula to deduce the distribution for the event *John leaves the airport*.

5.3. Dominating Events Model

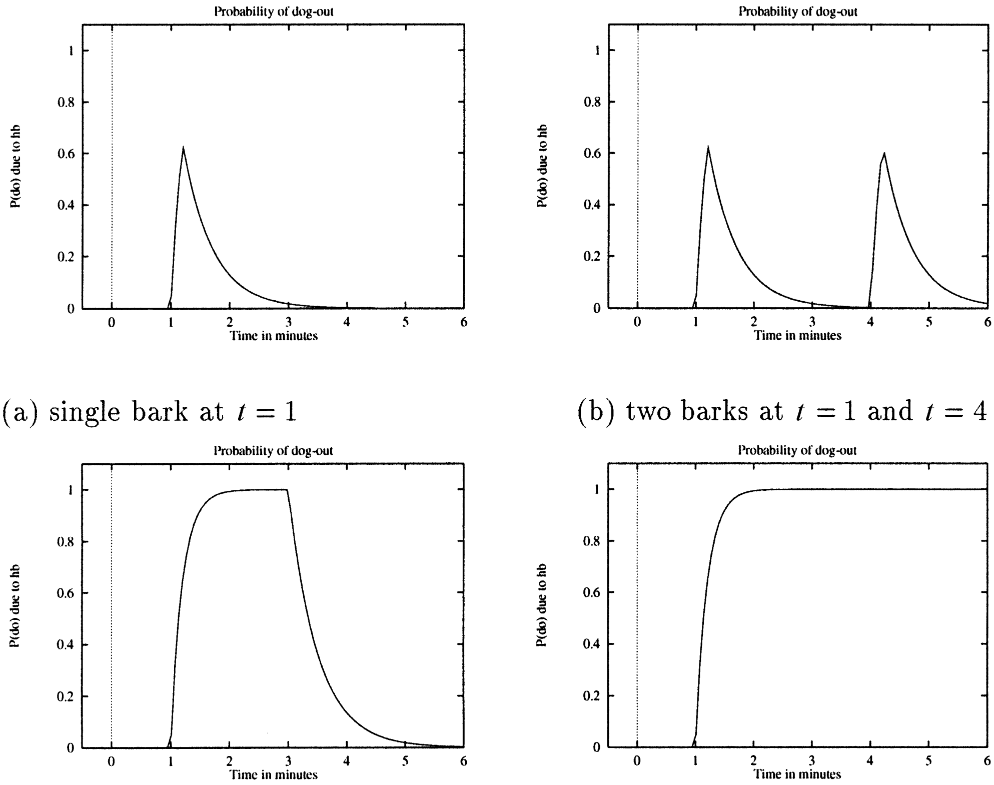
In this model, one event X tends to dominate other events $C_1, \ldots, C_n$ such that the probability of the outcome Y does not depend on $C_1, \ldots, C_n$ when X is true, or

$$P(Y|X, C_1, \ldots, C_n) = P(Y|X).$$

Here, Y is independent of $C_1, \ldots, C_n$ given X but Y depends on $C_1, \ldots, C_n$ given $\bar{X}$. For example, the death of a patient prevents the development of symptoms. The causal pattern captured by this model is as follows: X results in state S and the state $\bar{S}$ is necessary to enable the effect of $C_1, \ldots, C_n$. Rules to determine the dominating event are formulated explicitly along with the transfer function of this event. This model can be also used to capture the Markovian shielding property: given the causes of an effect, the effect becomes independent of all other events (Goldszmidt and Pearl 1992). For example, from a complete sequence of *changing-bulb* events, the probability of *burned-out-bulb* depends on the last *changing-bulb* and the lifetime of the bulbs. For this model to apply, the dominating event must make the dominated events irrelevant to the reasoning. Rules indicating domination, in the form "If X at t_i then ignore $C_1, \ldots, C_n$ at t_j in evaluating the probability of Y," allow domination to be stated explicitly. The model checks if the dominating event is known to be true, and ignores the dominated events accordingly. Constructs like *the most recent event*, *the earliest event*, and *the event resulting in a state Z* are usually useful in expressing the domination rules. This model is useful in modeling the interaction for the class of asymmetric blocking.

5.4. Storage Process with Additive Inputs

A storage process, like a reservoir, warehouse, or dam, is characterized by an inflow, a capacity, and release rules (Glynn 1989). If the events are additive inputs, the release rules are functions of the inputs and the storage level, and the level is the degree of



(a) single bark at $t = 1$

(b) two barks at $t = 1$ and $t = 4$

(c) barking continues from $t = 1$ to $t = 3$

(d) barking continuous from $t = 1$

FIGURE 7. Storage model results.

belief. Changes in the storage level reflect changes in belief with time. The level at any time t corresponds to the hazard $h(t)$. This model can be used for the *dog-out* and *hear-bark* causal relation. Every *hear-bark* token tends to fill belief in *dog-out* to a certain level and the release rule guarantees exponential decay of this belief. Following the conservation of mass principle, we cannot believe *dog-out* unless we hear barking at least once. But this does not mean $P(do) = 0$ since the calculated value represents the contribution of *hb* to the belief *do*, and the dog can be out and not barking.

Figure 7 illustrates how the probability of *dog-out* changes given different *hear-barking* event patterns. In this figure, the belief *dog-out* rises sharply whenever *hear-bark* takes place, and the degree of belief reached each time is slightly higher. If the barking is heard continuously over a period, the belief *dog-out* keeps rising during that period and then decays after barking ceases to be heard. The fourth event pattern in the figure deals with the case when the dog is heard continuously. In this case belief rises and then almost saturates.

For storage models, the causal model assumes collaboration and the events have an additive *RESULT* but limited persistence. The outflow represents the persistence.

This is not generally used as a survival model. It may be useful however for modeling failure due to stress build-up. Consider a reservoir with an incoming flow (inflow) and an outgoing flow (outflow). The incoming flow is the additive stress applied to the system and the outflow reflects the ability of the system to recover from stressful situations. Systems with adequate recovery can tolerate large stresses occurring over a long duration. They may fail however when the same stress is concentrated during a shorter duration. The special case of unlimited recovery corresponds to systems that would only fail if the instantaneous stress applied to them exceeds the maximum stress they can tolerate. Systems with no recovery let the stress add up until failure. The use of different release rules (for the outflow) can be motivated by the fact that most systems exhibit different patterns. Metals have unlimited recovery to low stresses within the elastic region and have no recovery once the stress exceeds a critical value causing permanent deformation.

6. PREDICTIVE AND EXPLANATORY REASONING

In a theory of action, inference is a two-stage process of explanation followed by prediction. The explanation stage is also the backward stage because it tries to find plausible events to explain an observation. The prediction or forward phase tries to find the possible consequences of an event.

Performing explanation or prediction based on new evidence usually changes the set of probabilities assigned to uncertain beliefs. Starting with a set of beliefs K and an event or observation A, a new set of beliefs incorporating possible explanations for A has to be formed first, then K is updated to incorporate the possible consequences of A and its preconditions.

6.1. Predictive Inference

Tawfik and Neufeld (1999) suggest a causal system that must consider eight cases corresponding to the four possible persistence patterns of causes and effects paired with immediate or delayed causation.

We assume that time of occurrence, duration of persistence of the cause, duration of persistence of the effect, and causation delay are all stochastically independent of each other. Let the cause starting time be given as a probability density $C_s(t)$, the cause persistence necessary to start producing effects is $C_{pc}(t)$, the causation delay $D(t)$, and the effect persistence $E_p(t)$. In this case, the probability of having effect E at time t is given by

$$E(t) = C_s(t) \otimes C_{pc}(t) \otimes D(t) \otimes E_p(t),$$

where $\otimes$ is the convolution operator.

Depending on the causation-persistence pattern, terms in the above equation may be omitted. For instance, if there is no causation delay, the delay term $D(t)$ is omitted. Similarly, $C_pc(t)$ and $E_p(t)$ disappear if the cause and/or the effect do not persist. The above equation corresponds to a persistent cause, persistent effect, and delayed causation. Expressions corresponding to other causation-persistence patterns can be derived in a similar fashion.

The quantities $C_s(t)$, $C_p(t)$, $D(t)$, and $E_p(t)$ are in general functions of the state as well as time. The dependence of these quantities on state reflects their conditional dependencies.

6.2. Explanatory Inference

For explanatory reasoning, it is assumed that the time when the effect manifests itself is known exactly or as a distribution. Now, given the effect start time T_{E_s}, we are interested in finding the cause time T_{C_s}, knowing the delay TD, the causation persistence of the cause $T_{C_{pc}}$, and the persistence of the effect T_{E_p}. The time separating T_{E_s} and T_{C_s} is the delay time. Thus, we have

$$T_{C_s} = T_{E_s} - T_D - T_{C_{pc}}.$$

A deconvolution operation can be performed using a transform to determine the distribution of the time of the cause knowing the distributions of the time of the effect and delay. Deconvolution is generally harder to evaluate than convolution (Mendel 1990).

6.3. From Subintervals to Intervals

Observing the status of a fluent or observing an event restricts the set of possible worlds to those in which the fluent can hold or the event can happen. This restriction is captured by the causal theory. The question that we address here is: What does the observation tell us about what is being observed? In general, an observation is made over a short duration (a subinterval); here we study the relationship between this subinterval and longer intervals. The study of temporal relationships between intervals and subintervals has been a point of interest in temporal logic (Shoham 1988). In a probabilistic theory of change, studying these relationships introduces a new set of issues. The first is that of a sensor model (or reliability of observation). The second issue results from the fact that different observations may belong to the same occurrence (same token) or to different tokens. The third issue deals with bidirectional persistence of the observed entity. How far in the past and in the future is it likely for the observation to hold?

The present treatment ignores the reliability of observations. Appropriate probabilistic solutions to this problem generally adjust the probability of evidence depending on the reliability of the observer (Hanks and McDermott 1994). For simplicity, we assume that observations are reliable.

Let X be a persistent cause or effect. An observation O at t_o indicates that X holds at this time point. First consider the persistence properties of X. If X has a limited persistence time L, it is possible that the observation corresponds to any point within a $2L$ duration. It is clear that a single observation indicates that X can start at most L time units before t_o and it can persist for at most L units after t_o. This results in X being possible over a $2L$ interval. This approach allows a formalization of the idea of regions of bidirectional persistence (Goodwin, Neufeld, and Trudel 1991).

Many logic-based AI formalisms model time as branching into the future at choice points and compute probabilities of future events by summing over choice points. However, frequently the past is as uncertain or unknowable as the future. This is the motivation behind the idea of regions of bidirectional persistence.

Example 2. Knowing that Joe is running now suggests he was also running a minute ago and will continue to run the next minute. Let $f_r(t)$ give the probability that Joe's running continues at least t units after Joe starts running. To simplify the arithmetic, assume that time is discrete, and that a run has a maximum duration of N units. We use the notation $P(e_k)$ for the probability that the run ends at time t_k, $P(r_k)$ for the probability that Joe is running at time t_k, and $P(s_k)$ for the probability that the run started at time t_k.

For a given starting time t_0, it is possible to predict the probability that Joe is still running at t given that he was running at t_i. To calculate this probability, we use survival analysis, considering stopping as a failure and $f_r(t)$ as a survival function. In this case the probability of stopping at time t is

$$P(e_j|r_i) = \frac{f_r(t_i - t_0) - f_r(t_j - t_0)}{f_r(t_i - t_0)}.$$

Therefore, the probability that the run continues at t_j is

$$P(r_j|r_i) = \frac{f_r(t_j - t_0)}{f_r(t_i - t_0)}.$$

If instead we are interested in an explanation of possible starting times, we assume prior indifference toward starting times, time invariance of f_r, and then use Bayes rule as follows:

$$\begin{aligned} P(s_0|r_i) &= \frac{P(s_0)P(r_i|s_0)}{P(r_i)} \\ &= \frac{P(s_0)f_r(t_i - t_0)}{\sum_{\Delta=0}^{N} P(s_{i-\Delta})f_r(\Delta)} \\ &= \frac{f_r(t_i - t_0)}{\sum_{\Delta=0}^{N} f_r(\Delta)}. \end{aligned}$$

The denominator in the above expression is constant; the numerator gives more weight for recent possible starting times (remember that survival functions are monotonically decaying functions of time). This supports our intuitions and the earlier result concerning the degeneration of relevance of information.

Two identical observations, O_1 and O_2 at times t_{o_1} and t_{o_2}, may belong to the same occurrence of X or to two different occurrences. Dynamic properties, such as the minimum duration required for X's recurrence or the duration between observations, can determine the answer.[7] It may also be possible to determine the answer by estimating the probabilities in each situation. If O_1 and O_2 belong to different tokens, then the analysis in the previous paragraph continues to apply. If O_1 and O_2 belong to a single occurrence of X, then the possible starting times are limited to the interval $\langle t_{o_2} - L, t_{o_1} \rangle$. Moreover, X is assumed to hold over the interval $\langle t_{o_1}, t_{o_2} \rangle$. Additional observations regarding the interval $\langle t_{o_1}, t_{o_2} \rangle$ do not add new information.

Example 3. The red traffic light at a given intersection persists for *four minutes*. Arriving at the intersection, a driver makes five equiprobable hypotheses regarding the time the light became red (corresponding to 4, 3, 2, 1, and 0 minutes ago). The driver of the car in the next lane says that the lights were red when she came two minutes earlier. This new information limits the possible hypotheses to three (4, 3, and 2 minutes).

In many practical situations, it is not possible or useful to assume a time limit for the persistence of X. Reasons for this may range from the possibility that X may hold indefinitely or that its persistence time varies widely. In such situations, it is not possible to proceed with the analysis discussed earlier. To proceed in such situations, additional

[7] It can be shown (Shahar 1997) that given a predefined confidence threshold δ and the transition probabilities p_1 and p_2 in a two-state Markov process, the probability that two observations belong to the same occurrence can be determined depending on the time separating the two observations and the transition probabilities. The result can be generalized to an n-state Markov process.

information about X is useful—in particular, the nearest time point when $\neg X$ was observed. It is possible to deduce that at least one transition from $\neg X$ to X, or vice versa, took place during the interval between the two observations.

Example 4. Parking a car in the driveway in the evening and not finding it the next morning may mean that it was stolen (Kautz 1986). Here, we have two observations at two different time points: car parked in the evening at t_i and car not parked in the morning at t_j. A change has occurred during the interval $[t_i, t_j]$ and some event C caused this change. $P_t(C)$ is a temporal probability distribution for the cause. This distribution has a nonzero value at least at one point during the interval $[t_i, t_j]$.[8] This condition guarantees that the cause could possibly have occurred, resulting in the effect. The probability of C given E is

$$P_{[t_i,t_j]}(C|E) = \frac{P_{[t_i,t_j]}(C)P_{[t_i,t_j]}(E|C)}{P_{[t_i,t_j]}(E|C)P_{[t_i,t_j]}(C) + P_{[t_i,t_j]}(E|\overline{C})P_{[t_i,t_j]}(\overline{C})}.$$

If C is the only possible cause for the change E then C is certain or $P_{[t_i,t_j]}(C|E) = [P_{[t_i,t_j]}(C)P_{[t_i,t_j]}(E|C)]/[P_{[t_i,t_j]}(E|C)P_{[t_i,t_j]}(C)] = 1$. If C results in E with the same probability regardless of the time of occurrence of the cause, the temporal profile of $P_t(C|E)$ from the above expression becomes a scaled version of $P(E)$. Therefore, the car was most likely stolen at the time of the night when most thefts occur.

The techniques described so far in this section serve as inference rules for our formalism. These rules are similar to logical inference rules in that they allow inference. Instead of inferring whether X is true or false, only the probability that X holds is inferred. A consistent theory reaches the exact same conclusion regardless of the inference rules used or the order of their application. To maintain the consistency of the present formalism, probabilities assigned have to be consistent. The treatment is consistent as long as it is true to the theory of probability.

7. RELATED RESEARCH

Reasoning about a changing uncertain world is a basic problem in AI. In addition to the work done in the subarea of temporal Bayes nets, surveyed in the introduction, there has been related work in other subareas.

Probabilities can represent the uncertainty of a dynamic world and can also handle unreliable observations (Hanks and McDermott 1994). An interruptible algorithm for temporal reasoning uses the more recent information first, then looks at the past for information that may affect the current conclusion. This is a very different solution for the information obsolescence problem from the one proposed here. Using the solution proposed by Hanks and McDermott, the probability of a fluent may oscillate as it looks at older information. This cannot happen with the solution proposed here because our information obsolescence criteria would consider all the information that may significantly affect the belief. Our approach however is not interruptible.

Markov processes are used to represent temporal phenomena for planning and diagnosis. The idea is to use a Markov chain consisting of a number of states where a transition matrix represents the possible state transitions as a result of an action or event. This model can predict the time when an observation is useful in a diagnosis

[8] If the cause can result in a delayed effect then a corresponding interval $[t_k, t_l]$, where $t_k = t_i - \Delta t$, $t_l = t_j - \Delta t$, and Δt, is the delay between event and effect. In this case, the techniques for delayed response can be used.

application (Portinale 1993). Dean (1994) uses a decision theoretic approach for choosing plans using this knowledge representation. Semi-Markov models are more general models, but they tend to be more complex. Starting with a semi-Markov model and simplifying the model seems to be a promising approach (Hanks, Madigan, and Garvin 1995). Approaches based on Markov models tend to ignore static causal relations. The approach proposed here allows the representation of static causal relations as well as dynamic ones.

A statistical event logic (Martin and Allen 1993) uses statistics to associate confidence intervals with the effects of an event in a planning application. This confidence interval is updated as the planner accumulates more experience. This statistical event logic uses Allen's (1984) temporal interval representation.

The use of simulation to perform temporal projections has attracted some attention (Yampratoom 1994). To calculate the probability of a fluent affected by a number of interacting events, events are generated based on their probability distribution. This approach has some advantages, including asymptotic convergence and interruptibility. The description of this system indicates that it supports to some extent concurrent interactions and takes into consideration the reliability of sensor observations. Stochastic simulation is also used to evaluate time-sliced Bayes nets (Kanazawa, Koller, and Russell 1995). Simulation can also be used within the proposed framework as a technique to calculate the probabilities.

8. DISCUSSION AND CONCLUSIONS

Getting temporal probabilities remains a challenging task. In the examples discussed here, some probabilities can be objectively determined from frequencies—for example, $P(fo)$. Others, such as $P(do|hb)$, are more subjective. Failures in circuits can be estimated from reliability models and may be thought of as propensities. Statistical models allow us to deduce the effect of infections. Such compromises are often forced on the Bayesian (Good 1983). The use of statistical techniques alleviates this problem to some extent due to the availability of software tools capable of deducing survival functions from data. The knowledge acquisition step would involve deducing survival functions and models of interactions from data.

Semi-Markov models are a generalization of survival models. Actually survival models can be deduced from semi-Markov models as special cases. They also generalize Markov models, which are becoming increasingly popular for probabilistic temporal representation. A semi-Markov model is similar to a Markov model but the sojourn time in a state is given by a survival function (Barlow and Proschan 1965). These models may be a good tool for more complex system. The choice of a knowledge representation is usually a question of choosing the least computationally demanding tool for the task at hand.

The knowledge representation consists of a static knowledge base and a dynamic one. The static knowledge base contains a description of event types, their survival functions, causal description of interactions, possible observations, the structure of possible cause-effect relationships in different contexts, background factors affecting each survival function, and Bayesian network structure. The dynamic knowledge base contains observations, events, and actions known about the time period under consideration. It also contains the updated probability profiles.

The techniques and results discussed earlier have many uses. For example, the idea of extraneousness can be used to divide the time line into related periods in virtually any temporal representation. The original intent of this work is to explore less computationally expensive techniques for probabilistic temporal reasoning. For readers interested in implementing some of the ideas discussed here, we propose a system structure consisting of the following modules:

1. An extraneousness test module that divides the time line into a set of related segments, and calculates the steady state probabilities.
2. An explanatory inference module that deduces the possible causes and preconditions for each observation.
3. A projection module that projects the effects of events, states, and deduced preconditions into the future.

The system described performs the two basic tasks in temporal reasoning: prediction and explanation. Many real-life problems including plan evaluation, medical diagnosis, and fault diagnosis in physical systems require these capabilities.

To evaluate a plan, the probabilities of different desirable and undesirable outcomes are needed. In medical diagnosis, information about the temporal progress of the symptoms may give strong evidence supporting a possible diagnosis. Lifetime information about the different components is very useful. Dynamic systems cannot be analyzed, simulated, or diagnosed without some type of temporal representation.

To summarize, the present article analyzes some aspects of temporal probabilistic reasoning. The results of this analysis are as follows:

1. If the duration between the observation time and the time point of interest is long enough, the observations can be ignored without loss in the accuracy of the conclusion.
2. The length of the duration required to make past observations extraneous depends on the dynamic nature of the system as reflected by the probability of transitions.
3. A temporal reasoning system has to perform three tasks: represent fluents, infer events from observations, and reason using interacting events.
4. Some interaction models can be used to reduce the elicitation and the inference complexity in probabilistic temporal reasoning.

In addition to the above, we have tried to provide useful categorization for the temporal relationship between causes and effects, and a classification for event interactions.

ACKNOWLEDGMENTS

This work has benefited from discussions with Mikelis Bickis. We thank the reviewers for their comments and we acknowledge the support of the University of Saskatchewan and the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

ALLEN, J. 1984. Towards a general theory of action and time. Artificial Intelligence, **23**(2):123–154.

ALLISON, P. 1984. Event History Analysis. Sage, Beverly Hills.

BARLOW, R., and F. PROSCHAN. 1965. Mathematical Theory of Reliability. John Wiley and Sons, New York.

BERZUINI, C. 1990. Representing time in causal probabilistic networks. *In* Uncertainty in Artificial Intelligence 5. Elsevier Science Publishers B.V., pp. 15–28.

CHARNIAK, E. 1991. Bayesian networks without tears. AI Magazine, **12**(4):51–63.

COX, D. 1972. Regression models and lifetables. Journal of the Royal Statistical Society, **B34**: 187–220.

DAGUM, P., A. GALPER, and E. HORVITZ. 1992. Dynamic network models for forecasting. *In* Proceedings of the 1992 Conference on Uncertainty in Artificial Intelligence, pp. 41–48.

DARWICHE, A., and M. GOLDSZMIDT. 1994. Action networks: A framework for reasoning about actions and change under uncertainty. *In* Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence, pp. 136–144.

DEAN, T. 1994. Expediting temporal inference by exploiting structure in time and state. Available by ftp from ftp.cs.brown.edu, file name pub/papers/ai/mdp.ps.Z.

DEAN, T., and K. KANAZAWA. 1989. A model for reasoning about persistence and causation. Computational Intelligence, **5**(3):142–150.

DEAN, T., and M. WELLMAN. 1991. Planning and Control. Morgan Kaufmann, San Mateo, CA.

DIACONIS, P. 1996. The cutoff phenomenon in finite Markov chains. *In* Proceedings of the National Academy of Science, **93**:1659–1664.

GLYNN, J. 1989. A discrete-time storage process with a general release rule. Journal of Applied Probability, **26**:566–583.

GOLDSZMIDT, M., and J. PEARL. 1992. Rank-based systems: A simple approach to belief revision, belief update and reasoning about evidence and action. *In* Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning, pp. 661–672.

GOOD, I. J. 1983. Good Thinking: The Foundations of Probability and Its Applications. University of Minnesota Press, Minneapolis.

GOODWIN, S., E. NEUFELD, and A. TRUDEL. 1991. Probabilistic regions of persistence. *In* Proceedings of the European Conference on Symbolic and Quantitative Approaches for Uncertainty.

HANKS, S., D. MADIGAN, and J. GARVIN. 1995. Probabilistic temporal reasoning with endogenous change. *In* Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence.

HANKS, S., and D. MCDERMOTT. 1994. Modeling a dynamic and uncertain world I: Symbolic and probabilistic reasoning about change. Artificial Intelligence, **66**(1):1–55.

HECKERMAN, D. 1993. Causal independence for knowledge acquisition and inference. *In* Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Mateo, CA, pp. 122–127.

HECKERMAN, D., and J. BREESE. 1994. A new look at causal independence. *In* Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Francisco, CA, pp. 286–292.

HITCHCOCK, C. 1993. A generalized probabilistic theory of causal relevance. Synthese, **97**(3): 335–364.

KALBFLEISCH, J., and R. PRENTICE. 1980. The Statistical Analysis of Failure Time Data. John Wiley and Sons, New York.

KANAZAWA, K. 1992. Reasoning about time and probability. Ph.D. thesis, Department of Computer Science, Brown University.

KANAZAWA, K., D. KOLLER, and S. RUSSELL. 1995. Stochastic simulation algorithms for dynamic probabilistic networks. *In* Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence.

KAUTZ, H. 1986. The logic of persistence. *In* Proceedings of the Fifth National Conference on Artificial Intelligence. Morgan Kaufmann, Los Altos, CA.

KEMENY, J., and J. SNELL. 1976. Finite Markov Chains. Springer-Verlag, New York.

KJÆRULFF, U. 1995. dHugin: a computational system for dynamic time-sliced Bayesian networks. International Journal of Forecasting, **11**(1):89–111.

MARTIN, N., and J. ALLEN. 1993. Statistical probabilities for planning. Tech. rept. 474, Department of Computer Science, University of Rochester.

MENDEL, J. 1990. Maximum-Likelihood Deconvolution: A Journey into Model-based Signal Processing. Springer-Verlag, New York.

NGO, L., P. HADDAWY, and J. HELWIC. 1995. A theoretical framework for context-sensitive temporal probability model with application to plan projection. *In* Proceedings of the Eleventh International Conference on Uncertainty in Artificial Intelligence.

PAZZANI, M. 1990. Creating a Memory of Causal Relationships: An Integration of Empirical and Explanation-Based Learning Methods. LEA Publishers, Hillsdale, NJ.

Pearl, J. 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo, CA.

Portinale, L. 1993. Selecting observation time in the monitoring and interpretation of time-varying data. *In* Advances in AI, Third Congress of the Italian Association for AI, (AI*AI'93). Springer-Verlag Lecture Notes in AI, pp. 314–325.

Shahar, Y. 1997. A framework for knowledge-based temporal abstraction. Artificial Intelligence, **90**(1/2):79.

Shoham, Y. 1988. Reasoning about Change: Time and Causation from the Standpoint of Artificial Intelligence. MIT Press series in AI. MIT Press, Boston.

Singer, J., and J. Willett. 1993. It's about time: Using discrete-time survival analysis to study duration and the timing of events. Journal of Educational Statistics, **18**(2):155.

Srinivas, S. 1993. A generalization of the noisy-or model. *In* Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Mateo, CA, pp. 208–215.

Tawfik, A., and T. Barrie. 2000. Degeneration of relevance in uncertain domains: An empirical study. *In* Advances in Artificial Intelligence: Proceedings of the Thirteenth Canadian Conference on Artificial Intelligence (AI'2000). *Edited by* H. Hamilton and Q. Yang.

Tawfik, A., and E. Neufeld. 1999. Changing times: A causal theory of probabilistic temporal reasoning. Journal of Experimental and Theoretical Artificial Intelligence, **11**:3–21.

Wei, L. 1992. The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. Statistics in Medicine, **11**(14/15):1871.

West, M. 1991. Modeling time-varying hazards and covariate effects. *In* Survival Analysis: State of the Art. *Edited by* J. Klein and P. Goel. NATO ASI Series, Kluwer Academic Publisher, Dordrecht.

Yampratoom, E. 1994. Using simulation-based projection to plan in an uncertain and temporally complex world. Tech. rept. 531, Department of Computer Science, University of Rochester.

APPENDIX: SURVIVAL ANALYSIS AND ARTIFICIAL INTELLIGENCE

Survival analysis is a relatively new subarea of statistics that emerged in response to the need to study events and event interaction. It differs from Markov models in that it uses distributions instead of a single transition probability. It differs from life-table methods in that it can account for time varying and time invariant factors affecting lifetimes. Survival analysis techniques, used in a wide variety of disciplines, are mostly the same whether the center of interest is the failure time of an engine or the success time in performing a given task in a learning process. The "failure" is any event of interest that can occur at any point in time. The "hazard" is the rate of occurrence of the event and a 'risk' is a potential cause.

Some earlier works have considered the use of survival models for temporal representation (e.g., Dean and Wellman 1991). The use of survival functions here is different. This difference is mainly attributed to our use of regressive survival functions. Regressive survival analysis expresses the probability of an event conditioned on observations and other events. To account for events interaction, survival models provide a compact and efficient solution. They avoid the problem of accounting for some factors more than once (Dean and Wellman).

In Section 5, survival models are introduced and used to predict the effect of event interactions. The advantages of using survival analysis are numerous, including the availability of survival analysis software tools to extract survival functions from historical data. They are also based on sound probabilistic theory and fit directly in the probabilistic reasoning framework. On the other hand, to make them fit into the AI setting, we generalize the models in more than one way. First, the concept of time here is not limited to clock time and utilization time but extends to include other forms such as the number of times[9] a certain fluent was true. Second, the assumption that a failure can occur only once corresponds to a single occurrence of events assumption. Often, such an assumption cannot be allowed; we have to allow multiple occurrences. Third, the assumption that everything fails eventually is equivalent to assuming that every event will occur at some future time.

Fortunately, these same assumptions are known to be problematic in social sciences (Allison 1984). Violating certain assumptions does not affect the models in any serious way. Others require some caution. The first assumption can be relaxed without problems. To allow the representation of repeated events, we consider each token of such events as a different event. The third assumption, the inevitability of death assumption, holds in a range of situations. For example, in model-based diagnosis, it is safe to assume that any component will eventually fail. In many other cases, this assumption is not justified. In such cases we can consider events that do not occur as happening at $t = \infty$. Most of the relationships and analysis continue to hold with some exceptions. We illustrate below, as an example of such exceptions, how the inevitability of death assumption affects the evaluation of survival functions. Some particular models make some assumptions regarding the mathematical properties of some functions (such as linearity). These assumptions can also be relaxed without serious problems.

In the following, T is the time of occurrence of an event E (or a failure). The temporal distribution of occurrence of an event can be expressed as a survival function $S_E(t)$, a probability density function $f_E(t)$, a probability distribution $F_E(t)$, or a hazard function $h_E(t)$. We drop the subscript if there is no confusion about the event. The probability density function $f(t)$, defined only for continuous time, is given by

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}.$$

The probability distribution function $F(t)$ is the probability that the event occurs by t and is defined in terms of $f(t)$ as

$$F(t) = P(T \leq t) = \int_0^t f(x)\,dx.$$

The survival function $S(t)$ is the probability that the failure has not occurred by time t:

$$S(t) = 1 - F(t).$$

If the inevitability of death assumption holds, we have that

$$S(t) = P(T > t) = \int_t^\infty f(t)\,dt.$$

If death could be avoided, which may be the case given a certain interpretation of death, then the first expression for $S(t)$ determines the occurrence of survival. The second expression fails to account for individuals who never die (or die at $t = \infty$). The inevitability of death assumption is sometimes woven into some survival models. One ought to be careful when using these models if this assumption does not hold.

[9] Using the count of times the key is turned as our time allows us to represent the decreasing belief that *the engine will start* as we repeat *turning the key* over and over. This example is from Darwiche and Goldszmidt (1994).

The hazard function $h(t)$ is the rate of occurrence of the failure and it is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t}.$$

We can also deduce that $h(t) = f(t)/S(t)$. By substitution and integration we get a useful general form for the survival function (correct even if the inevitability of death assumption does not hold):

$$S(t) = e^{-\int_0^t h(x)\,dx}.$$

From the above equations, we can draw several useful conclusions. First, simple relations exist among the four functions and knowing one fully specifies the others. Second, the survival function is monotonically decreasing because an event is more likely to happen as we allow for more time. For example, a constant hazard would result in an exponentially decreasing survival probability. Third, the equations described do not consider the effect of variables other than time. Solution of this problem requires the use of models. These models replace $h(t)$ by a conditional version $h(t|X_1, \ldots, X_n)$, where the X_i are in general time varying factors affecting the survival. These factors are sometimes called explanatory variables.

For small Δt, $h(t)\Delta t$, according to its definition, approximates the conditional probability that the event occurs during the period Δt. The conditional probability that an event E takes place during an interval Δt is given by

$$P_E(t < T \le t + \Delta | T > t) = \frac{S_E(t) - S_E(t + \Delta)}{S_E(t)}.$$