Irrelevance in Uncertain Temporal Reasoning

Ahmed Y. Tawfik and Eric M. Neufeld
Department of Computer Science
University of Saskatchewan,
Saskatoon, Saskatchewan S7N 5A9, Canada
{tawfik,eric}@cs.usask.ca

Abstract

In the presence of uncertainty, relevance of information degenerates as time evolves. This work shows that this degeneration occurs in probabilistic temporal reasoning. A mechanism for analyzing this phenomenon uses a Markov chain representation and a degree of relevance measure called temporal extraneousness.

Efficiency of probabilistic temporal reasoning can be improved by ignoring irrelevant and weakly relevant information. The analysis introduced here allows us to identify the portion of event history affecting the time instant of interest. The duration of relevant history depends on the dynamic nature of the system and the chosen relevance threshold. These notions are used to prune time-sliced Bayesian networks which constitute a popular probabilistic temporal reasoning knwolege representation.

1 Introduction

The success of intelligent systems in performing useful tasks depends to a large extent on the availability of relevant knowledge. A paradox arises in most practcal systems when the amount of knowledge grows beyond the capacity of the intelligent systems. Too much knowledge usually results in a significant degradation of the systems' performance. This performance degradation is usually attributed to the computational complexity of reasoning. Poor performance threatens the usefulness of intelligent systems and their success. Hence, it is desirable to have the knowledge base as concise as possible. Nevertheless, coverage of different situations and exceptions is an equally sought feature. The study of information relevance tries to determine what information is relevant to a particular task and what can be ignored without compromising the conclusion. For example, a query answering system has to identify and select from a knowledge base information relevant to answering a given query. Throughout this work, we assume, without loss of generality a query

answering task. The analysis and conclusions can be directly applied to any other task. In a query answering task, the subset of the theory relevant to a particular query is defined as follows:

Definition 1 *Given a theory Θ consisting of a set of predicates and a query Q consisting of a conjunction $Q = (q_1 \wedge q_2 \wedge \ldots \wedge q_n)$, the relevant theory Θ_Q is a minimal subset of Θ such that $\forall q_i \in Q$, q_i follows from Θ_q iff q_i also follows from Θ.*

To extend the above definition to probabilistic queries, we first consider that all q_i's in the query Q are independenet. In this case, a probabilistic version of the definition would require that the probability of each individual belief q_i remains unchanged in the full theory and the abridged one, or $P(q_i|\Theta) = P(q_i|\Theta_Q)$. In probabilistic terms, irrelevance of the information in the set $\Theta - \Theta_Q$ with respect to query Q implies that Q is independent of $\Theta - \Theta_Q$. In general, the conjunctive query Q may contain dependent beliefs. In this case, we require the abridged theory to provide the same values for the individual beliefs as well as for the joint distribution.

The objective of this work is to study irrelevance in temporal probabilistic domains. These domains are dynamic with uncertain temporal evolution. The ability to forecast and explain changes in such domains is required for a wide range of applications spanning planning, diagnosis, natural language understanding and scheduling. Efficient performance of such tasks requires a theory of relevance in dynamic uncertain domains. In such domains, two classes of irrelevant information fit directly into the notion of irrelevance as independence developed so far. The two classes are mutually independent beliefs and conditionally independent beliefs. Mutually independent beliefs do not affect each other in any way. For example, the probability of precipitation does not depend on the color of my pen. Two conditionally independent beliefs affect each other in general but become irrelevant once other

information is known. For example, the probability of precipitation depends on the temperature but the absence of clouds rules out the possibility of precipitation completely.

In dynamic uncertain domains, we identify a third class of irrelevant information. This class includes information that becomes obsolete with time due to the uncertain dynamic nature of change. For example, knowing that the cat was lying on the living room couch several hours ago, does not help in predicting its present location. The relevance of such information degenerates with time. Unfortunately, the notion of relevance developed so far does not capture this class of irrelevant information. It is not possible to claim that observing the cat on the couch is independent from its location. In our example, there is a slim chance that the cat remained on the couch all this time. It is not acceptable to claim that the probabilistic dependence fades away with time. Instead, we try to bridge the gap between our commonsense notion of irrelevance and the definition. We define a weaker temporal relevance criterion called *temporal extraneousness*.

Definition 2 *If the maximum effect of knowledge Θ at time t_j on belief q_l at time t_i is less than a small value δ, t_i and t_j are temporally extraneous with respect to q_l. The extraneousness level $\delta \ll 1$ is met when the inequality $|P(q_{l_i}|\Theta_j) - P(q_{l_i}|\neg\Theta_j)| \leq \delta$ holds.*

The strength of the degree of relevance as measured by temporal extraneousness can change according to the value of δ. A δ value of zero results in the know strong irrelevance notion of probabilistic independence. Irrelevance is weaker for higher δ settings. Weak temporal relevance corresponding to small δ values is of particular interest here. We redefine irrelevance using extraneousness instead of independence. The new definition allows us to ignore weak relevance.

Definition 3 *The theory Θ can be divided into a relevant theory Θ_Q and and extraneous information Θ_E to answer the query Q with accuracy δ iff $|P(Q|\Theta) - P(Q|\Theta_Q)| \leq \delta$ and for any $q \in Q$ $|P(q|\Theta) - P(q|\Theta_Q)| \leq \delta$.*

Starting with the complete theory Θ, we would like to identify the irrelevant and weakly relevant information. This information can then be discarded and reasoning can proceed with the concise theory Θ_Q. Here, we concentrate on the degeneration of the relevance of information as time evolves. Past information become irrelevant after a time duration T. This duration depends on the dynamic nature, the required

accuracy δ and the probability distribution describing the process.

2   Extraneousness Time

Before proceeding any further, it may be useful to examine more closely the performance savings that can be achieved by discarding irrelevant information. To this end, we start by considering the following example.

The cat is seen in the living room at 9:00 AM. We are interested in evaluating the probability that the cat is in the room at 2:00PM. The cat 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 $P(enter|outside) = 0.00773$.

Time-sliced Bayesian networks are an increasingly popular technique for solving this type of problems. A time-sliced Bayesian network is a sequence of instantiations each representing the state of different beliefs at a particular point in time. The arcs connecting two instantiations are sometimes called temporal arcs and are responsible for propagating temporal effects such as causation and persistence across time. There is no conscensus as to when a new instantiation is needed. DHugin networks add new instantiations at equal intervals [5]. Depending on the chosen temporal resolution, the network may contain a number of instantiations $n = T/\Delta$ where T is the total duration and Δ is the temporal resolution. Assuming a one minute resolution in our example, we can expect the network to have 300 instantiations. Other approaches may generate networks with only two instantiations. In either case, the computational complexity of inference in Bayesian Nets is high. Reducing the size of the network invariably results in significant performance gains.

Most people do not seem to find the information regarding the location of the cat at 9:00 AM relevant to finding it at 2:00PM. The following theorem supports this intuition. 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. Moreover for any arbitrarily small δ a different duration T can be found.

Theorem 1 *Consider a fluent f represented by the Markov Process in Figure 1 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 $1 > p_1, p_2 > 0$. 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*
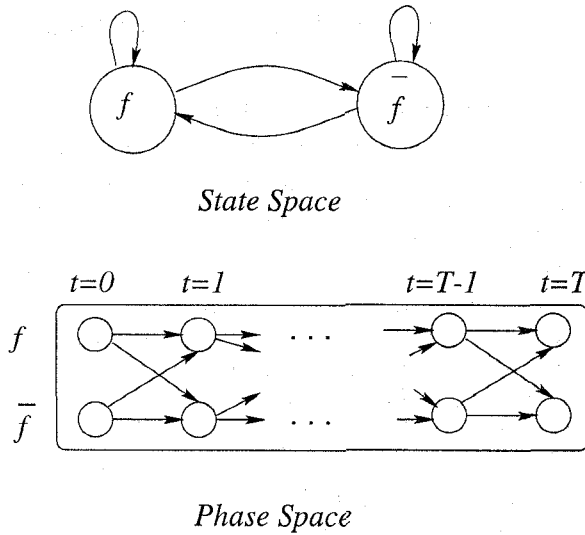
State Space

Phase Space

Figure 1: A two-state Markov Process

at time t be $P(f_t)$. The claim here is that for any $\delta \ll 1$ there exists T such that

$$|P(f_T|f_0) - P(f_T|\bar{f}_0)| < \delta$$

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 recurrent relation that can be solved using the iteration method to get

$$
\begin{aligned}
P(f_T|f_0) = \ & (1-p_1-p_2)^{T-1}P(f_1|f_0) + \\
& p_2(1-p_1-p_2)^{T-2} + \\
& p_2(1-p_1-p_2)^{T-3} + \ldots + \\
& p_2(1-p_1-p_2) + p_2
\end{aligned}
$$

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,

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

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

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 and p_2 the duration T that makes the difference less than δ can be determined as follows:

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

Applying the above result to the cat example, it turns out that knowing the location of the cat at 9:00 AM affects the belief *cat is in the living room* at 2:00 PM by less than 0.02. It is therefore acceptable to answer the query based on the prior probability of the cat being in the living room without considering a 300 stage Bayes net. This probability is 0.57.

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 transition probabilities p_1 and p_2 determine the dynamic properties of the system. Small transition probabilities indicate that states are persistent. A large p_1 and small p_2 lead the system to stay longer in state $\bar{f}$. Hence, the transition probabilities determine 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. This results in the deterministic commensense law of inertia [8]. 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. For example, if the cat could not come back to the room once it leaves it, then the calculated T gives the duration when the cat would have left with certainty $1 - \delta$. This probability is sometimes called time to absorption. Another extreme case occurs when $p_1 = 1$ and $p_2 = 1$, the system oscillates
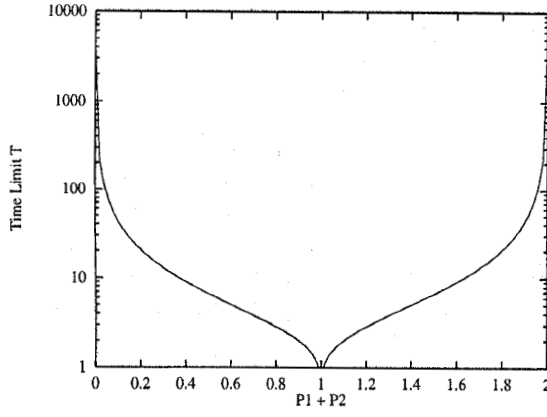
198

Figure 2: Extraneousness Time vs. Sum of Probabilities

and so does the probability difference [1]. In this case, the system is fully predictable given an initial state.

Figure (2) 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 state previous states. The following corollary shows this mathematically.

Corollary 1 *Any two time points t and $t + \Delta$ in a process as described in the theorem are temporally independent if $p_2 = 1 - p_1$.*
   Proof

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

But $P(f_{t+\Delta}|f_t) = P(f_{t+\Delta}|\bar{f}_t) = 1 - p_1 = p_2$ then,

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

The proof of the theorem assumes a stationary process (i.e. time invariant transition probabilities), but the result holds for time varying non-zero transition probabilities as well. In the case of time varying probabilities, an upper bound on T can be found by considering the lowest value possible value for $p_1 + p_2$ if this sum is less than one or the highest possible value for the sum greater than one.

It is possible to generalize the proof to non-binary variables; for a variable with n truth values, the proof

_____

[1] A negative difference simply indicates that the probability $P(f_T|\bar{f}_0)$ is larger than the probability $P(f_T|f_0)$. As an extraneousness criterion, the absolute value of the difference is what matters.

proceeds in a similar fashion except that for the number of probabilities as n^2 transition probabilities are needed. The result should also hold for continuous time. In the case of continuous time the probabilities are calculated by integrating the probability density functions over an interval Δt. The proof also assumes that the transitions take no time, which is not the case in most practical situations. Let I_{1-2} be the transition time from state f to state $\bar{f}$ and I_{2-1} the transition time from state $\bar{f}$ to state f.

Corollary 2 *A fluent f modeled 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 $1 > p_1, p_2 > 0$. The prior probability that the system is in a state f_i is $\frac{p_2}{p_1 + p_2}$.*
   Proof *By taking the limit as T tends to infinity in $P(f_T|f_0)$ and $P(f_T|\bar{f}_0)$ in the proof of the theorem.*

The convergence of probabilities in a regular Markov chain as time progresses [4] is related to the above result. The convergence property implies that a longer time duration allows a higher accuracy (smaller δ).

2.1 Multiple States Markov Processes

If P is a transition matrix, the element in the j-th row of the i-th column of the matrix P^T gives the probability of being in the state designated by the column at time T assuming that the process started from the state designated by the row. Checking for relevance of initial information in an m states Markov process can be performed in $O(m^3 log(T))$ using this property. The procedure for relevance checking squares the transition matrix until reaching the power T. The difference between the highest and lowest probabilities gives the maximum effect an initial state can have. If this difference is less that the chosen δ, the initial state can be considered irrelevant.

Example 1 *Consider a three states Markov process with the following transition matrix*

$$P := \begin{bmatrix} .1 & .1 & .8 \\ .4 & .3 & .3 \\ .6 & .2 & .2 \end{bmatrix}.$$

We would like to determine the relevance of information gathered 50 time units ago if the required accuracy $\delta = 0.01$. To check the relevance of this information we calculate P^{50}

$$P^{50} := \begin{bmatrix} .3759398497 & .1804511278 & .4436090226 \\ .3759398497 & .1804511278 & .4436090226 \\ .3759398497 & .1804511279 & .4436090227 \end{bmatrix}.$$
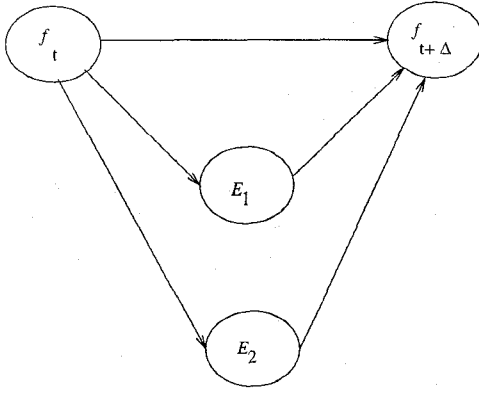
Figure 3: Generic Time-Sliced Bayesian Network

It is clear that the effect of this information is almost nil and there is not need to take it into consideration.

Note that the procedure described above can stop if the required accuracy is met before reaching T.

3 Irrelevance in Temporal Bayes Nets

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. Many of these use multiple instantiation of a static network. New copies are instantiated at different time points. The irrelevance criterion introduced here is particularly useful in such networks. Dynamic nets [1], temporal Bayes nets [2], and dHugin time-sliced Bayesian nets [6] are all based on the duplication approach.

These networks maintain Markov property which means that each node depends only on its immediate predecessor. This property means that nodes representing fluents at time t depend on some fluents at time $t - \Delta t$. A fluent at time $t - \Delta t$ can affect a node at time t through two mechanisms: persistence or causation. Persistence represent the probability that the fluent at time $t - \Delta t$ maintains its truth values at time t. In the absence of extreme probabilities, persistence is bound to decay as in the case of Markov chains. Causation may trigger a change in truth value immediately or after a causation delay. Instantaneous causation is weaker than evidence because the causal mechanism may fail. Therefore the bound on relevance time continues to hold. Delayed causation means that a fluent at time t causes another fluent to became true (or false) at time $t + D$. Delayed causation becomes irrelevant after a duration $T + D$ where T is the relevance time and D is the causation delay.

To apply the above theorem to temporal Bayesian networks, estimation of p_1 and p_2 is necessary. Those probabilities can be directly obtained in virtually any temporal probabilistic formalism.

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, in Figure 3 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 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. In our cat example, event e_1 corresponds to the cat entering the room during any minute, e_2 corresponds to the cat leaving the room. It is possible that the cat enters and leaves within the same minute or stay away from the room.

The transition probabilities p_1 and p_2 are conditional probabilities. Therefore, the knowledge K available about different states during the interval $[t, t + \Delta]$ affect the probabilities. In general, $p_1 = P(\bar{f}_{t+\Delta}|f_t, K_{[t,t+\Delta]})$ and $p_2 = P(f_{t+\Delta}|\bar{f}_t, K_{[t,t+\Delta]})$. For example, whether a car is in running condition or not, affects the probability of it staying parked or being driven away.

4 Discussion

The analysis of temporal degeneration of information relevance seems to suggest that maintaining a consistent level of predictive accuracy requires a continuous range of memory retention periods. Implementing such variable term retention can be done by attaching to each proposition a time-stamp and an expiry tag.

The lack of information during the duration between the observation time and the query time made the observation insignificant. In general, it is possible

to have fully reliable information or uncertain information about this interval. In the case of fully reliable information, the Markov property allows us to consider only the most recent one. However in the case of uncertain information, it is not generally possible to ignore the old observation or the recent one. The effect of the recent information on the belief may depend on the degree of belief resulting from the old information. In some special cases it is possible to ignore the old information. These cases include situations when old and new information interact according to the additive storage model [10] where the total belief is given by the sum of individual beliefs. In such situations a weak relevance continues to have a weak effect.

5 Related Research

Lin and Reiter [7] update a logical knowledge base by removing irrelevant information. Information irrelevance in their framework, corresponds to the first two types of irrelevance mentioned in the introduction. Irrelevance resulting from the dynamic nature of the process is not taken into account.

Ngo, Haddawy and Helwig [9] introduce a technique for finding information relevant to a particular query in a given context. This is close to Lin and Reiter notion of irrelevance. It also results in substantial reduction in the size of the problem. This technique can be used in conjunction to the extraneousness criterion introduced here resulting in a more manageable solution to probabilistic temporal reasoning.

An algorithm for temporal reasoning, proposed by Hanks and McDermott [3], uses the most recent information first, then looks in the past for information that may affect the current conclusion. This is a very different solution for the information obsolescence problem to the one proposed here. The algorithm cannot identify extraneous events and eventually considers them. Another problem occurs when events are not extraneous: the probability of a fluent may oscillate as it looks at older information which is a serious problem for such interruptible algorithm. This cannot happen with the solution proposed here because our information obsolescence criterion would consider all the information that may significantly affect the belief.

Backward smoothing is another alternative technique for dealing with the irrelevance of past information [6]. According to this approach, time is divided into a present window of arbitrary size, a past and a future. Statistical time series analysis techniques are used to perform backward smoothing and forward forecasting for the past and the future respectively. This approach does not take into account the dynamic nature of the process and the choice of the present window size does not depend on the probabilities or accuracy.

6 Conclusions

This study analyzed irrelevance in temporal probabilistic reasoning. The results of this analysis are:

1. Independence, conditional independence and extraneousness allow us to identify irrelevance and weak relevance in temporal domains.

2. If the duration between the observation time and the time point of interest is long enough, the observations can be ignored without significant loss in the accuracy of the conclusion.

3. 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.

Acknowledgements

References

[1] P. Dagum, A. Galper, and E. Horvitz. Dynamic network models for forecasting. In *Proceedings of the 1992 Conference on Uncertainty in Artificial Intelligence*, pages 41–48, 1992.

[2] T. Dean and K. Kanazawa. A model for reasoning about persistence and causation. *Computational Intelligence*, 5(3):142–150, August 1989.

[3] S. Hanks and D. McDermott. Modeling a dynamic and uncertain world i: Symbolic and probabilistic reasoning about change. *Artificial Intelligence*, 66(1):1–55, March 1994.

[4] J. Kemeny and J. Snell. *Finite Markov Chains*. Springer-Verlag, New York, 1976.

[5] U. Kjaerulff. A computational scheme for reasoning in dynamic probabilistic networks. In *Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence*, pages 121–129, San Mateo, CA, 1992. Morgan Kauffmann Publishers.

[6] U. Kjaerulff. A computational system for dynamic time-sliced bayesian networks. To appear in the International Journal of Forecasting, 1995.

[7] F. Lin and R. Reiter. How to progress a database (and why) i: Formal foundations. In *Proc. Fourth Int. Conf. on Principles of Knowledge Representation and Reasoning.*, 1994.

[8] J. McCarthy. Applications of circumscription to formalizing commonsense knowledge. *Artificial Intelligence*, 28:89–116, 1986.

[9] L. Ngo, P. Haddawy, and J. Helwig. A theoretical framework for context-sensitive temporal probability model with application to plan projection. In *Proceedings of the 11th International Conference on Uncertainty in Artificial Intelligence*, 1995.

[10] A. Tawfik and E. Neufeld. Temporal bayesian networks. In *Proceedings of Time-94: International Workshop on Temporal Knowledge Representation and Reasoning*, pages 85–92, Pensacola, Florida, 1994.