Taming Reasoning in Temporal Probabilistic Relational Models

Marcel Gehrke, Ralf Möller, and Tanya Braun
Institute of Information Systems, University of Lübeck, Lübeck
{gehrke, braun, moeller}@ifis.uni-luebeck.de

Abstract
Evidence often grounds temporal probabilistic relational models over time, which makes reasoning infeasible. To counteract groundings over time and to keep reasoning polynomial by restoring a lifted representation, we present temporal approximate merging (TAMe), which incorporates (i) clustering for grouping submodels as well as (ii) statistical significance checks to test the fitness of the clustering outcome. In exchange for faster runtimes, TAMe introduces a bounded error that becomes negligible over time. Empirical results show that TAMe significantly improves the runtime performance of inference, while keeping errors small.

Introduction
Temporal probabilistic relational models express relations between objects, modelling uncertainty as well as temporal aspects. Within one time step, a temporal model is considered static. When time advances, the current model state transitions to a new state. Performing inference on such models requires algorithms to efficiently handle the temporal aspect to be able to efficiently answer queries.

Reasoning in lifted representations has a complexity polynomial in domain sizes. But, models dissolve into ground instances through evidence, which no longer permits reasoning in polynomial time, making query answering infeasible for any reasoning algorithm, exact or approximate. Thus, a key challenge during inference in temporal models is to restore a lifted, i.e., non-grounded, representation. Therefore, we formulate and study the problem of keeping reasoning polynomial (KRP) in temporal models to tame the effect of evidence for efficient query answering.

First-order probabilistic inference leverages the relational aspect of a static model, using representatives for groups of indistinguishable, known objects, also known as lifting. Poole (2003) presents parametric factor graphs as relational models and proposes lifted variable elimination (LVE) as an exact inference algorithm on relational models. Taghipour et al. (2013) extend LVE to its current form. To benefit from the ideas of the junction tree algorithm (Lauritzen and Spiegelhalter 1988) and LVE, Braun and Möller (2016) present the lifted junction tree algorithm (LJT) for exact inference given a set of queries. To answer multiple temporal queries, Gehrke, Braun, and Möller (2018) present the lifted dynamic junction tree algorithm (LDJT), which combines the advantages of the interface algorithm (Murphy 2002) and LJT. Other approaches for temporal relational models perform approximate inference. Ahmadi et al. (2013) propose a colour passing scheme to obtain a lifted representation of a dynamic Markov logic network (DMLN) using exact symmetries and extend lifted belief propagation (Singla and Domingos 2008) for temporal approximate inference. Further inference algorithms for DMLNs exist (Geier and Biundo 2011; Papai, Kautz, and Stefankovic 2012). But, to the best of our knowledge, none of these approaches tackle the KRP problem.

For static relational models, approaches exist to approximate symmetries as evidence may ground even a static model (Van den Broeck and Davis 2012). Van den Broeck and Darwiche (2013) approximate lifted binary evidence. Singla, Nath, and Domingos (2014) propose approximate lifting techniques, which group together distinguishable objects and treat them identically. Venugopal and Gogate (2014) form clusters of objects and project the marginal distribution of one object to all objects of a cluster. Both approaches introduce an unknown bias into the distributions of the groups. Van den Broeck and Niepert (2015) present an unbiased approach for approximating symmetries. However, these approaches do not account for temporal aspects.

Thus, we present temporal approximate merging (TAMe) as an approach to solve the KRP problem in temporal models. Specifically, TAMe incorporates (i) clustering to group submodels and (ii) statistical significance checks to test the groups to be merged. Model structure and behaviour are captured in a set of functions that define local distributions for the random variables (randvars) in the model. Clustering forms groups of functions based on the similarity between local distributions. The significance checks allow for determining the fitness of the clustering outcome. If the clustering is deemed fit, each group is merged, yielding an unbiased approximation. In exchange for faster runtime, TAMe introduces a bounded error, which becomes negligible over time.

Boyen and Koller (1998) show that for stationary processes, evidence can lead to conditional dependences in temporal probabilistic propositional models, making inference infeasible. They propose to introduce additional randvars to achieve conditional independences between subprocesses.
We may omit \(|T|\) in \(A_{\uparrow T}\) or \(L_{\uparrow T}\). The term \(R(A)\) denotes the possible values (range) of a PRV \(A\). An event \(A = a\) denotes the occurrence of \(PRV A\) with range value \(a \in R(A)\).

We use the random variables \(A, D, R,\) and \(Pub\) for attending conference, does research, reputation, and publishes in journals, respectively, and \(L = \{X, J\}\) with \(D(X) = \{x_1, x_2, x_3\}\) (people) and \(D(J) = \{j_1, j_2\}\) (journals). We build boolean PRVs \(A(X), D(X), R(X),\) and \(Pub(X, J)\). A parametric factor (parfactor) describes a function, mapping argument values to real values (potentials).

**Definition 2.** We denote a parfactor \(g\) by \(\phi(A)|_C\) with \(A = (A^1, \ldots, A^n)\) a sequence of PRVs, \(\phi : \times^n_{i=1} R(A^i) \rightarrow \mathbb{R}^+\) a function with name \(\phi \in \Phi\), and \(C\) a constraint on the logvars of \(A\). We may omit \(|\top|\) in \(\phi(A)|_C\). The term \(lv(Y)\) refers to the logvars in some element \(Y\), a PRV, a parfactor or sets thereof. The term \(gr(Y|C)\) denotes the set of all instances of \(Y\) w.r.t. constraint \(C\). A set of parfactors forms a model \(G := \{g^i\}_{i=1}^n\). The semantics of \(G\) is given by grounding and building a full joint distribution. With \(Z\) as the normalisation constant, \(G\) represents \(P_G = \frac{1}{Z} \prod_{f \in gr(G)} f\).

Let us build the PM \(G_{ex} = \{g^i\}_{i=0}^3\), shown in Fig. 1, with \(g^0 = \phi_0(R(X), A(X), Pub(X, J))|_{\top}\) and \(g^1 = \phi_1(R(X), A(X), D(X))|_{\top}\), each with eight input-output pairs (omitted). Next, we present a temporal extension of a PM.

### Parameterised Probabilistic Dynamic Models

We define PDMs based on the first-order Markov assumption. Further, the underlying process is stationary.

**Definition 3.** A PDM \(G\) is a pair of PMs \((G_0, G_{\rightarrow})\) where \(G_0\) is a PM representing the first time step and \(G_{\rightarrow}\) is a two-slice temporal parameterised model representing \(A_{\rightarrow-1}\) and \(A_{\rightarrow}\) where \(A_{\rightarrow}\) is a set of PRVs from time slice \(\pi\). The semantics of \(G\) is to instantiate \(G\) for a given number of time steps, resulting in a PM as defined above.

Figure 2 shows \(G_{\rightarrow\pi}\) consisting of \(G_{\rightarrow\pi}\) for time slice \(t = \pi\) and \(G_{\rightarrow\pi}\) with \(inter\)-slice parfactors for the behaviour over time. The parfactor \(g^R\) is the \(inter\)-slice parfactor. For example, we can observe AAAI conference attendance, which changes over time as, unfortunately, getting papers accepted at consecutive conferences is difficult. Nonetheless, people with high attendance usually have a good reputation.

In general, a query asks for a probability distribution of a randvar given fixed events as evidence.

**Definition 4.** Given a PDM \(G\), a query term \(Q\) (ground PRV), and events \(E_{\circ\top} = \{E^i = e^i\}_{i=t}\), the expression \(P(Q|E_{\circ\top})\) denotes a query w.r.t. \(P_G\).

The problem of answering a query \(P(Q|E_{\circ\top})\) w.r.t. the model is called filtering for \(\pi = t\) and prediction for \(\pi > t\). In this paper, we focus on such temporal queries.

### Query Answering Algorithm: LDJT

The important property of LDJT [Gehrke, Braun, and Möller 2018] for this paper is that LDJT constructs first-order junction trees (FO jtrees) to efficiently answer multiple queries using LVE. The FO jtrees in LDJT contain a minimal set
of PRVs to m-separate time steps, which means that information about these PRVs renders FO jtrees independent from each other. Let us now define an FO jtree, with parameterised clusters (parclusters) as nodes, and present how formation about these PRVs renders FO jtrees independent.

**Definition 5.** Let $X$ be a set of logvars, $A$ a set of PRVs with $lv(A) \subseteq X$, and $C$ a constraint on $X$. Then, $A_{|C}$ denotes a parcluster. We omit $|C$ if $C = T$ and $lv(A) = X$. An FO jtree for a model $G$ is a cycle-free graph $J = (V, E)$, where $V$ is the set of nodes, i.e., parclusters, and $E$ the set of edges. $J$ must satisfy three properties: (i) A parcluster $C^i$ is a set of PRVs from $G$. (ii) For each parfactor $\phi(A)_{|C}$ in $G$, $A$ must appear in some parcluster $C^i$. (iii) If a PRV from $G$ appears in two parclusters $C^i$ and $C^j$, it must also appear in every parcluster $C^k$ on the path connecting nodes $i$ and $j$ in $J$. The parameterised set $S^1$, called separator of edge $\{i, j\}$, is defined by $C^i \cap C^j$. Each $C^i \in V$ has a local model $G^i$ and $\forall g \in G^i: vg(g) \subseteq C^i$. The $G_i$’s partition $G$.

Querying a minimal set of PRVs with LVE in an FO jtree combines all information to m-separate time steps. To obtain the minimal set, LDJT uses interface PRVs $I_t$ of $G_{t+1}$.

**Definition 6.** The forward interface $I_{t-1}$ is given by

$$I_{t-1} = \{A^i_{t-1} | \exists \phi(A)_{|C} \in G : A^i_{t-1} \in A \land \exists A^j_{t-1} \in A\}.$$ 

PRVs $R_{t-1}(X)$ and $A_{t-1}(X)$ from $G^{ex}_{t-1}$, shown in Fig. 2, make up $I_{t-1}$. While constructing FO jtree structures, LDJT ensures that the FO jtree $J_t$ for time step $t$ has a parcluster containing $I_{t-1}$, which is called in-cluster, and a parcluster containing $I_t$, which is called out-cluster. The in- and out-clusters allow for reusing the FO jtree structures.

To proceed in time, LDJT calculates a forward message $m_t$ over $I_t$ using the out-cluster of $J_t$. Hence, $m_t$ contains exactly the necessary information, as a set of parfactors, to be able to answer queries in the next time step. Afterwards, LDJT adds $m_t$ to the local model of the in-cluster of $J_{t+1}$.

Figure 3 depicts passing on the current state from time step 3 to 4. To capture the state at $t = 3$, LDJT sums out the non-interface PRV $D(X)$ from the local model and received messages of $C^3_3$ and saves the result in message $m_3$. Increasing $t$ by one, LDJT adds $m_3$ to $C^3_3$’s local model.

**Temporal Approximate Merging**

In a temporal probabilistic relational model, evidence can slowly ground the model over time by introducing splits. We propose to name the problem of finding how to undo splits to retain a lifted solution over time, keeping any error unbiased and acceptable, as the KRP problem. Retaining a lifted solution over time means that lifted algorithms run in polynomial time w.r.t. the domain size if a lifted solution exists [Niepert and Van den Broeck 2014]. To solve the KRP problem, an approach is required to identify any number of clusters based on how similar $\phi$’s of parfactors are and combine them. To keep the error unbiased and acceptable, the groundings need to be accounted for and the identified cluster means have to discriminate the clusters. Unfortunately, to combine similar $\phi$’s, we cannot use the colouring algorithm [Ahmadi et al. 2013] as it uses exact symmetries. Before presenting TAMe, let us formulate the problem in terms of PDMs.

Even though LDJT instantiates vanilla FO jtree structures, $m_t$ carries over splits caused by evidence. Formally, the problem is that in a model $G_t = \{g_t^n\}_{n=1}^m$ at time step $t$, many parfactors are split. Whenever evidence leads to a split of a parfactor, the split carries over to subsequent time steps. Thus, $G_t$ has the following form:

$$\{g_{t+1}^{i_1}, \ldots, g_{t+1}^{i_m}\}_{i=1}^m, m \in \mathbb{N}^+.$$

For each $i$, the different $g_{i,j}^{t+1} = \phi_{i,j}^{t+1}(A^i)$, $1 \leq j \leq m$, have the same arguments $A^i$ but different constraints $C^{i,j}$ and varying functions $\phi_{i,j}^{t+1}$ as a result of evidence. The assumption is that some $g_{i,j}^{t+1}$ have similar $\phi$’s as differences introduced by evidence are minimal or otherwise are overcome by model behaviour over time, i.e., potentials align again. Then, one can combine similar $\phi$’s while introducing only a small and bounded error in exchange for faster reasoning. In the following, we show that the assumption holds,
by showing that φ’s converge, allowing them to be merged, and that the error TAMe introduces is bounded.

The idea for restoring a lifted representation is to merge those \(g_{i,j}^{t,k}\) with similar \(φ\)’s into one parfactor

\[
g_{i,k}^{t} = φ_{i,k}(A_{i}^{t})_{C_{i}^{t,k}}
\]

(2)

where \(φ_{i,k}\) represents a merged version of the combined \(φ_{i,j}^{t,k}\) and \(C_{i}^{t,k}\) is a union of the combined \(C_{i}^{t,j}\). Merging all parfactors that behave similarly for each \(i\) leads to a \(G_{i}\) of the following form with parfactors as in Eq. (2) and \(l < m\):

\[
\{g_{i,1}^{t}, \ldots, g_{i,l}^{t}\}_{i=1}^{n}
\]

(3)

With TAMe, we present a merging scheme that takes a model \(G\) as given in Eq. (1) and computes a model \(G’\) as given in Eq. (3). It is reasonable to apply TAMe to \(G\) when transitioning from time step \(t\) to \(t+1\) as the transition transfers any splits as well. In general, \(G\) may be any parfactor model and one may also transfer the idea to a DMLN model (Ahmadi et al. 2013). But, models may be very large, e.g., the union of all local models of an FO jtree \(J_{i}\), such that finding groups for each \(i\) is too costly. Therefore, we propose to make TAMe a subroutine of LDJT. Transitioning from \(t\) to \(t+1\) requires computing message \(m_{t}\), which provides a state description of \(t\) that is relevant to \(t+1\). Applying TAMe to \(m_{t}\) prepares a message with fewer groups, leading to fewer splits in \(J_{t+1}\). Additionally, \(m_{t}\) normally has considerably fewer parfactors than \(G_{i}\). Next, we explain in detail how to get from Eq. (1) to Eq. (3) with TAMe.

**Keeping Reasoning Polynomial with TAMe**

Algorithm 1 outlines TAMe to solve the KRP problem. Inputs are a model \(G\), possibly \(m_{t}\), as well as two additional parameters, radius \(\epsilon\) and significance level \(α\), which become important later on. The first step is to preprocess \(G\) for easier handling in subsequent steps. The main loop describes how a clustering algorithm identifies groups for merging and how groups are merged if TAMe deems the clusters to fit. The upcoming paragraphs discuss the individual steps of Alg. 1.

**Model Partitioning**

The preprocessing of \(G\) is a consequence of the following considerations. A challenge that arises from a model as in Eq. (1) is that merging parfactors for each \(i\) independent of each other may lead to different groups that cause splits again, undoing any merging efforts. Using an \(i\) at random and transferring the grouping of the \(i\) parfactors to all other parfactors may lead to unreasonable groups for the other \(i\)’s. A safe option is to multiply parfactors with overlapping constraints into one parfactor which in a worst case leads to \(n = 1\) and very large parfactors that no longer explicitly represent independencies and may complicate calculations for messages and queries. Within LDJT, one could trace back if a set of parfactors in \(m_{t}\) originates from the message that has come from the direction of the in-cluster to the out-cluster as this message contains information about the past and is the origin of the most splits in \(m_{t}\). Therefore, it may be possible to identify a unique \(i\) in \(m_{t}\) as a reasonable source for merging. However, there are no guarantees to find such an \(i\). Instead, we opt to partition the parfactors in \(G\) based on the logvars appearing in \(G\) into a set \(P\) of sets of parfactors. Each partition \(P \subseteq \{g_{i,1}^{t}, \ldots, g_{i,m}^{t}\}_{i=1}^{n}\)

(4)

where \(lv(g_{i,j}^{t}) \subseteq X_{p}\). The next step is to identify groups of parfactors in each partition that behave similarly.

**Parfactor Clustering**

After partitioning \(G\), each partition \(P \subseteq \{g_{i,1}^{t}, \ldots, g_{i,m}^{t}\}_{i=1}^{n}\) has parfactors whose constraints overlap between all \(t\) for each \(j\). Therefore, TAMe multiplies all parfactors with overlapping constraints into one parfactor before starting with identifying groups. If \(lv(g_{i,j}^{t}) \subseteq X_{p}\), each \(i\) refers to \(m\) parfactors with the same constraint over all \(i\)’s for each \(j\), i.e., the constraints are the same at position \(j\) for all \(i\)’s. Then, multiplication in \(P\) to combine PRVs with the same constraints boils down to

\[
P = \left\{ \prod_{i=1}^{n} g_{i,1}^{t}, \ldots, \prod_{i=1}^{n} g_{i,m}^{t} \right\} = \left\{ g_{p,1}^{t}, \ldots, g_{p,m}^{t} \right\}
\]

(5)

where multiplying parfactors corresponds to the LVE operation of multiply, c.f. (Taghipour et al. 2013).

To identify groups of parfactors with similar behaviour, one needs to specify (i) what “similar behaviour” means and (ii) how to find such groups automatically. We first consider the second item, which influences specifying the first item. TAMe needs to identify an unknown number of groups based on how similar \(φ\)’s are. Density-based clustering groups similar points into an unknown number of groups. Therefore, TAMe uses density-based clustering. For the evaluation, we instantiate TAMe with DBSCAN (Ester et al. 1996, Schubert et al. 2017) as the clustering approach. In the following, we illustrate how density-based clustering fits into the overall scheme of TAMe using DBSCAN. DBSCAN identifies data points as core points if in their neighbourhood, determined by a radius \(\epsilon\) around a point, lie a certain number \(\text{minPts}\) of other data points. A core data point makes up a cluster along with all the data points in its neighbourhood, which recursively proceeds with the next core data point in the neighbourhood. To determine data points in a neighbourhood, DBSCAN requires a distance function as an input. DBSCAN is able to detect outliers, which do not occur in any neighbourhood. For the purpose of clustering parfactors, we set \(\text{minPts}\) to 2 to be able to cluster even two parfactors. The distance measure should assess

Algorithm 1: Temporal Approximate Merging

\[
\text{procedure TAMe(Model } G, \text{ Radius } \epsilon, \text{ Significance } α)\]
\[\text{P} \leftarrow \text{partitioning of } G \text{ based on logvars} \quad \triangleright \text{Eq. (4)}\]
\text{for each partition } P \in \text{P do}\]
\[\text{P} \leftarrow \text{multiply overlapping parfactors} \quad \triangleright \text{Eq. (5)}\]
\text{K} \leftarrow \text{DBSCAN}(P, \epsilon, 2, \text{rsim})\]
\text{if ANOVA(K, rsim', α) rejects } H_0 \text{ then}\]
\[G \leftarrow G \setminus P\]
\text{for each cluster } K \in \text{K do}\]
\[G \leftarrow G \cup \{K \text{ merged}\} \quad \triangleright \text{Eq. (7)}\]
The same case arises for two parfactors mapping to $g_1$ and $g_2$, grounding each, the marginals for true assignment about twice as much as the second. Assuming both the similarity of two parfactors $g_1$ and $g_2$ angle between them, i.e., a high cosine similarity, which $\phi_{t,i,j}$ mapping to $d$ distributions and behave similarly. For example, a parfactor $\phi_{t,i,j}$ behaves assignment. Thus, in case the ratio of the potentials of the other hand specify the current weight for each possible assignment. Thus, in case the ratio of the potentials of two parfactors are similar, they also have similar marginal distributions and behave similarly. For example, a parfactor mapping to 4 and 2 and another parfactor mapping to 8.1 and 3.9 behave similarly. Both parfactors weight the first assignment about twice as much as the second. Assuming both parfactors are independent from the rest and only have one grounding each, the marginals for true and false will be 0.667 and 0.075 respectively, i.e., less than 0.01 apart from each other. The same case arises for two parfactors mapping to (4, 2) and (4, 1, 9) respectively.

Such potentials, when thought of as vectors, have a small angle between them, i.e., a high cosine similarity, which we use to specify “similar behaviour”. For the setup of the similarity of two parfactors $g_{t,i,j} = \phi_{t,i,j}^k(A^i)_{|C_i,i,j}$ and $g_{t,i,j}^{k,j} = \phi_{t,i,j}^{k,j}(A^i)_{|C_i,j}$, we use a function $rsim : (\times_{i=1}^n range(A^i) \rightarrow \mathbb{R}^+, \times_{i=1}^n range(A^i) \rightarrow \mathbb{R}^+) \rightarrow \mathbb{R}^+$ that is defined as follows:

$$rsim(\phi_{t,i,j}^k, \phi_{t,i,j}^{k,j}) = \left[1 - \frac{\sum_{a \in range(A^i)} \phi_{t,i,j}^k(a) \cdot \phi_{t,i,j}^{k,j}(a)}{\sqrt{\sum_{a \in range(A^i)} \phi_{t,i,j}^k(a)^2} \cdot \sqrt{\sum_{a \in range(A^i)} \phi_{t,i,j}^{k,j}(a)^2}}\right]$$  

(6)

The result of Eq. (6) lies in the interval $[0, 1]$. We calculate 1 minus the fraction to get a “distance” measure, in which a lower value means a closer distance.

As a consequence of $rsim$ with its codomain $[0, 1]$ as the distance function for DBSCAN, $\epsilon$ needs to be $\leq 1$. Overall, the inputs of DBSCAN for clustering parfactors are a partition $P$ of parfactors, $\epsilon$, $minPts = 2$, and $rsim$. $\epsilon$ trades off cluster sizes with accuracy. The output is a clustering (partitioning) of $P$, i.e., a set $K$ of sets in which each $K \in K$ is a set of parfactors that are assumed to behave similarly.

**Fitness of Clustering** The question that remains after clustering is: How good is the clustering? The clustering is highly influenced by the choice of $\epsilon$, which leads to large clusters if set to a high value but may also blur the potentials in the merged parfactor to a higher degree.

One could calculate the error introduced by the clustering w.r.t. a given PRV $A$ by comparing marginal distributions of $A$ before and after merging. However, if a model already is highly shatted, the computational effort can be very high to compute marginal distributions before merging.

DBSCAN clusters together parfactors with a small angle between them. So a clustering fits if the variance of angles within clusters is low and the variance of angles between clusters is high. Analysis of variance (ANOVA) is a statistical method to test for significance of a clustering. In our setup, ANOVA computes the variance of each parfactor in a cluster $K \in K$ to the mean parfactor of $K$ as well as the variance of the mean parfactor of $K$ to the mean parfactor of all points in $K$. Hence, it provides an indication of how good the clustering separates parfactors.

ANOVA is used to accept or reject hypotheses. The default hypothesis is that the means of all clusters are equal. For our problem, the default hypothesis $H_0$ is that the mean parfactors of the clusters are equal, i.e., are not statistically significant to discriminate clusters. The goal is to be able to reject $H_0$, that is to say there is more difference between than within clusters. In case TAMe can reject $H_0$, at least one cluster is significantly different from the others.

To compute a mean parfactor of a cluster $K$, TAMe calculates the average of all potentials while accounting for groundings. Formally, given a set of parfactors $\{\phi_{t,i,j}^k(A^i)_{|C_i,i,j}\}_{j=1}^m$, a mean parfactor $\bar{\phi}_{t,i,j}^k = \phi_{t,i,j}^k(A^i)_{|C_i,k}$ is determined by

$$\bar{\phi}_{t,i,j}^k(a) = \frac{\sum_{j=1}^m gr(\phi_{t,i,j}^k(a)_{|C_i,i,j})}{gr(\phi_{t,i,j}^k(a)_{|C_i,k})}$$  

(7)

for each $a \in range(A^i)$ and $C_i,k$ is a union of the different $C_i,j$. Thus, TAMe goes through all potentials and for each assignment, adds the current potential, which is multiplied by the number of groundings of the current parfactor. After all potentials are added up, TAMe divides the potential by the number of overall groundings to obtain a mean potential. To illustrate Eq. (7), consider a cluster with 3 parfactors. The first parfactor maps to the potentials 2 and 1 with 2 groupings, the second maps to 3.9 and 1.9 with 5 groupings, and the third maps to 8.1 and 4 with 1 grounding. To calculate the mean potential, TAMe calculates for the first mapping $(2 \cdot 2 + 5 \cdot 3.9 + 1 \cdot 8.1)/8 = 3.95$ and for the second mapping $(2 \cdot 1 + 5 \cdot 1.9 + 1 \cdot 4)/8 = 1.9375$. Thus, the mean parfactor maps to 3.95 and 1.9375 with 8 groupings.

To calculate variances of parfactors, TAMe uses $rsim$ as the clusters have been built based on $rsim$. The intuition behind the choice is that if two parfactors have a very small angle between their potentials, then the variance of the potentials would be close to 0. The variance increases with the angle between potentials. As the number of groundings influences the new potentials, we also include the number of groundings while calculating a variance as the function should reflect that a parfactor that represents more groundings has a greater weight than one parfactor with one grounding, i.e., semantically we have that factor more of the number of groundings while calculating a variance as the clusters have been built based on $rsim$. However, marginal distributions and behave similarly.

One could calculate the error introduced by the clustering w.r.t. a given PRV $A$ by comparing marginal distributions of $A$ before and after merging. However, if a model already is highly shatted, the computational effort can be very high to compute marginal distributions before merging.

DBSCAN clusters together parfactors with a small angle between them. So a clustering fits if the variance of angles within clusters is low and the variance of angles between clusters is high. Analysis of variance (ANOVA) is a statistical method to test for significance of a clustering. In our setup, ANOVA computes the variance of each parfactor in a cluster $K \in K$ to the mean parfactor of $K$ as well as the variance of the mean parfactor of $K$ to the mean parfactor of all points in $K$. Hence, it provides an indication of how good the clustering separates parfactors.

ANOVA is used to accept or reject hypotheses. The default hypothesis is that the means of all clusters are equal. For our problem, the default hypothesis $H_0$ is that the mean parfactors of the clusters are equal, i.e., are not statistically significant to discriminate clusters. The goal is to be able to reject $H_0$, that is to say there is more difference between than within clusters. In case TAMe can reject $H_0$, at least one cluster is significantly different from the others.
\[ MSG = \frac{1}{l-1} \sum_{k \in K} g^T(k) \cdot (r \cdot \text{sim}(g^T_k, g^T_{k,m}))^2 \]
\[ MSE = \frac{1}{m-1} \sum_{k \in K} \sum_{g_{i,j} \in K} \frac{g^{i,j}}{g^{i,j}} \cdot (r \cdot \text{sim}(g^{i,j}_k, g^{i,j}_l))^2 \]

where \( l = |K| \), i.e., number of clusters, and \( m = |\text{gr}(K)| \), i.e., number of overall groundings. Computing \( F = \frac{MSG}{MSE} \), ANOVA compares \( F \) against a critical value \( F_{\text{crit}} \), which depends on \( \alpha, l-1 \), and \( m-1 \) and can be looked up in a pre-computed table. If \( F \leq F_{\text{crit}} \), TAMe accepts \( H_0 \) and discards the clustering. In case TAMe rejects \( H_0 \), i.e., \( F > F_{\text{crit}} \), there is more difference between clusters than within clusters and TAMe proceeds to merging parfactors.

**Merging Parfactors** The new parfactor for each cluster \( K \in K \) is the mean parfactor \( g^T_{i,k} \) already computed by ANOVA. TAMe replaces \( P \) in \( G \) with the merged parfactors. Then, TAMe proceeds with the next partition, identifying and checking a clustering for the new partition, until all partitions are processed. The result is a model whose parfactors are merged versions of the input model, partially restoring a lifted representation. Given a forward message \( m_t \), the output is a message that possibly contains fewer groups within logvars and thus, prevents ongoing splitting over time.

**Application Cycle** As ANOVA may determine that the clustering is not fit enough, TAMe may incur overhead if TAMe cannot merge groups. Therefore, in most cases, TAMe should not be applied at every time step. Normally, the model is slowly grounded over time with evidence, but if the groups behave similarly, which is the case due to the impact of the model, the reoccurring application of the model behaviour results in the potentials being similar enough for TAMe to merge them. Thus, based on how much evidence splits up the model, the interval of how often TAMe should be used as a subroutine needs to be determined.

Next, we look at theoretical implications of TAMe.

**Theoretical Analysis** We show that TAMe introduces an acceptable, unbiased, and bounded error and that TAMe keeps reasoning polynomial.

**Proposition 1.** TAMe errors are acceptable and unbiased.

Due to a density-based clustering, TAMe clusters parfactors with similar \( \phi \)’s. ANOVA determines the fitness of clusterings to prevent unacceptable errors. By accounting for groundings during merging, the error is unbiased.

Knowing that TAMe produces acceptable and unbiased errors, let us have a look at theoretical bounds of the approximation error TAMe introduces as well as whether groups with only slightly different evidence do converge, allowing TAMe to keep reasoning polynomial.

**Theorem 1.** TAMe introduces a bounded error.

**Proof sketch.** A PDM is a Markov process and \( G_{\rightarrow} \) describes a temporal transitions model. Given the semantics of a PM, \( G_{\rightarrow} \) forms a stochastic transition model \( Q \), which has a so-called minimal mixing rate \( \gamma_Q \in [0,1] \) \( \text{(Boyen and Koller 1998)} \). \( \gamma_Q \) is the minimal extent to which the model behaviour causes an approximation to converge to the true belief state while transitioning from one time step to the next. TAMe approximates the belief state of the interface \( I_t \) and LDJT computes the transition from \( t \) to \( t+1 \). Thus, the approximation error \( \delta \) is reduced by the factor \( (1-\gamma_Q) \) with each transition. Assuming that TAMe introduces an error of at most \( \delta \) for each time step, the expected error up to time step \( t \) accumulates to \( \delta + (1-\gamma_Q) \delta + \ldots + (1-\gamma_Q)^{t-1} \delta = \sum_{i=0}^{t} (1-\gamma_Q)^i \leq \sum_{i=0}^{\infty} (1-\gamma_Q)^i = \delta/\gamma_Q \). For the last step, we apply the geometric series, i.e., \( \sum_{i=0}^{\infty} \delta \cdot (1-\gamma_Q)^i = \delta/\gamma_Q \). Thus, the error is indefinitely bounded by \( \delta/\gamma_Q \).

For TAMe the significance check influences the approximation error \( \delta \). Before TAMe merges parfactors and thereby, approximates a belief state, TAMe uses a significance check to determine the fitness of a proposed clustering. Therefore, one can use the significance check to obtain a small \( \delta \). Now, we prove that TAMe keeps reasoning polynomial.

**Theorem 2.** TAMe keeps reasoning polynomial.

**Proof sketch.** Evidence introduces a discrepancy between two distributions of the same origin. The minimal mixing rate \( \gamma_Q \) ensures that these two distributions converge again. Therefore, TAMe will merge these two distributions at some point in time. Merging distributions ensures that LDJT calculates a solution in polynomial time w.r.t. domains.

Now, we use Thm.2 to restore an original representation.

**Corollary 1.** Without new evidence, TAMe obtains a fully lifted representation with the true belief state.

**Proof sketch.** During each transition from \( t \) to \( t+1 \), \( \gamma_Q \) ensures that approximated distributions converge to the true distribution as the distributions converge at least by the factor \( (1-\gamma_Q) \). Thus, the approximated distributions converge to the true belief state without new evidence provided. Further, all groups have the same origin. Therefore, all groups converge to the same true belief state. Hence, TAMe can merge all groups and thereby, again obtain a fully lifted representation at some point in time.

Thus, TAMe solves the KRP problem. Since the underlying distributions of \( \phi \)’s converge, TAMe is able to merge \( \phi \)’s, allowing TAMe to keep reasoning polynomial. Further, TAMe introduces a bounded, unbiased, and acceptable error.

**Evaluation** For the evaluation, we compare runtimes of LDJT with and without TAMe and have a look at the introduced error. We use the model \( G^* \) with \( |D(X)| = 100 \) and divide these 100 persons equally into symmetry groups, where members of each group behave identically over time. For one time step, each symmetry group has the same evidence, but the evidence can change from one time step to the next. To break symmetries within a group, evidence may be missing with a probability of 0.1 for each person. We split \( D(X) \) into 2
Table 1: Introduced error; \( I = 2, \epsilon = 5 \cdot 10^{-2}, 10 \) groups

| \( \pi \) | Max | Min | Average |
|---|---|---|---|
| 0 | 0.0001537746121 | 0.0000000001720 | 0.0000091206488 |
| 2 | 0.00000000851654 | 0.0000000000001 | 0.00000000111949 |
| 4 | 0.0000000080478 | 0 | 0.0000000000068 |

Table 2: Introduced error; with and without significance test

| | w | w/o |
|---|---|---|
| Max | 0.000225927071 | 0.0000000000000 |
| Min | 0.0000000000168 | 0.0000000000168 |
| Average | 0.0000104567643 | 0.0000137870835 |

Figure 4 shows runtimes of LDJT without TAMe and with TAMe for the three options. The number of symmetry groups is plotted on the x-axis. With more symmetry groups, LDJT without TAMe is slightly faster. As LDJT sends between time steps as this message is smaller than the model and causes splits in the next time step. To identify similar instances, TAMe uses density-based clustering with the cosine similarity as a distance measure, which captures similarity of potentials. TAMe applies ANOVA to the clustering result to check if the cluster means significantly differ. In summary, even by only merging parfactors that hardly differ, TAMe merges enough parfactors to improve runtimes of LDJT. TAMe with Options 2 and 3 improves runtimes of LDJT significantly. Overall, TAMe is able to save runtime of LDJT of up to 2 orders of magnitude. Knowing that TAMe can significantly improve the performance of LDJT, we look at the costs of the speed up, namely the introduced error.

Table 2 shows the error in the marginals for 10 symmetry groups for the most aggressive option, when performing filtering, 2 time step prediction, and 4 time step prediction for each instance and each time step. For filtering queries, the error is already negligible and decreases for prediction queries. Thus, the empirical evaluation underscores that TAMe can keep reasoning polynomial, introducing only a negligible error. Further, the error converges to the true belief state without new evidence as the prediction queries show. Next, we take a look at the significance check.

To empirically evaluate the significance check, we run LDJT with TAMe on a model once with and once without the significance check. Table 2 shows the introduced errors for these runs. The maximum error hardly differs between the two runs, which is is due to the error being bounded. Further, the minimum error is lower with the significance check as the significance check does not accept all proposed clusters. Discarding a clustering and thus, not following through with another approximation, the current approximation and the true belief state continue to converge based on the mixing rate. Lastly, the average error without the significance check is around 32% higher. Even though in this case both average errors are negligible on an absolute scale, the average error on a relative scale without the significance check does increase significantly.

Overall, we show empirically that TAMe does not introduce any unacceptable error due to the significance check and that TAMe keeps reasoning polynomial for LDJT.

**Conclusion**

Evidence often grounds a temporal model over time. Consequently, inference runtimes suffer. Thus, the idea is to use approximate symmetries to restore a lifted representation and thereby, keep reasoning polynomial by taming evidence. To the best of our knowledge, we present the first approach solving the KRP problem for temporal relational probabilistic models, which can be used within any (exact or approximate) temporal inference algorithm. The main idea is that instances of parfactors with similar ratios between potentials behave similarly. To merge parfactors, TAMe uses a message LDJT sends between time steps as this message is smaller than the model and causes splits in the next time step. To identify similar instances, TAMe uses density-based clustering with the cosine similarity as a distance measure, which captures similarity of potentials. TAMe applies ANOVA to the clustering result to check if the cluster means significantly discriminate the clusters. We show that TAMe can merge parfactors as their distributions converge and that TAMe introduces a bounded error. Additionally, the approximated distributions converge to the true distributions and TAMe can obtain a fully lifted representation again without new evidence. Empirical results show that LDJT with
TAME significantly outperforms LDJT without TAME. The results support our analysis that TAME retains a lifted solution, while keeping the introduced error negligible. Hence, LDJT with TAME produces fast and precise results.

Future work includes how to approximate evidence (Van den Broeck and Darwiche 2013) to cause fewer splits in temporal models as well as learning temporal models.

Acknowledgement This research originated from the Big Data project being part of Joint Lab 1, funded by Cisco Systems, at the centre COPICOH, University of Lübeck

References

[Ahmadi et al. 2013] Ahmadi, B.; Kersting, K.; Mladenov, M.; and Natarajan, S. 2013. Exploiting Symmetries for Scaling Loopy Belief Propagation and Relational Training. Machine learning 92(1):91–132.

[Benjamin et al. 2018] Benjamin, D. J.; Berger, J. O.; Johannesson, M.; Nosek, B. A.; Wagenmakers, E.-J.; Berk, R.; Bollen, K. A.; Bremsb, B.; Brown, L.; Camerer, C.; et al. 2018. Redefine statistical significance. Nature Human Behaviour 2(1):6.

[Boyen and Koller 1998] Boyen, X., and Koller, D. 1998. Tractable inference for complex stochastic processes. In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, 33–42. Morgan Kaufmann Publishers Inc.

[Braun and Möller 2016] Braun, T., and Möller, R. 2016. Lifted Junction Tree Algorithm. In Proceedings of the Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz), 30–42. Springer.

[Braun and Möller 2018] Braun, T., and Möller, R. 2018. Parameterised Queries and Lifted Query Answering. In Proceedings of IJCAI 2018, 4980–4986.

[Ester et al. 1996] Ester, M.; Kriegel, H.-P.; Sander, J.; and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd, 226–231. AAAI Press.

[Fisher 1925] Fisher, R. 1925. Statistical Methods for Research Workers. Edinburgh Oliver & Boyd.

[Gehrke, Braun, and Möller 2018] Gehrke, M.; Braun, T.; and Möller, R. 2018. Lifted Dynamic Junction Tree Algorithm. In Proceedings of the 3rd International Conference on Conceptual Structures, 55–69. Springer.

[Gehrke, Braun, and Möller 2019] Gehrke, M.; Braun, T.; and Möller, R. 2019. Relational Forward Backward Algorithm for Multiple Queries. In Proceedings of the 32nd International Florida Artificial Intelligence Research Society Conference (FLAIRS-19), 464–469. AAAI Press.

[Geier and Biundo 2011] Geier, T., and Biundo, S. 2011. Approximate Online Inference for Dynamic Markov Logic Networks. In Proceedings of the 23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI), 764–768. IEEE.

[Lauritzen and Spiegelhalter 1988] Lauritzen, S. L., and Spiegelhalter, D. J. 1988. Local Computations with Graphical Structures and their Application to Ex-pert Systems. Journal of the Royal Statistical Society. Series B (Methodological) 50(2):157–224.

[Milch et al. 2008] Milch, B.; Zettlemoyer, L. S.; Kersting, K.; Haines, M.; and Kaelbling, L. P. 2008. Lifted Probabilistic Inference with Counting Formulas. In Proceedings of AAAI, volume 8, 1062–1068.

[Murphy 2002] Murphy, K. P. 2002. Dynamic Bayesian Networks: Representation, Inference and Learning. Ph.D. Dissertation, University of California, Berkeley.

[Niepert and Van den Broeck 2014] Niepert, M., and Van den Broeck, G. 2014. Tractability through exchangeability: A new perspective on efficient probabilistic inference. In AAAI, 2467–2475.

[Papai, Kautz, and Stefankovic 2012] Papai, T.; Kautz, H.; and Stefankovic, D. 2012. Slice Normalized Dynamic Markov Logic Networks. In Proceedings of the Advances in Neural Information Processing Systems, 1907–1915.

[Poole 2003] Poole, D. 2003. First-order probabilistic inference. In Proceedings of IJCAI, volume 3, 985–991.

[Schubert et al. 2017] Schubert, E.; Sander, J.; Estep, M.; Kriegel, H. P.; and Xu, X. 2017. Dbscan revisited, revisited: why and how you should (still) use dbscan. ACM Transactions on Database Systems (TODS) 42(3):19.

[Singla and Domingos 2008] Singla, P., and Domingos, P. M. 2008. Lifted first-order belief propagation. In AAAI, volume 8, 1094–1099.

[Singla, Nath, and Domingos 2014] Singla, P.; Nath, A.; and Domingos, P. M. 2014. Approximate lifting techniques for belief propagation. In AAAI, 2497–2504.

[Taghipour et al. 2013] Taghipour, N.; Fierens, D.; Davis, J.; and Blockeel, H. 2013. Lifted Variable Elimination: Decoupling the Operators from the Constraint Language. Journal of Artificial Intelligence Research 47(1):393–439.

[Van den Broeck and Darwiche 2013] Van den Broeck, G., and Darwiche, A. 2013. On the complexity and approximation of binary evidence in lifted inference. In Advances in Neural Information Processing Systems, 2868–2876.

[Van den Broeck and Davis 2012] Van den Broeck, G., and Davis, J. 2012. Conditioning in first-order knowledge compilation and lifted probabilistic inference. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, 1–7. AAAI Press.

[Van den Broeck and Niepert 2015] Van den Broeck, G., and Niepert, M. 2015. Lifted probabilistic inference for asymmetric graphical models. In AAAI, 3599–3605.

[Venugopal and Gogate 2014] Venugopal, D., and Gogate, V. 2014. Evidence-based clustering for scalable inference in markov logic. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 258–273. Springer.