

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). 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.

Boyan 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

even under evidence. Further, Boyen and Koller show that, for any approximation scheme of belief state representations, the error contracts exponentially as the process evolves, making the introduced error bounded indefinitely (Boyen and Koller 1998). Their approach and TAME are related as in both cases evidence makes inference infeasible. However, TAME aims at *automatically* restoring a lifted representation. In summary, the cause, namely evidence, is the same for both problems but the problems are different and the means to make inference possible again differ highly.

TAME is applicable in different formalisms and algorithms. However, we discuss TAME as part of LDJT for two reasons: First, when advancing in time, LDJT computes a minimal message that is the source of the most splits of the next time step. Applying TAME on this message tackles the KRP problem at its root. Second, using TAME with an exact algorithm allows for attributing errors to merging rather than imprecisions during reasoning. Additionally, TAME is deterministic in its approximation, thereby, avoiding problems with sampling rates or ergodicity. Empirical results show that TAME significantly improves performances of LDJT, while keeping errors small and attributable to merging.

In the following, we recapitulate parameterised probabilistic dynamic models (PDMs) as a formalism for specifying temporal probabilistic relational models and LDJT for efficient query answering in PDMs. Then, we present TAME, which includes clustering, significance checks, and merging. Lastly, we evaluate TAME theoretically and empirically.

Preliminaries

We shortly present parameterised probabilistic models (PMs) (Braun and Möller 2018), then extend PMs to the temporal case, resulting in PDMs, and recapitulate LDJT (Gehrke, Braun, and Möller 2018; 2019), a *smoothing*, *filtering*, and *prediction* algorithm for PDMs.

Parameterised Probabilistic Models

PMs combine first-order logic with probabilistic models, using logical variables (logvars) as parameters to represent sets of indistinguishable constructs. As an example, we set up a PM to model the reputation of researchers, inspired by the competing workshop example (Milch et al. 2008), with a logvar representing researchers. A reputation is influenced by activities such as publishing, doing active research, and attending conferences. A randvar parameterised with logvars forms a parameterised randvar (PRV).

Definition 1. Let \mathbf{R} be a set of randvar names, \mathbf{L} a set of logvar names, Φ a set of factor names, and \mathbf{D} a set of constants (universe). All sets are finite. Each logvar L has a domain $\mathcal{D}(L) \subseteq \mathbf{D}$. A *constraint* is a tuple $(\mathcal{X}, C_{\mathbf{X}})$ of a sequence of logvars $\mathcal{X} = (X^1, \dots, X^n)$ and a set $C_{\mathcal{X}} \subseteq \times_{i=1}^n \mathcal{D}(X_i)$. The symbol \top for C marks that no restrictions apply, i.e., $C_{\mathcal{X}} = \times_{i=1}^n \mathcal{D}(X_i)$. A PRV $R(L^1, \dots, L^n)$, $n \geq 0$ is a syntactical construct of a randvar $R \in \mathbf{R}$ possibly combined with logvars $L^1, \dots, L^n \in \mathbf{L}$. If $n = 0$, the PRV is parameterless and forms a propositional randvar. A PRV A or logvar L under constraint C is given by $A|_C$ or $L|_C$, respectively. We may omit \top in $A|_{\top}$ or $L|_{\top}$. The term $\mathcal{R}(A)$ denotes the

possible values (range) of a PRV A . An *event* $A = a$ denotes the occurrence of PRV A with range value $a \in \mathcal{R}(A)$.

We use the randvar names A , D , R , and Pub for attends conference, does research, reputation, and publishes in journals, respectively, and $\mathbf{L} = \{X, J\}$ with $\mathcal{D}(X) = \{x_1, x_2, x_3\}$ (people) and $\mathcal{D}(J) = \{j_1, j_2\}$ (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 $\phi(\mathcal{A})|_C$ with $\mathcal{A} = (A^1, \dots, A^n)$ a sequence of PRVs, $\phi : \times_{i=1}^n \mathcal{R}(A^i) \mapsto \mathbb{R}^+$ a function with name $\phi \in \Phi$, and C a constraint on the logvars of \mathcal{A} . We may omit \top in $\phi(\mathcal{A})|_{\top}$. 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 $G_{ex} = \{g^i\}_{i=0}^1$, 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 \mathbf{A}_{t-1} and \mathbf{A}_t where \mathbf{A}_{π} a set of PRVs from time slice π . 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}^{ex} consisting of G^{ex} for time slice $t-1$ and t 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 $\mathbf{E}_{0:t} = \{E_t^i = e_t^i\}_{i,t}$, the expression $P(Q_t | \mathbf{E}_{0:t})$ denotes a *query* w.r.t. P_G .

The problem of answering a query $P(A_{\pi}^i | \mathbf{E}_{0:t})$ 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

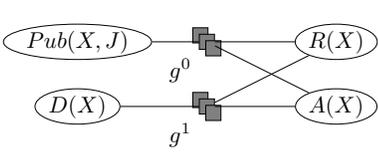


Figure 1: Parfactor graph for G^{ex}

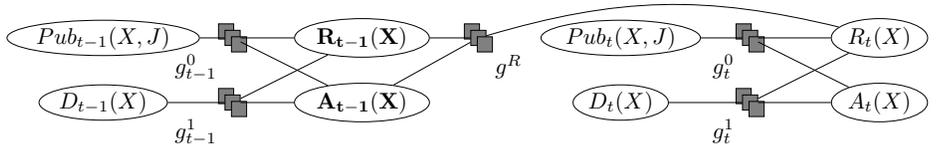


Figure 2: G_{\rightarrow}^{ex} the two-slice temporal parfactor graph for model G^{ex}

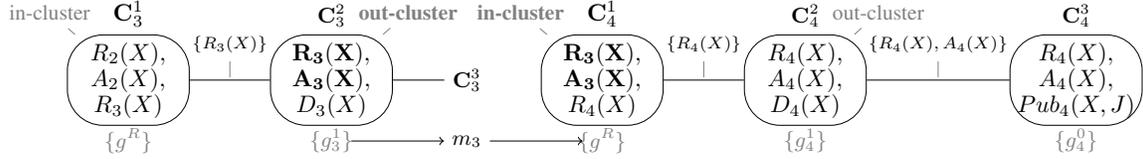


Figure 3: FO jtree J_3 without C_3^3 and FO jtree J_4 connected with m_3

from each other. Let us now define an FO jtree, with parameterised clusters (parclusters) as nodes, and present how LDJT proceeds in time.

Definition 5. Let \mathbf{X} be a set of logvars, \mathbf{A} a set of PRVs with $lv(\mathbf{A}) \subseteq \mathbf{X}$, and C a constraint on \mathbf{X} . Then, $\mathbf{A}|_C$ denotes a *parcluster*. We omit $|C$ if $C = \top$ and $lv(\mathbf{A}) = \mathbf{X}$. 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 $\phi(\mathbf{A})|_C$ in G , \mathbf{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^{ij} , called *separator* of edge $\{i, j\} \in E$, is defined by $C^i \cap C^j$. Each $C^i \in V$ has a *local model* G^i and $\forall g \in G^i: rv(g) \subseteq C^i$. 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 \mathbf{I}_t of G_{\rightarrow} .

Definition 6. The forward interface \mathbf{I}_{t-1} is given by

$$\mathbf{I}_{t-1} = \{A_t^i \mid \exists \phi(\mathbf{A})|_C \in G : A_{t-1}^i \in \mathcal{A} \wedge \exists A_t^j \in \mathcal{A}\}.$$

PRVs $R_{t-1}(X)$ and $A_{t-1}(X)$ from G_{\rightarrow}^{ex} , shown in Fig. 2, make up \mathbf{I}_{t-1} . While constructing FO jtree structures, LDJT ensures that the FO jtree J_t for time step t has a parcluster containing \mathbf{I}_{t-1} , which is called *in-cluster*, and a parcluster containing \mathbf{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 \mathbf{I}_t using the *out-cluster* of J_t . Hence, m_t contains exactly the necessary information, as a set of parfactored, 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_3^2 and saves the result in message m_3 . Increasing t by one, LDJT adds m_3 to C_4^1 '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 ϕ 's of parfactored 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 ϕ '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 $G_t = \{g_t^i\}_{i=1}^n$ at time step t , many parfactored 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:

$$\{g_t^{i,1}, \dots, g_t^{i,m}\}_{i=1}^n, m \in \mathbb{N}^+. \quad (1)$$

For each i , the different $g_t^{i,j} = \phi_t^{i,j}(\mathcal{A}^i)|_{C^{i,j}}$, $1 \leq j \leq m$, have the same arguments \mathcal{A}^i but different constraints $C^{i,j}$ and varying functions $\phi_t^{i,j}$ as a result of evidence. The assumption is that some $g_t^{i,j}$ have similar ϕ '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 ϕ '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 $g_t^{i,j}$ with similar ϕ 's into one parfactor

$$g_t^{i,k} = \phi_t^{i,k}(\mathcal{A}^i)|_{C^{i,k}} \quad (2)$$

where $\phi_t^{i,k}$ represents a merged version of the combined $\phi_t^{i,j}$ and $C^{i,k}$ is a union of the combined $C^{i,j}$. Merging all par-factors that behave similarly for each i leads to a G'_t of the following form with parfactors as in Eq. (2) and $l < m$:

$$\{g_t^{i,1}, \dots, g_t^{i,l}\}_{i=1}^n \quad (3)$$

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_t , 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_t . 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 ϵ 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 \mathbf{P} of sets of parfactors. Each partition $P \in \mathbf{P}$ has a set of logvars \mathbf{X}_p that has been affected in the same way by splitting due to evidence. Formally, P has the form

$$P = \{g_t^{i,1}, \dots, g_t^{i,m}\}_{i=1}^{n_p} \quad (4)$$

with $lv(g_t^{i,j}) \subseteq \mathbf{X}_p$. The next step is to identify groups of parfactors in each partition that behave similarly.

Parfactor Clustering After partitioning G , each partition $P \in \mathbf{P}$ of the form in Eq. (4) has parfactors whose constraints overlap between all i for each j . Therefore, TAME multiplies all parfactors with overlapping constraints into one parfactor before starting with identifying groups. If $lv(g_t^{i,j}) = \mathbf{X}_p$, 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

$$P = \left\{ \prod_{i=1}^{n_p} g_t^{i,1}, \dots, \prod_{i=1}^{n_p} g_t^{i,m} \right\} = \{g_t^{p,1}, \dots, g_t^{p,m}\} \quad (5)$$

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 neighbourhoods, determined by a radius ϵ around a point, lie a certain number *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 *minPts* to 2 to be able to cluster even two parfactors. The distance measure should assess how similarly parfactors behave, with 0 meaning identical behaviour and larger values meaning less similar behaviour.

To determine the similarity of the behaviour of two parfactors, one could calculate marginal distributions for a PRV that occurs with split constraints and compare if the marginals are in a certain δ area. However, marginal distributions could result from completely different potentials

Algorithm 1 Temporal Approximate Merging

procedure TAME(Model G , Radius ϵ , Significance α)
 $\mathbf{P} \leftarrow$ partitioning of G based on logvars \triangleright Eq. (4)
for each partition $P \in \mathbf{P}$ **do**
 $P \leftarrow$ multiply overlapping parfactors \triangleright Eq. (5)
 $\mathbf{K} \leftarrow$ DBSCAN($P, \epsilon, 2, rsim$)
 if ANOVA($\mathbf{K}, rsim', \alpha$) rejects H_0 **then**
 $G \leftarrow G \setminus P$
 for each cluster $K \in \mathbf{K}$ **do**
 $G \leftarrow G \cup \{K \text{ merged}\}$ \triangleright Eq. (7)

and be similar by chance. The potentials of a parfactor on 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* would be 0.667 and 0.675 respectively, i.e., less than 0.01 apart from each other. The same case arises for two parfactors mapping to $\langle 4, 2 \rangle$ and $\langle 4.1, 1.9 \rangle$ 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 $g_t^{i,j_1} = \phi_t^{i,j_1}(\mathcal{A}^i)_{|C^{i,j_1}}$ and $g_t^{i,j_2} = \phi_t^{i,j_2}(\mathcal{A}^i)_{|C^{i,j_2}}$, we use a function *rsim* : $(\times_{i=1}^n \text{range}(A^i) \mapsto \mathbb{R}^+, \times_{i=1}^n \text{range}(A^i) \mapsto \mathbb{R}^+) \mapsto \mathbb{R}^+$ that is defined as follows:

$$rsim(\phi_t^{i,j_1}, \phi_t^{i,j_2}) = 1 - \frac{\sum_{\mathbf{a} \in \text{range}(A^i)} \phi_t^{i,j_1}(\mathbf{a}) \cdot \phi_t^{i,j_2}(\mathbf{a})}{\sqrt{\sum_{\mathbf{a} \in \text{range}(A^i)} \phi_t^{i,j_1}(\mathbf{a})^2} \cdot \sqrt{\sum_{\mathbf{a} \in \text{range}(A^i)} \phi_t^{i,j_2}(\mathbf{a})^2}} \quad (6)$$

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, ϵ needs to be ≤ 1 . Overall, the inputs of DBSCAN for clustering parfactors are a partition P of parfactors, ϵ , $minPts = 2$, and *rsim*. ϵ trades off cluster sizes with accuracy. The output is a clustering (partitioning) of P , i.e., a set \mathbf{K} of sets in which each $K \in \mathbf{K}$ 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 ϵ , 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 shattered, 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) (Fisher 1925) is a statistical method to test for significance of a clustering. In our setup, ANOVA computes the variance of each parfactor in a cluster $K \in \mathbf{K}$ 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 \mathbf{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 $\{\phi_t^{i,j}(\mathcal{A}^i)_{|C^{i,j}}\}_{j=1}^m$, a mean parfactor $g_t^{i,k} = \phi_t^{i,k}(\mathcal{A}^i)_{|C^{i,k}}$ is determined by

$$\phi_t^{i,k}(\mathbf{a}) = \frac{\sum_{j=1}^m gr(\phi_t^{i,j}(\mathbf{a})_{|C^{i,j}}) \phi_t^{i,j}(\mathbf{a})}{gr(\phi_t^{i,k}(\mathbf{a})_{|C^{i,k}})} \quad (7)$$

for each $\mathbf{a} \in \text{range}(A^i)$ 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 groundings, the second maps to 3.9 and 1.9 with 5 groundings, 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 groundings.

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 often and therefore, in the ground case the variance would be calculated more often.

After computing a mean parfactor $g_t^{i,k}$ for each cluster $K \in \mathbf{K}$ and an overall mean parfactor $g_t^{i,m}$ based on all parfactors in \mathbf{K} , ANOVA proceeds to compute the variation between groups, i.e., *MSG*, and within groups, i.e., *MSE*, using Eq. (6) and the groundings of parfactors:

$$MSG = \frac{1}{l-1} \sum_{K \in \mathbf{K}} gr(g_t^{i,k}) \cdot (rsim(g_t^{i,k}, g_t^{i,m}))^2$$

$$MSE = \frac{1}{m-l} \sum_{K \in \mathbf{K}} \sum_{g_t^{i,j} \in K} gr(g_t^{i,j}) \cdot (rsim(g_t^{i,j}, g_t^{i,k}))^2$$

where $l = |\mathbf{K}|$, i.e., number of clusters, and $m = |gr(\mathbf{K})|$, i.e., number of overall groundings. Computing $F = \frac{MSG}{MSE}$,

ANOVA compares F against a critical value F_{crit} , which depends on α , $l - 1$, and $m - l$ and can be looked up in a pre-computed table. If $F \leq F_{crit}$, TAME accepts H_0 and discards the clustering. In case TAME rejects H_0 , i.e., $F > F_{crit}$, there is more difference between clusters than within clusters and TAME proceeds to merging parfactors.

Merging Parfactors The new parfactor for each cluster $K \in \mathbf{K}$ 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 ϕ '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 $\gamma_Q \in]0, 1]$ (Boyen and Koller 1998). γ_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 \mathbf{I}_t and LDJT computes the transition from t to $t + 1$. Thus, the approximation error δ is reduced by the factor $(1 - \gamma_Q)$ with each transition. Assuming, that TAME introduces an error of at most δ for each time step, the expected error up to time step t accumulates to $\delta + (1 - \gamma_Q) \cdot \delta + \dots + (1 - \gamma_Q)^{t-1} \cdot \delta =$

$\sum_{i=0}^t \delta \cdot (1 - \gamma_Q)^i \leq \sum_{i=0}^{\infty} \delta \cdot (1 - \gamma_Q)^i = \delta / \gamma_Q$. For the last step, we apply the geometric series, i.e., $\sum_{i=0}^{\infty} \delta \cdot (1 - \gamma_Q)^i = \delta / (1 - (1 - \gamma_Q)) = \delta / \gamma_Q$ (Boyen and Koller 1998). Thus, the error is indefinitely bounded by δ / γ_Q . \square

For TAME the significance check influences the approximation error δ . 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 δ . 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 γ_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. \square

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 $t + 1$, γ_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. \square

Thus, TAME solves the KRP problem. Since the underlying distributions of ϕ 's converge, TAME is able to merge ϕ '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^{ex} 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 to 10 symmetry groups and generate evidence for 20 time steps. For each symmetry group i , LDJT answers $A_{t+\pi}(x_i)$ for $\pi = \{0, 5, 10\}$ in each time step t for all 20 time steps.

We vary ϵ and the interval I of how often LDJT applies TAME. α is fixed to 0.005. Based on the problem at hand, an appropriate α needs to be determined in advance (Benjamin et al. 2018). The three options we evaluate are, from conservative to aggressive: 1) $I = 5$, $\epsilon = 5 \cdot 10^{-14}$, 2) $I = 5$, $\epsilon = 5 \cdot 10^{-2}$, and 3) $I = 2$, $\epsilon = 5 \cdot 10^{-2}$. TAME with Option 1 mostly merges parfactors that only differ in a scaling factor. TAME with Options 2 and 3 also merges parfactors that

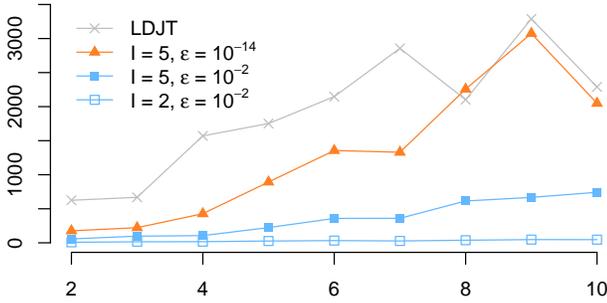


Figure 4: Runtimes [seconds], x-axis: #symmetry groups

π	Max	Min	Average
0	0.0001537746121	0.0000000001720	0.0000191206488
2	0.0000000851654	0.0000000000001	0.0000000111949
4	0.0000000000478	0	0.0000000000068

Table 1: Introduced error; $I = 2$, $\epsilon = 5 \cdot 10^{-2}$, 10 groups

slightly differ in their ratio. With $I = 2$, LDJT calls TAME every other time step, and with $I = 5$ every fifth time step.

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, evidence can ground the model faster over time. Thus, the runtimes correlate to the number of groups. For 5 symmetry groups, LDJT without TAME takes about twice as long as LDJT with TAME using the conservative option (1), answering 300 queries for the 20 time steps. However, for 8 symmetry groups, LDJT without TAME is slightly faster. As merging depends on evidence, which here is randomly generated, TAME may not always be able to trade off its overhead. TAME with Option 2 merges more parfactors. Hence, every fifth time step, LDJT answers queries on fewer groups, which are then again split up by evidence. With the most aggressive option (3), LDJT applies TAME every other time step and thus answers queries on highly lifted models.

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 1 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

	Max	Min	Average
w	0.0002259927071	0.0000000000000	0.0000104567643
w/o	0.0002260554389	0.0000000000168	0.0000137870835

Table 2: Introduced error; with and without significance test

the significance check. Table 2 shows the introduced errors for these runs. The maximum error hardly differs between the two runs, which 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, 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, D. J.; Berger, J. O.; Johannesson, M.; Nosek, B. A.; Wagenmakers, E.-J.; Berk, R.; Bollen, K. A.; Brembs, B.; Brown, L.; Camerer, C.; et al. 2018. Redefine statistical significance. *Nature Human Behaviour* 2(1):6.
- 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, 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, T., and Möller, R. 2018. Parameterised Queries and Lifted Query Answering. In *Proceedings of IJCAI 2018*, 4980–4986.
- 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, R. 1925. *Statistical Methods for Research Workers*. Edinburgh Oliver & Boyd.
- Gehrke, M.; Braun, T.; and Möller, R. 2018. Lifted Dynamic Junction Tree Algorithm. In *Proceedings of the 23rd International Conference on Conceptual Structures*, 55–69. Springer.
- 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, 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, S. L., and Spiegelhalter, D. J. 1988. Local Computations with Probabilities on Graphical Structures and their Application to Expert Systems. *Journal of the Royal Statistical Society. Series B (Methodological)* 50(2):157–224.
- Milch, B.; Zettlemoyer, L. S.; Kersting, K.; Haimes, M.; and Kaelbling, L. P. 2008. Lifted Probabilistic Inference with Counting Formulas. In *Proceedings of AAAI*, volume 8, 1062–1068.
- Murphy, K. P. 2002. *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph.D. Dissertation, University of California, Berkeley.
- Niepert, M., and Van den Broeck, G. 2014. Tractability through exchangeability: A new perspective on efficient probabilistic inference. In *AAAI*, 2467–2475.
- 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, D. 2003. First-order probabilistic inference. In *Proceedings of IJCAI*, volume 3, 985–991.
- Schubert, E.; Sander, J.; Ester, 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, P., and Domingos, P. M. 2008. Lifted first-order belief propagation. In *AAAI*, volume 8, 1094–1099.
- Singla, P.; Nath, A.; and Domingos, P. M. 2014. Approximate lifting techniques for belief propagation. In *AAAI*, 2497–2504.
- 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, 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, 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, G., and Niepert, M. 2015. Lifted probabilistic inference for asymmetric graphical models. In *AAAI*, 3599–3605.
- 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.