Seventh IEEE International Conference on Data Mining - Workshops

Granularity Conscious Modeling for Probabilistic Databases

Eirinaios Michelakis[†] Daisy Zhe Wang[†] Minos Garofalakis[†‡] Joseph M. Hellerstein[†]

[†]UC Berkeley EECS Dept.
Berkeley, CA 94720
{ireneos, daisyw, hellerstein}@cs.berkeley.edu

[‡]Yahoo! Research
Santa Clara, CA 95054
minos@yahoo-inc.com

Abstract

*The convergence of embedded sensor systems and stream query processing suggests an important role for database techniques, in managing data that only partially – and often inaccurately – capture the state of the world. Reasoning about uncertainty as a first class citizen, inside a database system, becomes an increasingly important operation for processing non deterministic data. An essential step for such an approach lies in the choice of the appropriate uncertainty model, that captures the probabilistic information in the data, both accurately and at the right semantic detail level. This paper introduces Hierarchical First-Order Graphical Models (*HFGMs*), an intuitive and economical representation of the data correlations stored in a Probabilistic Data Management system, in a hierarchical setting.* HFGM *semantics allow for an efficient summarization of the probabilistic model that can be induced from a dataset at various levels of granularity, effectively controlling the trade-off of the model's complexity vs its accuracy.*

1 Introduction

Many real-world applications need to process large quantities of inherently uncertain data. Noise, incomplete information, or statistical models that fail to adequately represent the underlying data distribution, are typical instances of uncertainty that applications have to deal with.

The area of Machine Learning has lent itself quite naturally to the task of modeling uncertainty. The major shortcoming of ML techniques has been their inability to scale to large datasets. While data mining algorithms attempted historically to address the scalability issues, they exhibited limited integration with the traditional query processing infrastructure that was employed by commercial data management systems. The result of their execution – often a specialized statistical model – was rarely made visible to the data processing architecture, but rather fed to customary statistical analysis modules outside the data management system, that knew how to interpret these output models.

During the last decade, a small number of Probabilistic Database research projects attempted to incorporate and query uncertain data in a relational database [3, 1]. In an effort though to simplify the semantics and the query processing involved, they made the unrealistic assumption of *unconditional independence* between the data items.

On the modeling front, the representation level commonly used to model the data values per se, i.e. *per tuple*, often results in limiting the applicability of the Probabilistic Database. According to that uncertainty model, every tuple has a probability of belonging to the table associated with it. Although this approach offers an intuitive representation of the probability distribution of the probabilistic database's possible instances (*possible worlds semantics*), it tends to model the data items in the finest detail level, which makes the system highly impractical to use, in a realistic setting.

Some fairly recent approaches attempted to address the limitations imposed by the oversimplifying independence assumption, by modeling explicitly the correlations among the stored entities. Sen et al. [4] make use of statistical models to achieve this goal, maintaining though the tuple-level semantics of their predecessors. The resulting models end up containing *one random variable per tuple*, either explicitly stored in the database, or created by a join of two or more tables. As a result, the usability of such models while learning the correlations between the random variables, or even inferencing for query processing purposes, gets restricted to small sized databases.

This work considers a rather orthogonal dimension of the statistical model construction process, which has to do with the *granularity* in which the model entities are represented. In the rest of the paper we focus on an a particular class of statistical models, called *Probabilistic Graphical Models* [2]. Their power lies in their ability to capture the joint probability distribution of the domain entities, in an intuitive and compact form, by factorizing them into conditionally independent subsets. The participating entities are represented as *Random Variables* (RVs) of the stochastic process with which the problem domain is simulated.

0-7695-3019-2/07 $25.00 © 2007 IEEE
DOI 10.1109/ICDMW.2007.63

Unlike the majority of the most popular statistical models (classification, clustering, association rules), in graphical models the entity selection process is rarely automated, but rather guided by expert knowledge. When the detail level of the RVs is refined, the resulting graphical model ends up being composed of a large set of fine-grained RVs. Consequently, the degrees of freedom in modeling the underlying data distribution increase, and so does the complexity of the graphical model, and with it the time to learn or perform inference on it (both problems are known to be NP-hard [2]). Furthermore, although its fitness to the training dataset tends to be high, this often leads to overly optimistic estimates of the model's accuracy. The latter is then said to *overfit* the training data, and as a result, its generalization ability to unobserved data instances gets reduced.

This paper introduces an economical and intuitive representation of the correlation among the data stored in a probabilistic data management system. Given a schema and the probabilistic entities of a domain, we define the class of *Hierarchical First-Order Graphical Models* (or HFGMs for short) as a hierarchical extension of Probabilistic Graphical Models. HFGMs utilize a combination of *first-order* and *aggregate* semantics, to model the correlations among the probabilistic entities of a domain, in the granularity that captures the underlying data distribution as faithfully as possible. First order graphical modeling in the context of databases has been addressed in [5], but that approach was not able to capture correlations among different instances of the same entity.

HFGMs are based on a hierarchy of RVs, which is formed by partitioning the instances of the probabilistic database, according to a set of "dimension" attributes. A graphical model is derived for each level of the hierarchy, that captures the correlations among the probabilistic attributes (or RVs), as expressed in that particular level of granularity. We formally define HFGMs and explain how their semantics accelerate the process of deriving a finer-grained graphical model when its coarser-grained parents are already constructed, by *inheriting* common structure and parameters. We finally touch upon the issues involved in learning an HFGM from a probabilistic database.

2 The Digital Home

In this section we introduce a running example to which we will refer throughout the paper. Source of our inspiration has been the *Digital Home*, a vision of a house, equipped with sensors, able to interpret the activities of its occupants and enhance their living quality.

Let us assume that we have a set of environmental sensors, deployed at various places in a house (see Figure 1). Sensors S_1 to S_6 measure temperature, voltage and light levels, reporting their readings as integers in the range of $[1, 100]$, appropriately quantized. We also consider four RFID readers (RF_1 to RF_4), used to report the number of occupants detected within their effective detection range.

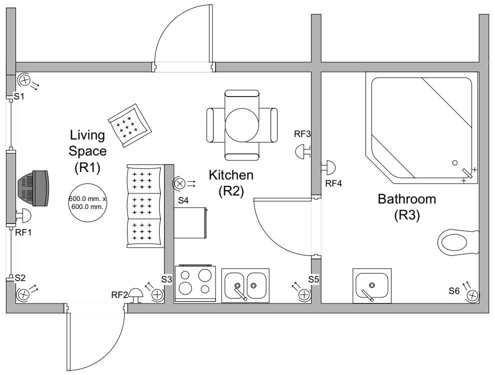

Figure 1. Floor plan of a Digital Home.

EnvirReadings(sID, time, temp$^\rho$, light$^\rho$, volt$^\rho$)
RFIDReadings(sID, time, numOccupants$^\rho$)
Sensors(sID, type)
SensorLocation(sID, room, xCoord, yCoord)

Figure 2. Schema of the Digital Home DB.

All these readings are stored in a relational database, with its schema shown in Figure 2. Primary keys have been underlined, and foreign keys can be easily identified, as they share the same names across tables. Attributes with the superscript ρ correspond to "probabilistic" attributes, encoding measurable quantities that are uncertain in our world. The rest are treated as deterministic information.

Due to the proximity of the sensors and the various characteristics of our floor plan, a lot of correlation patterns could be derived between the sensed quantities. For instance, since S_1 and S_2 are so close to the windows, it is not unreasonable to expect that during daytime, the temperature readings of those sensors to be slightly higher than those collected from S_3 or S_4. Furthermore, at a specific point in time, the temperature and light levels across the living room and the kitchen should present small variability, while at all times, temperature and voltage levels in all sensors are expected to be positively correlated.

These interesting patterns can be adequately represented by a Bayesian or a Markov Network, which can give probabilistic estimates of the sensor values at various locations in the room. The question that arises is the granularity with which these correlations are represented. The temperature–voltage and temperature–light correlations hold *for each pair* of readings in the living room and kitchen (modulo some exceptions – e.g. low light levels and increased temperature reported from S_5, since its closer to the oven), but not in the bathroom. If the number of such exceptional cases is small, instead of selecting as RVs the individual sensor readings, we could model the above patterns in a coarser granularity, at the room level. Abstracting even further, if a particular correlation pattern is consistent across rooms, we could capture it among the *averaged values* of, say, temperature–light reading pairs of *all the sensors in the house*, at a specific point in time.

HFGMs attempt to address the trade-off of expressibility vs model accuracy, by defining RVs (and thus Graphical Models) at various levels of granularity; the models at finer levels refine the probability distribution modeled at coarser level models, with respect to fitting a pool of training data. The next section gives a formal definition of HFGMs and their semantics.

3 Hierarchical FO Graphical Models

A HFGM is composed of two parts: a hierarchy of random variables (RVs) $\mathcal{H}$ corresponding to the probabilistic attributes of the domain, and a number of layered First-Order Graphical Models $\mathcal{M}^* = \{\mathcal{M}_1, \ldots, \mathcal{M}_n\}$, which describe the dependencies among the RVs of $\mathcal{H}$. The RVs at each $\mathcal{M}_i$ reside in the same hierarchical level in $\mathcal{H}$. Before introducing the HFGM formalism, we will briefly define our notion of a Probabilistic Relational Database.

A Probabilistic Relational Schema $\mathcal{S}^p$ is an extension of a Relational Schema that identifies for each participating relation $\mathcal{R} = \{\mathcal{A}_1, \ldots, \mathcal{A}_n\}$, a non-empty subset $\mathcal{A}^d \in \mathcal{R}$ of *deterministic* and a subset $\mathcal{A}^p \in \mathcal{R}$ of *probabilistic* attributes. For the deterministic attributes of $\mathcal{R}$ there is no uncertainty concerning their values. In that respect, their functionality is very similar to the the dimension attributes of the multidimensional OLAP paradigm. For the probabilistic attributes, it is assumed that their values are *uncertain*; thus, they can be seen as instances or measurements, taken from random experiments of a *stochastic process*.

Thus, we formally define a *Probabilistic Database* with schema $\mathcal{S}^p$, as a set of *independent and identically distributed* (IID) trials of a *stochastic process*. This interpretation allows us to group the values of the probabilistic attributes in the database into a number of distinct, independent *trials* of this process.

3.1 Random Variables Hierarchy $\mathcal{H}$

A *domain* $\mathcal{D}$ consists of a set of basic entities, $\mathcal{E}^* = \{\mathcal{E}_1, \ldots, \mathcal{E}_l\}$, representing the probabilistic attributes $\{\mathcal{A}^p_1, \ldots, \mathcal{A}^p_l\}$ of a Probabilistic Relational Schema $\mathcal{S}^p$. Each entity $\mathcal{E}_i$ can be further refined into a *family* of entities of various granularities, $\mathcal{E}_i = \{\mathcal{E}^1_i, \ldots, \mathcal{E}^k_i\}$, that share the same semantics (i.e. model the same physical world quantity), but in different levels of detail. An entity $\mathcal{E}^j_i$ can participate in a graphical model in the form a random variable (RV), $\mathcal{R}^j_i$, characterized by a set of domain values $\mathcal{V}_i$, common to all the RVs of the same family.

As for the deterministic attributes $\mathcal{A}^d$, each of them can be associated with a *set of domains*, that specify values for that attribute at different levels of granularity. An ordering of these domains, from the coarsest to the finest level, defines a *dimension chain*, which can be used to arrange the allowable values of the attribute into a *hierarchy* (see Figure 3(a) for an example). Each node in this hierarchy specifies a *unique horizontal partition* of the tuples of that table. We call the value of each node a *partition key*.

We can model the data falling in these horizontal partitions probabilistically, by assigning to each node of the hierarchy, a RV per probabilistic attribute, which captures the values that correspond to the latter at the level of granularity that this node resides in the hierarchy. Thus, considering a single dimension attribute[1] and k probabilistic ones, will give rise to k identical RV hierarchies, each corresponding to one of the probabilistic attributes. Each such hierarchy is associated with a particular family of RVs, and hence, its nodes are the RVs that belong in the family.

More formally, a *Hierarchy* $\mathcal{H}(\mathcal{R}_i)$ of a family of RVs $\mathcal{R}_i$ defines a partial ordering relationship $\prec$ between its member RVs. For each $\mathcal{R}^j_i \in \mathcal{R}_i$, $\mathcal{C}(\mathcal{R}^j_i)$ denotes the set of the immediate subclasses (or *children*) of $\mathcal{R}^j_i$, as defined by the partial ordering $\prec$. $x \prec_{\mathcal{C}} y$, denotes that x is an immediate subclass of y, while with $\prec^*$ we symbolize the transitive closure of $\prec$; if $x \prec y$, x is a (not necessarily immediate) subclass of y.

The *scope* $\mathcal{S}(\mathcal{R}^j_i)$ of the RV $\mathcal{R}^j_i$ specifies its value *for every independent trial* (or instance) stored in the database. This value is an aggregate over the values of its immediate children $\mathcal{R}^k_i$, $\mathcal{R}^k_i \in \mathcal{C}(\mathcal{R}^j_i)$, that fall in the horizontal partition of the hierarchical level l in which $\mathcal{R}^j_i$ belongs. More formally, we define $\mathcal{S}(\mathcal{R}^j_i)$ in terms of Relational Algebra operators as: $\mathcal{S}(\mathcal{R}^j_i) = \gamma(\sigma(< \textit{list of predicates} >), \mathcal{A}_{id})$, where γ represents the aggregation operator, σ an equality selection over the list of dimension attributes which define $\mathcal{R}^j_i$, and $\mathcal{A}_{id}$ the deterministic attribute used to identify the independent trials of the stochastic process that the probabilistic database models.

We conclude this definition section with the notions of *compatibility* and *identity*, which are key to HFGM semantics. Two hierarchies are *compatible* if they share the same dimension chain. Two hierarchies are *identical* if they are compatible and share the same tree structure (ie. there is a one-to-one mapping between their RVs at each level, and the partition key of each RV pair is the same).

[1] For simplicity of exposition, we consider a single dimension attribute per probabilistic relation. An extension of this scheme to the multidimensional model introduced by the OLAP paradigm, represents an interesting future work direction.

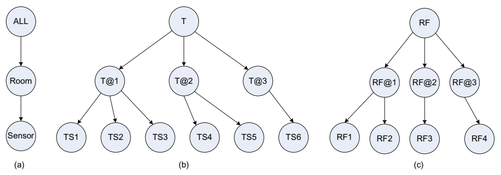

Figure 3. The hierarchy of the RVs of the Digital Home domain: (a)The common dimension chain of $\mathcal{H}(T), \mathcal{H}(V), \mathcal{H}(L) \& \mathcal{H}(RF)$; (b) $\mathcal{H}(T)$; (c) $\mathcal{H}(RF)$.

Example In our Digital Home domain (see Fig. 2), the basic probabilistic entities are the temperature (T), sensor voltage (V), light levels (L) and the number of people present (RF) at a specific region of the house. Their RV-equivalents correspond to the roots of four hierarchies of RVs. Their scope consists of all the *aggregate* temperature, voltage, light level readings and person counts respectively, from all the sensors in the house, at the time instances the data collection took place.

At a finer detail level, partitioning *SensorReadings* according to the *sensor's location* would be a reasonable refinement. Each of our base hierarchies would be extended with a second level of RVs, corresponding to the aggregated temperature / voltage / light-level / people-presence readings for a particular room (e.g. for the temperature hierarchy, the RVs $T@1$, $T@2$ and $T@3$ would be created, for the temperature readings in the living room, kitchen and the bathroom respectively, with the partial ordering relationship $\{T@1, T@2, T@3\} \prec_{\mathcal{C}} T$ holding). Further refining the detail of the representation to the sensor level is quite straightforward. Figures 3(b) and 3(c) depict the hierarchies of the T and RF RVs of our example.

The notion of *granularity* of a RV R_i is ultimately realized by *aggregating* at every independent trial the values of its children RVs, $\mathcal{C}(R_i)$. Let $\gamma(P, \mathcal{A}_{id})$ denote the aggregation of partition P on the deterministic attribute set $\mathcal{A}_{id}$. In our example, a stochastic trial is uniquely identified by its time-stamp (attributes *EnvirReadings.time* and *RFIDReadings.time* in our example). Thus, the RV $T@1$ is defined as the average of its children ($TS1, TS2$ and $TS3$), per time instance, thus representing a coarser-level abstraction – a probabilistic summary – of the temperature readings collected in room 1.

3.2 First-Order Graphical Models

The second component of an HFGM is a set of First-Order Graphical Models, each corresponding to a level of the hierarchy of RVs presented in the previous section. A Probabilistic Graphical Model is a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ between a set of RVs (the nodes of the graph), and a set of conditional independence relationships among them, encoded by the edges set $\mathcal{E}$. A First-Order Graphical Model extends the above formalism by encoding dependencies between *populations* of RVs. Its definition is tightly integrated with the existence of a hierarchy of RVs $\mathcal{H}$, in the sense that each RV R_i can have a set of RVs R_i^j that are regarded its descendants, defined at a finer granularity level than R_i in $\mathcal{H}$ – more formally, $R_i^j \in \mathcal{C}(R_i)$.

Two or more RVs can be joined by edges that represent their correlation patterns, only when they belong to the same granularity level of *two compatible* hierarchies (or the same one). Nevertheless, such a correlation can have two distinct interpretations, based on the relation of the connected RVs with respect to their hierarchies:

- **Templating**: If the RVs have *identical* hierarchies, the correlation also holds for the children of the correlated variables (e.g. $T@1 - L@1$ is translated as $TS_i - LS_i$, for $i = 1, \ldots, 3$). The model is said to display "first-order" logic semantics, since each dependency structure is *universally quantified* over its children.

- **Aggregation**: If the RVs have simply *compatible* hierarchies but *different partition keys*, the correlation specified at that level represents an *aggregate* correlation, through which – and *not directly* – the children of these RVs are correlated (e.g. $T@1 - T@2$). No one-to-one correlation (ie. templating) between the children of the RVs involved can be assumed. Thus, in our example, the children of $T@1$ and $T@2$ are correlated implicitly, through their parents, at a coarser granularity level.

First-Order Graphical Models constitute a more expressive and economic representation of the correlation patterns among populations of RVs that share similar characteristics and interactions. In this paper we focus on directed graphical models. The following section presents the extension of the widely used *Bayesian Network* formalism, by incorporating the aforementioned semantics.

3.2.1 First-Order Bayesian Networks

As in the classical Bayesian Networks, by taking into account the conditional independence assumptions encoded in the graph's structure, we can define a probability distribution over all the valid instantiations of the RVs of the model, as a product of the local conditional probabilities of its nodes, given their parents. Thus:

$$p(R_V) = \prod_{u \in V} p(R_u | Pa(R_u)), \tag{1}$$

where $Pa(R_u)$ is the set of the parental nodes of R_u.

Definition 1 A *First-Order Bayesian Network* (FBN) is defined as follows. Given a number of compatible entity hierarchies $\mathcal{H}$, and a set of RVs $\mathcal{R}$, drawn from a specific level l of these hierarchies, we introduce for each RV $R_i \in \mathcal{R}$:

- a set of parents nodes $Pa(R_i) = \{R^1_{\pi_i}, \ldots, R^k_{\pi_i}\}$;
- a conditional probability table $\phi_l(R_i)$ (CPT) that represents the local probability distribution defined by the RV R_i and its parents, $p(R_i|Pa(R_i))$;
- an *inherited* indicator, specifying a CPT $\phi_{l'}(R_i)$ from the FBN $M_{l'}$ of a previous level $l' < l$, which may act as a template according to the first order semantics;
- an *aggregate* indicator, which points to an ancestor RV at a coarser hierarchical level $l' < l$, iff that RV participates in an aggregate correlation at level l'■

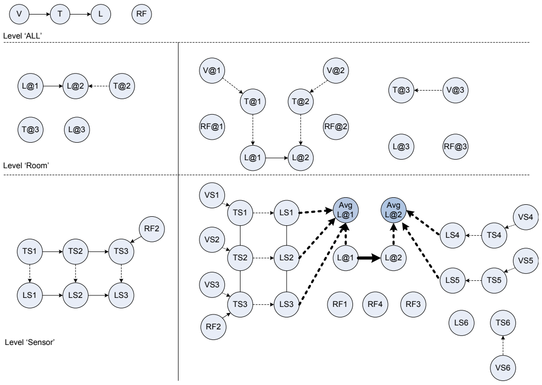

Figure 4. A 3-level HFBN for the Digital Home domain.

3.3 Hierarchical FO Bayesian Networks

Having described their constituents, we define a HFBN as a hierarchy of First-Order Bayesian Networks $\mathcal{M}^* = \{\mathcal{M}_1, \ldots, \mathcal{M}_n\}$. Each FBN corresponds to the aggregation level of the *dimension chain* of the hierarchy $\mathcal{H}(R_V)$, that the participating RVs $R_V \subseteq \mathcal{R}$ belong to.

Based on the first-order semantics described above, we can represent the FBNs that comprise an HFBN in a more intuitive way: interconnections of RVs at levels $\{1, \ldots, i-1\}$ that form *templates*, imply the same sets of correlations, with the same CPTs, between their decedent RVs which participate at the next level network $\mathcal{M}_i$. These finer-grained RVs do not need to be explicitly represented. As a result, we gain significant computational and spatial savings during both learning and inferencing, since the inherited factors do not need to be explicitly replicated or recomputed. Aggregated correlations at more abstract levels are again not represented in $\mathcal{M}_i$.

HFBNs do not preclude the existence of exceptions to an abstract correlation pattern that is captured at a coarser granularity network. Let us look at an example HFBN, which is depicted in Figure 4. Inherited and aggregate correlations are symbolized by dotted and bolded edges respectively.

At the most abstract level, $\mathcal{M}_1$ encodes a pairwise correlation pattern between average temperature, light level and voltage readings among all the environmental sensors in the deployment. In addition, RV RF is assumed to be independent of the all the others. While this is true for the majority of the sensors, there exist some cases, like $T@3$ and $L@3$ (average temperature and light at room 3) that the above correlation patterns do not hold. Since the high "compression ratio" imposed by the coarse granularity of the RVs in $\mathcal{M}_1$ cannot support modeling information in finer detail levels, we expand our HFBN by introducing this refinement in a finer granularity level (model $\mathcal{M}_2$). Apart from exceptions, further structural refinements may be present. In the same example, we see that $L@1$ and $L@2$ appear correlated, an example of an *aggregate* correlation, which could not be captured by $\mathcal{M}_1$ at the previous abstraction level.

As Figure 4 suggests (left column of the figure for the FBNs at levels 2 and 3), correlations that carry over from coarser granularity FBNs, do not need to be explicitly modeled again. Only exceptions have to be specified at each level of granularity. This compact representation, not only results in a more intuitive graphical model formalism, but also is indicative of the amount of the state that is required to store the FBNs at each level of the hierarchy.

With regards to inferencing, for a probabilistic query that involves RVs from various granularity levels, one has to go to the FBN at the level corresponding to the finest-grained RV in the query, $\mathcal{M}_i$, and create a *ground* BN, incorporating inherited and aggregated correlations recursively, from all the coarser grained FBNs ($\mathcal{M}_1, \ldots, \mathcal{M}_{i-1}$). In Figure 4 the ground BNs for the second and third level appear to the right of their corresponding FBNs. First-order inferencing techniques can also be applied, so that inference can be accelerated [6].

For inherited correlations, grounding results in replication of the same correlation patterns among the descendants of the abstracted RVs, and copying the CPTs of the parent nodes to their descendants. Inherited correlations *can be overridden* by exceptions at any finer level FBN.

Grounding aggregated correlations proves to be a more complicated process. In Figure 4, at the second level, the edge between $L@1$ and $L@2$ represents an aggregate correlation between the average light levels of rooms 1 and 2. If we ground the FBN $\mathcal{M}_3$, we notice that the keys of the partitions corresponding to the finer grained RVs $LS1, LS2, LS3$ and $LS4, LS5$ are not the same (clearly here, no semantically valid one-to-one mapping can be found between these two groups of RVs). Although these

5 RVs can not be directly correlated, they are correlated at a more abstract level, through $L@1$ and $L@2$.

In terms of parameterization, the factorization of the ground Bayesian Network of level 3 will also contain the aggregated conditional probability between the coarser level RVs $L@1$ and $L@2$. The problem lies in binding the contribution of the aggregate potential $p(L@2|L@1)$ with the finer grained potentials in the overall parameterization. Essentially, the coarse-grained RV $L@1$ and its children $LS1, LS2$ and $LS3$ are *deterministically* dependent on each other by the aggregate function $\gamma(.)$ used to define the scope of $L@1$ (in our example $\mathcal{S}(L@1) = \gamma(LS1, LS2, LS3) = avg(LS1, LS2, LS3)$). The same holds for $L@2$ and its children RVs.

Mixing deterministic and probabilistic correlations in a graphical model conflicts with its default parameterization, as the deterministic correlations prohibit all incompatible assignments of values to the nodes in the network (e.g. any assignment on $LS1$, $LS2$, $LS3$ and $L@1$ which invalidates the deterministic constraint $\frac{LS1+LS2+LS3}{3} = L@1$). To overcome this problem we introduce two auxiliary binary RVs, $AvgL@1$ and $AvgL@2$, that capture the deterministic constraints of $L@1$ and $L@2$ with their respective children.

More formally, for every deterministic constraint $\mathcal{D}(\mathcal{R})$ that involves a set of RVs $\mathcal{R}$, we introduce the binary auxiliary variable $AuxR$, with CPT defined as follows:

$$Pr(AuxR = 1|\mathcal{R}) = \begin{cases} 1 & \text{if } \mathcal{R} \text{ satisfies } \mathcal{D}(\mathcal{R}); \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

One can easily verify that the FBN resulting from the incorporation of these auxiliary variables, can be used to perform inference on the set of RVs of the problem's domain, by simply conditioning the auxiliary RVs $AuxR$ to be true (i.e. $Pr(\mathcal{R}) = Pr(\mathcal{R}|AuxR = \vec{1})$).

4 Learning HFGMs

Graphical model learning from complete datasets has been proven a substantially challenging task. It involves inducing the structure and the parameters of the model, and because of the NP-Hardness of the process [2], the problem is attacked by heuristically navigating through the big search space of candidate models, for the one that maximizes the likelihood to the training data.

Learning HFBNs involves a considerably larger search space; as the RV hierarchies get deeper or the number of dimensions gets larger, the number of RVs at each level increases dramatically, with negative impact to the structure learning process. Also, in order to support exceptions to the correlations inherited by coarser-grained FBNs, the learning process has to consider, apart from edge additions as its basic operator, removals of inherited edges, and even recomputations of the inherited or aggregate CPTs.

Our learning algorithm essentially performs a top-down greedy search to construct the FBN at each hierarchical level, by first inheriting the existing template or aggregate correlations from the previous level FBN. The resulting model gets evaluated and then possible perturbations of the inherited structure and parameters are considered, to identify exceptions. The algorithm terminates when a prespecified accuracy vs. model complexity condition is met. A more detailed description of the algorithm is provided in the full version of the paper.

5 Conclusions and Future Work

In this paper we presented HFGMs, a Probabilistic Graphical Modeling formalism which provides the means for an economical and intuitive representation of the correlations among the data, in a Probabilistic Database System. Guided by a set of hierarchical dimensions that partition the probabilistic values of the database in regions of varying granularity, HFGMs utilize a combination of *first-order* and *aggregate* semantics, to model the correlations among these values, in the granularity that captures the underlying data distribution as faithfully as possible. Through the succinct representation of uncertain data correlation they employ, HFGMs enable the construction of practical Probabilistic Database Systems, which can scale to large datasets, without having to resort to unrealistic statistical models, like full independence, which limit their applicability. We plan to extend our formalism to multiple hierarchical dimensions, and to also support Markov Networks as an additional underlying graphical modeling framework, so that a larger family of probability distributions can be modeled.

References

[1] O. Benjelloun, A. D. Sarma, A. Halevy, and J. Widom. Uldbs: Databases with uncertainty and lineage. In *Proceedings of the 32nd VLDB Conference*, Seoul, Korea, 2006.

[2] D. Chickering, D. Geiger, and D. Heckerman. Learning Bayesian Networks is NP-Hard. Technical Report MSR-TR-94-17, 1994.

[3] N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. *VLDB Journal*, 2006.

[4] A. Deshpande and P. Sen. Representing and querying correlated tuples in probabilistic databases. In *23rd International Conference on Data Engineering*, Istanbul, Turkey, 2007.

[5] L. Getoor, N. Friedman, D. Koller, and A. Pfeffer. Learning probabilistic relational models. *Relational Data Mining*, 2001.

[6] D. Poole. First-order probabilistic inference. In *Proceedings of the 2003 International Joint Conferene on Artificial Intelligence (IJCAI)*, 2003.