Learning Bayes Nets for Relational Data With Link Uncertainty Extended Abstract

Oliver Schulte and Zhensong Qian

School of Computing Science
Simon Fraser University
Vancouver-Burnaby, Canada

Abstract

We present an algorithm for learning correlations among link types and node attributes in relational data that represent complex heterogeneous networks. The link correlations are represented in a Bayes net structure. The current state of the art algorithm for learning relational Bayes nets captures only correlations among entity attributes *given* the existence of links among entities. The models described in this paper capture a wider class of correlations that involve uncertainty about the link structure. Our base line method learns a Bayes net from join tables directly. This is a statistically powerful procedure that finds many correlations, but does not scale well to larger datasets. We compare join table search with a hierarchical search strategy. A key challenge for relational learning that scales with data size is to compute event counts in a relational database (sufficient statistics), especially when these involve negated relationships. We describe how the fast Möbius transform provides a scalable solution for this problem.

1 Introduction

Link analysis for heterogenous networks with multiple link types is a challenging problem in network science. We describe a method for learning a Bayes net that captures simultaneously correlations between link types, link features, and attributes of nodes. Previous work on learning Bayes nets for relational data was restricted to correlations among attributes given the existence of links [19]. The larger class of correlations examined in our new algorithms includes two additional kinds: [1]

1. Dependencies between different types of links.
2. Dependencies among node attributes given the *absence* of a link between the nodes.

Discovering such dependencies is useful for several applications.

Knowledge Discovery Dependencies provide valuable insights in themselves. For instance, a web search manager may wish to know whether if user searches for a video in Youtube for a product, they are also likely to search for it on the web.

Relevance Determination Once dependencies have been established, they can be used as a relevance filter for focusing further network analysis only on statistically significant associations. For example, the classification and clustering methods of [21] for heterogeneous networks assume that a set of "meta-paths" have been found that connect link types that are associated with each other.

Query Optimization The Bayes net model can also be used to estimate relational statistics, the frequency with which statistical patterns occur in the database [20]. This kind of statistical model can be applied for database query optimization [7].

Approach We consider three approaches to multiple link analysis with Bayes nets.

Flat Search Apply a standard Bayes net learner to a single large join table. This table is formed as follows: (1) take the cross product of entity tables. (An entity table lists the set of nodes of a given type.) (2) For each tuple of entities, add a relationship indicator whose value "true" or "false" indicates whether a certain relationship holds among the entities.

Hierarchical Search Conducts bottom-up search through the lattice of table joins hierarchically. Dependencies (Bayes net edges) discovered on smaller joins are propagated to larger joins. The different table joins include information about the presence or absence of relationships as in the flat search above. This is an extension of the current state of the art Bayes net learning algorithm for relational data [19].

[1] This research was supported by a Discovery Grant to Oliver Schulte from the Canadian Natural Sciences and Engineering Council. And Zhensong Qian was also supported by a grant from the China Scholarship Council. This is a preliminary version of a paper that will appear in the post proceedings of the IJCAI 2013 GKR workshop.

Evaluation. We compare the learned models using standard scores (e.g., Bayes Information Criterion, log-likelihood). These results indicate that both flat search and hierarchical search are effective at finding correlations among link types. Flat search can on some datasets achieve a higher score by exploiting attribute correlations that depend on the absence of relationships. Structure learning time results indicate that hierarchical search is substantially more scaleable.

Contributions

1. To our knowledge this is the first application of Bayes net learning to modelling correlations among different types of links.
2. Extension of a lattice search strategy for link type modelling, with a comparison to a flat search join approach.

Paper Organization We describe Bayes net models for relational data (Poole's Parametrized Bayes Nets). Then we present the learning algorithms, first flat search then hierarchical search. We compare the models on four databases from different domains.

2 Related Work

To our knowledge, there are no implementations of structure learning algorithms for directed graphical models that consider correlations among different link types, let alone together with node attributes. Such implementations exist, however, for other types of graphical models, specifically Markov random fields (undirected models) [2] and dependency networks (directed edges with cycles allowed) [14]. Structure learning programs for Markov random fields include Alchemy [2] and MLN-Booster *et al.* [11]. Neither of these programs is able to return a result on half of our datasets because they are too large. For space reasons we restrict the scope of this paper to directed graphical models and do not go further into undirected model. For an extensive comparison of the learn-and-join Bayes net learning algorithm with Alchemy please see [19].

3 Background and Notation

Poole introduced the Parametrized Bayes net (PBN) formalism that combines Bayes nets with logical syntax for expressing relational concepts [15]. We adopt the PBN formalism, following Poole's presentation.

3.1 Bayes Nets for Relational Data

A **population** is a set of individuals. Individuals are denoted by lower case expressions (e.g., *bob*). A **population variable** is capitalized. A **functor** represents a mapping $f : \mathcal{P}_1, \ldots, \mathcal{P}_a \rightarrow V_f$ where f is the name of the functor, and V_f is the output type or **range** of the functor. In this paper we consider only functors with a finite range, disjoint from all populations. If $V_f = \{T, F\}$, the functor f is a (Boolean) **predicate**. A predicate with more than one argument is called a **relationship**; other functors are called **attributes**. We use uppercase for predicates and lowercase for other functors.

A **Bayes Net (BN)** is a directed acyclic graph (DAG) whose nodes comprise a set of random variables and conditional probability parameters. For each assignment of values to the nodes, the joint probability is specified by the product of the conditional probabilities, $P(child|parent\_values)$. A **Parametrized random variable** is of the form $f(X_1, \ldots, X_a)$, where the populations associated with the variables are of the appropriate type for the functor. A **Parametrized Bayes Net** (PBN) is a Bayes net whose nodes are Parametrized random variables [15]. If a Parametrized random variable appears in a Bayes net, we often refer to it simply as a node.

3.2 Databases and Table Joins

We begin with a standard **relational schema** containing a set of tables, each with key fields, descriptive attributes, and possibly foreign key pointers. A **database instance** specifies the tuples contained in the tables of a given database schema. A relational structure can be visualized as a complex heterogeneous network [16, Ch.8.2.1]: individuals are nodes, attributes of individuals are node labels, relationships correspond to (hyper)edges, and attributes of relationships are edge labels. Conversely, a complex heterogeneous network can be represented using a relational database schema.

We assume that tables in the relational schema can be divided into *entity tables* and *relationship tables.* This is the case whenever a relational schema is derived from an entity-relationship model (ER model) [23, Ch.2.2]. In our university example, there are two entity tables: a *Student* table and a *Course* table. There is one relationship table *Registered* with foreign key pointers to the *Student* and *Course* tables whose tuples indicate which students have registered in which courses.

The functor formalism is rich enough to represent the constraints of an ER schema by the following translation: Entity sets correspond to types, descriptive attributes to functions, relationship tables to predicates, and foreign key constraints to type constraints on the arguments of relationship predicates.

Table 1 shows a relational schema for a database related to a university. Figure 1 displays a small database instance for this schema together with a Parametrized Bayes Net (omitting the *Teaches* relationship for simplicity.)

The **natural table join**, or simply join, of two or more tables contains the rows in the Cartesian products of the tables whose values match on common fields. In logical terms, a join corresponds to a conjunction [23].

4 Bayes Net Learning With Link Correlation Analysis

We outline the two methods we compare in this paper, flat search and hierarchical search.

(a)

Student		
s-id	Intelligence	Ranking
Jack	3	1
Kim	2	1
Paul	1	2

(b)

Registration			
s-id	C.id	Grade	Satisfaction
Jack	101	A	1
Jack	102	B	2
Kim	102	A	1
Kim	103	A	1
Paul	101	B	1
Paul	102	C	2

(c)

Course					
c-id	Prof	Rating	Difficulty	T-a-prof	P-prof
101	Oliver	3	1	1	3
102	David	2	2	1	2
103	Oliver	3	2	1	3

(d)

attribute-relation table								
Intelligence	Rank	Rating	Diff	P-prof	t-a-prof	Reg(S,C)	R.Grade	R.Sat
3	1	3	1	3	1	T	A	1
3	1	2	2	2	1	T	B	2
3	1	3	2	3	1	F	⊥	⊥
2	1	3	1	3	1	F	⊥	⊥
2	1	2	2	2	1	T	A	1
2	1	3	2	3	1	T	A	1
1	2	3	1	3	1	T	B	1
1	2	2	2	2	1	T	C	2
1	2	3	2	3	1	F	⊥	⊥

(e)

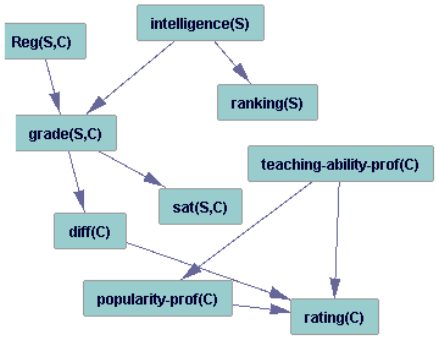

Figure 1: Database Table Instances: (a) *Student*, (b) *Registered* (c) *Course*. To simplify, we added the information about professors to the courses that they teach. (d) The attribute-relation table $Registered^+$ derived from *Registered*, which lists for each pair of entities their descriptive attributes, whether they are linked by *Registered*, and the attributes of a link if it exists. (e) A Parametrized Bayes Net for the university schema.

$Student(\underline{student\_id}, intelligence, ranking)$
$Course(\underline{course\_id}, difficulty, rating)$
$Professor\ (\underline{professor\_id}, teaching\_ability, popularity)$
$Registered\ (\underline{student\_id}, \underline{course\_id}, grade, satisfaction)$
$Teaches(\underline{professor\_id}, \underline{course\_id})$

Table 1: A relational schema for a university domain. Key fields are underlined. An instance for this schema is given in Figure 1.

4.1 Flat Search

The basic idea for flat search is to apply a standard propositional or single-table Bayes net learner to a single large join table. To learn correlations between link types, we need to provide the Bayes net with data about when links are present *and* when they are absent. To accomplish this, we add to each relationship table a **link indicator column**. This columns contains T if the link is present between two entities, and F if the link is absent. (The entities are specified in the primary key fields.) We add rows for all pairs of entities of the right type for the link, and enter T or F in the link indicator column depending on whether a link exists or not. We refer to relationship tables with a link indicator column as **extended** tables. Extended tables are readily computed using SQL queries. If we omit the entity Ids from an extended table, we obtain the **attribute-relation** table that lists (1) all attributes for the entities involved, (2) whether a relationship exists and (3) the attributes of the relationship if it exists. If the attribute-relation table is derived from a relationship R, we refer to it as R^+.

The attribute-relation table is readily defined for a set of relationships: take the cross-product of all populations involved, and add a link indicator column for each relationship in the set. For instance, if we wanted to examine correlations that involve both the *Registered* and the *Teaches* relationships, we would form the cross-product of the entity types *Student*, *Course*, *Professor* and build an attribute-relation table that contains two link indicator columns $Registered(S, C)$ and $Teaches(P, C)$. The **full join** is the attribute-relation table for all relationships in the database.

The **flat search Bayes net learner** takes a standard Bayes net learner and applies it to the full join table to obtain a single Parametrized Bayes net. The results of [17] can be used to provide a theoretical justification for this procedure; we outline two key points. (1) The full join table correctly represents the *sufficient statistics* of the database: using the full join table to compute the frequency of a joint value assignment for Parametrized Random Variables is equivalent to the frequency with which this assignment holds in the database. (2) Maximizing a standard single-table likelihood score from the full join table is equivalent to maximizing the *random selection pseudo likelihood*. The random selection pseudo log-likelihood is the expected log-likelihood assigned by a Parametrized Bayes net when we randomly select individuals from each population and instantiate the Bayes net with attribute values and relationships associated with the selected individuals.

4.2 Hierarchical Search

Khosravi *et al.* [19] present the learn-and-join structure learning algorithm. The algorithm upgrades a single-table Bayes net learner for relational learning. We describe the fundamental ideas of the algorithm; for further details please see [19]. The key idea is to build a Bayes net for the entire database by level-wise search through the *table join lattice*. The user chooses a single-table Bayes net learner. The learner is applied to table joins of

size 1, that is, regular data tables. Then the learner is applied to table joins of size $s, s+1, \ldots$, with the constraint that larger join tables inherit the absence or presence of learned edges from smaller join tables. These constraints are implemented by keeping a global cache of forbidden and required edges. Algorithm 1 provides pseudocode for the previous learn-and-join algorithm (LAJ) [18].

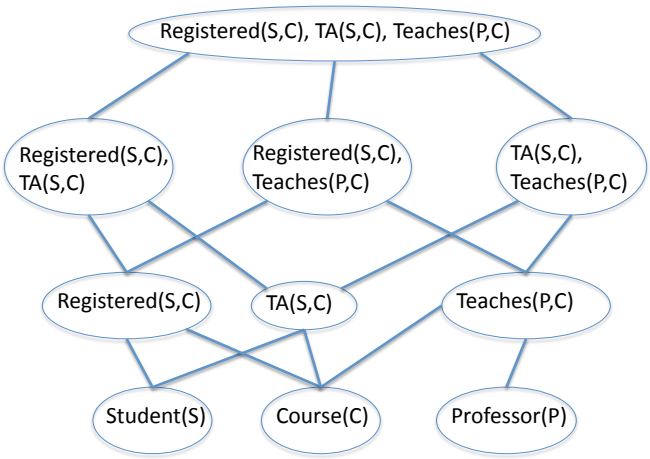

Figure 2: A lattice of relationship sets for the University schema of Table 1. Links from entity tables to relationship tables correspond to foreign key pointers.

To extend the learn-and-join algorithm for link analysis, we replace the natural join in line 7 by the extended join (more precisely, by the attribute-relation tables derived from the extended join). The natural join contains only tuples that appear in all relationship tables. Compared to the extended join, this corresponds to considering only rows where the link indicator columns have the value T. When the propositional Bayes net learner is applied to such a table, the link indicator variable appears like a constant. Therefore the BN learner cannot find any correlations between the link indicator variable and other nodes, nor can it find correlations among attributes conditional on the link indicator variable being F. Thus the previous LAJ algorithm finds only correlations between entity attributes conditional on the existence of a relationship. In sum, hierarchical search with link correlations can be described as follows.

1. Run the previous LAJ algorithm (Algorithm 1) using natural joins.

2. Starting with the constraints from step 1, extend them with the LAJ algorithm where extended joins replace natural joins. That is, for each relationship set shown in the lattice of Figure 2, apply the single-table Bayes net learner to the extended join for the relationship set.

Algorithm 1 Pseudocode for previous Learn-and-Join Structure Learning for Lattice Search.

Input: Database $\mathcal{D}$ with $E_1, ..E_e$ entity tables, $R_1, ...R_r$ Relationship tables,
Output: Bayes Net for $\mathcal{D}$
Calls: PBN: Any propositional Bayes net learner that accepts edge constraints and a single table of cases as input.
Notation: PBN(T, Econstraints) denotes the output DAG of PBN. Get-Constraints(G) specifies a new set of edge constraints, namely that all edges in G are required, and edges missing between variables in G are forbidden.

1: Add descriptive attributes of all entity and relationship tables as variables to G. Add a Boolean indicator for each relationship table to G.
2: Econstraints = $\emptyset$ [Required and Forbidden edges]
3: **for** m=1 to e **do**
4: Econstraints += Get-Constraints(PBN(E_m , $\emptyset$))
5: **end for**
6: **for** m=1 to r **do**
7: N_m := natural join of R_m and entity tables linked to R_m
8: Econstraints += Get-Constraints(PBN(N_m, Econstraints))
9: **end for**
10: **for all** N_i and N_j with a foreign key in common **do**
11: K_{ij} := join of N_i and N_j
12: Econstraints += Get-Constraints(PBN(K_{ij}, Econstraints))
13: **end for**
14: **return** Bayes Net defined by Econstraints.

5 Evaluation

All experiments were done on a QUAD CPU Q6700 with a 2.66GHz CPU and 8GB of RAM. The LAJ code and datasets are available on the world-wide web [10]. We made use of the following single-table Bayes Net search implementation: GES search [1] with the BDeu score as implemented in version 4.3.9-0 of CMU's Tetrad package (structure prior uniform, ESS=10; [22]).

Methods Compared We compared the following methods.

LAJ The previous LAJ method without link correlations (Algorithm 1).

LAJ+ The new LAJ method that has the potential to find link correlations (Algorithm 1 with the extended join instead of natural join).

Flat Applies the single-table Bayes net learner to the full join.

Performance Metrics We report learning time, log-likelihood, Bayes Information Criterion (BIC), and the Akaike Information Criterion (AIC). We write

$$L(\hat{G}, \mathbf{d})$$

for the log-likelihood score, where $\hat{G}$ is the BN G with its parameters instantiated to be the maximum likelihood

Dataset	#tuples
University	662
Movielens	1585385
Mutagenesis	1815488
Hepatitis	2965919
Small-Hepatitis	19827

Table 2: Size of datasets in total number of table tuples.

Dataset	Flat	LAJ+	LAJ
University	1.916	1.183	0.291
Movielens	38.767	18.204	1.769
Mutagenesis	3.231	3.448	0.982
Small-Hepatitis	9429.884	8.949	10.617

Table 3: Model Structure Learning Time in seconds.

estimates given the dataset $\mathbf{d}$, and the quantity $L(\hat{G}, \mathbf{d})$ is the log-likelihood of $\hat{G}$ on $\mathbf{d}$.

The BIC score is defined as follows [1; 17]

$$BIC(G, \mathbf{d}) = L(\hat{G}, \mathbf{d}) - par(G)/2 \times ln(m)$$

where the data table size is denoted by m, and $par(G)$ is the number of free parameters in the structure G. The AIC score is given by

$$AIC(G, \mathbf{d}) = L(\hat{G}, \mathbf{d}) - par(G).$$

BIC and AIC are standard scores for Bayes nets [1]. AIC is asympotically equivalent to selection by cross-validation, so we may view it as a closed-form approximation to cross-validation, which is computationally demanding for relational datasets.

Datasets We used one synthetic and three benchmark real-world databases, with the modifications described by Schulte and Khosravi [19]. See that article for more details.

University Database. We manually created a small dataset, based on the schema given in Table 1. The dataset is small and is used as a testbed for the correctness of our algorithms.

MovieLens Database. A dataset from the UC Irvine machine learning repository. The data are organized in 3 tables (2 entity tables, 1 relationship table, and 7 descriptive attributes).

Mutagenesis Database. A dataset widely used in ILP research. It contains two entity tables and two relationships.

Hepatitis Database. A modified version of the PKDD'02 Discovery Challenge database. The data are organized in 7 tables (4 entity tables, 3 relationship tables and 16 descriptive attributes). In order to make the learning feasible, we under sampled Hepatitis database to keep the ratio of positive and negative link indicator equal to one.

5.1 Results

Learning Times Table 3 provides the model search time for each of the link analysis methods. This does not include the time for computing table joins since this is essentially the same for all methods (the cost of the full join). On the smaller and simpler datasets, all search strategies are fast, but on the medium-size and more complex datasets (Hepatitis, MovieLens), hierarchical search is much faster due to its use of constraints. Adding prior knowledge as constraints could speed the structure learning substantially.

University	BIC	AIC	log-likelihood	# Parameter
Flat	-17638.27	-12496.72	-10702.72	1767
LAJ+	-13495.34	-11540.75	-10858.75	655
LAJ	-13043.17	-11469.75	-10920.75	522

MovieLens	BIC	AIC	log-likelihood	# Parameter
Flat	-4912286.87	-4911176.01	-4910995.01	169
LAJ+	-4911339.74	-4910320.94	-4910154.94	154
LAJ	-4911339.74	-4910320.94	-4910154.94	154

Mutagenesis	BIC	AIC	log-likelihood	# Parameter
Flat	-21844.67	-17481.03	-16155.03	1289
LAJ+	-47185.43	-28480.33	-22796.33	5647
LAJ	-30534.26	-25890.89	-24479.89	1374

Hepatitis	BIC	AIC	log-likelihood	# Parameter
Flat	-7334391.72	-1667015.81	-301600.81	1365357
LAJ+	-457594.18	-447740.51	-445366.51	2316
LAJ	-461802.76	-452306.05	-450018.05	2230

Table 4: Performance of different Searching Algorithms by dataset.

Statistical Scores As expected, adding edges between link nodes improves the statistical data fit: the link analysis methods LAJ+ and Flat perform better than the learn-and-join baseline in terms of log-likelihood on all datasets shown in table 4, except for MovieLens where the Flat search has a worse score. On the small synthetic dataset University, flat search appears to overfit whereas the hierarchical search methods are very close. On the medium-sized dataset MovieLens, which has a simple structure, all three methods score similarly. Hierarchical search finds no new edges involving the single link indicator node (i.e., LAJ and LAJ+ return the same model).

The most complex dataset, Hepatitis, is a challenge for flat search, which seems to overfit severely with a huge number of parameters that result in a model selection score that is an order of magnitude worse than for hierarchical search. Because of the complex structure of the Hepatitis schema, the hierarchical constraints appear to be effective in combating overfitting.

The situation is reversed on the Mutagenesis dataset where flat search does well: compared to attribute-only search, it manages to fit the data better with a less complex model. Hiearchical search performs very poorly compared to flat search (lower likelihood yet many more parameters in the model). Investigation of the models

shows that the reason for this phenomenon is a special property of the Mutagenesis dataset: whereas generally relationships are sparse—very few pairs of entities are actually linked—in Mutagenesis most entities whose type allows a link are linked. As a result, we find strong correlations between attributes conditional on *the absence of relationships*. The LAJ+ algorithm is constrained so that it cannot add Bayes net edges between attribute nodes at its second stage, when absent relationships are considered. As a result, it can represent attribute correlations conditional on the absence of relationships only indirectly through edges that involve link indicators. A solution to this problem would be to add a phase to the search so that we first learn edges between attributes conditional on the existence of relationships, then conditional on their nonexistence. The last phase then would consider edges that involve relationship nodes. We expect that with this change, hierarchical search would be competitive with flat search on the Mutagenesis dataset as well.

6 Computing Relational Sufficient Statistics

The learning algorithms described in this paper rely on the availability of the extended relational tables R^+ (see Figure 1). Our current implementation constructs this tables using standard joins. While this was sufficient for our experiments, the cross-products carry a quadratic costs for binary relations, and therefore do not scale to large datasets. Moreover, the hierarchical search requires joins of the extended tables. In this section we describe a "virtual join" algorithm that computes the required data tables without the quadratic cost of materializing a cross-product.

Our starting point is the observation that a statistical learning algorithm like a Bayes net learner does not require an enumeration of individuals tuples, but only *sufficient statistics* [8; 17]. Consider a list of relationship nodes $R_1, R_2, \ldots, R_m$, and attribute nodes $f_1, \ldots, f_j$. For example, in Figure 3 we have $m = 2, j = 1$ and $f_1 = gender(X)$. The sufficient statistics for this set of random variables are the database probabilities

$$P_{\mathcal{D}}(R_1 = b_1, R_2 = b_2, \ldots, R_m = b_m; f_1 = v_1, \ldots, f_j = v_j) \quad (1)$$

where the b_i values are Boolean and each v_j is from the domain of f_j. Bayes net algorithms can construct a Bayes net when provided with a table as input that lists these sufficient statistics. In what follows, we suppose that there are r possible assignments of the form shown in Equation (1) and therefore r sufficient statistics to be specified.

So long as a database probability involves only positive relationships, the computation is straightforward. For example, in $P_{\mathcal{D}}(gender(X) = M, Friend(X, Y) = T)$, the value $\#_{\mathcal{D}}(gender(X) = M, Friend(X, Y) = T)$, the count of friendship pairs (x, y) where x is male and the

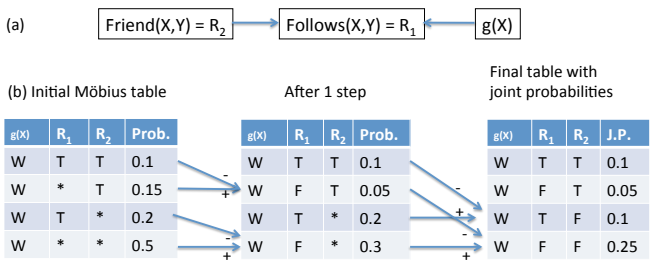

Figure 3: (a) A Bayes net with two relationship nodes. (b) An illustrative trace of the lattice Möbius transform (see text).

Friend relationship is true, can be computed by regular table joins or optimized virtual joins [24].

Computing joint probabilities for a family containing one or more negative relationships is harder. A naive approach would explicitly construct new data tables that enumerate tuples of objects that are *not* related. However, the number of unrelated tuples is too large to make this scalable (think about the number of user pairs who are *not* friends on Facebook). In their work on learning Probabilistic Relational Models with existence uncertainty, Getoor et al. provided a subtraction method for the special case of estimating a joint probability with only a single negated relationship [5, Sec.5.8.4.2]. They did not treat parameter learning with multiple negated relationships, which we consider next.

6.1 Statistics With Multiple Negated Relationships: The Fast Möbius Transform

The general case of multiple negative relationships can be efficiently computed using the **fast Möbius transform** (FMT), or Möbius transform for short. We compute the r joint probabilities (1) by first computing the r Möbius parameters of the joint distribution, then using the lattice Möbius transform to transform the Möbius parameters into the desired joint probabilities. Figure 4 provides an overview of the computation steps. Because the Möbius parameters involve probabilities for events with *positive relationships only*, they can be estimated directly from the data. We next define the Möbius parameters, then explain the FMT.

Let $\mathbb{B} = B_1, \ldots, B_m$ be a set of binary random variables with possible values 0 or 1, and P be the joint distribution that specifies 2^m probabilities, one for each possible assignment of values to the m binary variables. For any subset $\mathbf{B} \subseteq \mathbb{B}$ of the variables, let $P(\mathbf{B} = \mathbf{1})$ denote the probability that the variables in $\mathbf{B}$ are assigned the value 1, leaving the value of the other variables unspecified. The **Möbius parameters** of the distribution P are the values $P(\mathbf{B} = \mathbf{1})$ for all subsets $\mathbf{B} \subseteq \mathbb{B}$[4, Sec.3]. There are 2^m Möbius parameters for m binary variables, with $0 \leq P(\mathbb{B} = \mathbf{1}) \leq P(\mathbf{B} = \mathbf{1}) \leq P(\emptyset = \mathbf{1}) = 1$.

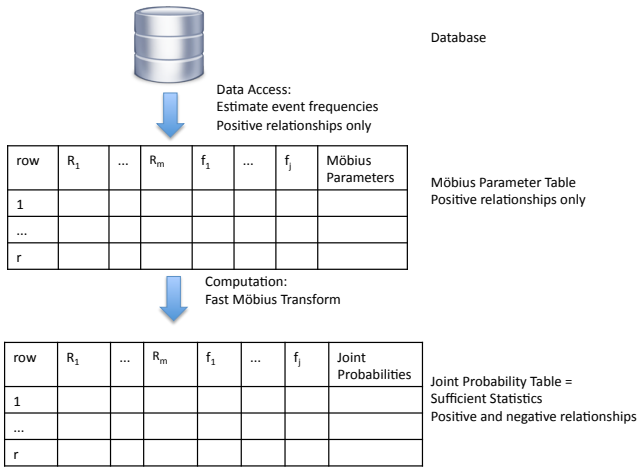

Figure 4: Computation of joint probabilities in a relational database. (1) Estimate Möbius parameters using standard table join operations. (2) Transform the Möbius parameters into joint probabilities. Only the first step involves data access.

If we fix the values $v_1, \ldots, v_j$ of the attribute atoms, the sufficient statistics correspond to a joint distribution over m Boolean relationship random variables:

$$P(R_1 = \cdot, R_2 = \cdot, \ldots, R_m = \cdot; f_1 = v_1, \ldots, f_j = v_j).$$

We refer to the Möbius parameters of this joint distribution as the **Möbius parameters for the attribute values** $v_1, \ldots, v_j$.

Example. For the Bayes net of Figure 3 (Top), fix the attribute condition $gender(X) = W$. The four Möbius parameters for this attribute condition are

$P(gender(X) = W)$
$P(Friend(X, Y) = T; gender(X) = W)$
$P(Follows(X, Y) = T; gender(X) = W)$
$P(Friend(X, Y) = T, Follows(X, Y) = T; gender(X) = W)$

6.2 The Fast Möbius Transform

The Möbius extension theorem entails that the joint probabilities can be computed from the Möbius parameters [12, Sec.4.4.2.1]. The Möbius Transform is an optimal algorithm for carrying out this computation, using the local update operation.

$$P(R = F, \mathbf{R}) = P(\mathbf{R}) - P(R = T, \mathbf{R}) \qquad (2)$$

where $\mathbf{R}$ is a conjunction of relationship specifications, possibly with both positive and negative relationships. The equation holds for any fixed set of attribute conditions $f_1 = v_1, \ldots, f_j = v_j$. Eq. 2 generalizes the subtraction trick: the joint probabilities on the right hand side each involve exactly one less false relationship than the joint probability on the left.

Algorithm 2 The Möbius transform for computing sufficient statistics with link uncertainty.

Input: database $\mathcal{D}$; a set of nodes divided into attribute nodes $f_1, \ldots, f_j$ and relationship nodes $R_1, \ldots, R_m$.
Output: joint probability table specifying the data frequencies for each joint assignment to the input nodes.

```
for all attribute value assignments f_1 := v_1, ..., f_j := v_j do
    initialize the table: set all relationship nodes to either T or *; find joint frequencies with data queries.
    for i = 1 to m do
        Change all occurrences of R_i = * to R_i = F.
        Update the joint frequencies using (2).
    end for
end for
```

Figure 4 illustrates the control flow for computing sufficient statistics. The FMT initializes the Möbius parameter values with frequency estimates from the data (top table). It then goes through the relationship nodes $R_1, \ldots, R_m$ in order, at stage i replacing all occurrences of $R_i = *$ with $R_i = F$, and applying the local update equation to obtain the probability value for the modified row. At termination, all $*$ values have been replaced by F and the table specifies all joint frequencies (bottom table).

Complexity Analysis For big data analysis, the key property of the FMT is that it accesses only *existing* links, never nonexisting links. The number of updates is $O(m \times r)$ [9]. If the number m of relationship nodes is small enough to be treated as a constant, the number of updates is therefore proportional to the number r of sufficient statistics.[2]

So far we have discussed the Möbius transform for a single fixed list of random variables. The Möbius transform could be applied dynamically during learning or off-line prior to learning. A pre-computation approach is attractive for analyzing large heterogeneous networks because it separates the problem of computing event frequencies/counts from the problem of statistical model selection. (Moore and Lee present a classic pre-computation approach for single-table data [13]). Moreover, pre-computing sufficient statistics for the entire database could take advantage of the lattice structure illustrated in Figure 2 to reuse computation results as much as possible.

7 Conclusion

We described different methods for extending relational Bayes net learning to correlations involving links. Sta-

[2]For general m, the problem of computing a sufficient statistic in a relational structure—a joint probability of the form (1)—is #P-complete [3, Prop.12.4].

tistical measures indicate that Bayes net methods succeed in finding relevant correlations. There is a trade-off between statistical power and computational feasibility (full table search vs constrained search). Hierarchical search often does well on both dimensions, but needs to be extended to handle correlations conditional on the absence of relationships.

A key issue for scalability is that most of the learning time is taken up by forming table joins, whose size is the cross product of entity tables. These table joins provide the sufficient statistics required in model selection. To improve scalability, computing sufficient statistics needs to be feasible for cross product sizes in the millions or more. A possible solution may be the virtual join methods that compute sufficient statistics without materializing table joins, such as the Fast Möbius Transform [20; 24].

References

[1] D. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2003.

[2] Pedro Domingos and Daniel Lowd. *Markov Logic: An Interface Layer for Artificial Intelligence*. Morgan and Claypool Publishers, 2009.

[3] Pedro Domingos and Matthew Richardson. Markov logic: A unifying framework for statistical relational learning. In *Introduction to Statistical Relational Learning* [6].

[4] Mathias Drton and Thomas S. Richardson. Binary models for marginal independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):287–309, 2008.

[5] Lise Getoor, Nir Friedman, Daphne Koller, Avi Pfeffer, and Benjamin Taskar. Probabilistic relational models. In *Introduction to Statistical Relational Learning* [6], chapter 5, pages 129–173.

[6] Lise Getoor and Ben Taskar. *Introduction to statistical relational learning*. MIT Press, 2007.

[7] Lise Getoor, Benjamin Taskar, and Daphne Koller. Selectivity estimation using probabilistic models. *ACM SIGMOD Record*, 30(2):461–472, 2001.

[8] D. Heckerman, D. Geiger, and D. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

[9] Robert Kennes and Philippe Smets. Computational aspects of the Möbius transformation. In *UAI*, pages 401–416, 1990.

[10] H. Khosravi, T. Man, J. Hu, E. Gao, and O. Schulte. Learn and join algorithm code. URL = `http://www.cs.sfu.ca/~oschulte/jbn/`.

[11] Tushar Khot, Jude Shavlik, and Sriraam Natarajan. Boostr, 2013. URL = `http://pages.cs.wisc.edu/~tushar/Boostr/`.

[12] Daphne Koller and Nir. *Probabilistic Graphical Models*. MIT Press, 2009.

[13] Andrew W. Moore and Mary S. Lee. Cached sufficient statistics for efficient machine learning with large datasets. *J. Artif. Intell. Res. (JAIR)*, 8:67–91, 1998.

[14] Sriraam Natarajan, Tushar Khot, Kristian Kersting, Bernd Gutmann, and Jude W. Shavlik. Gradient-based boosting for statistical relational learning: The relational dependency network case. *Machine Learning*, 86(1):25–56, 2012.

[15] David Poole. First-order probabilistic inference. In *IJCAI*, pages 985–991, 2003.

[16] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2010.

[17] Oliver Schulte. A tractable pseudo-likelihood function for Bayes nets applied to relational data. In *SIAM SDM*, pages 462–473, 2011.

[18] Oliver Schulte. Challenge paper: Marginal probabilities for instances and classes. ICML-SRL Workshop on Statistical Relational Learning., June 2012.

[19] Oliver Schulte and Hassan Khosravi. Learning graphical models for relational data via lattice search. *Machine Learning*, 88(3):331–368, 2012.

[20] Oliver Schulte, Hassan Khosravi, Arthur Kirkpatrick, Tianxiang Gao, and Yuke Zhu. Modelling relational statistics with bayes nets. In *Inductive Logic Programming (ILP)*, 2012.

[21] Yizhou Sun and Jiawei Han. *Mining Heterogeneous Information Networks: Principles and Methodologies*, volume 3. Morgan & Claypool Publishers, 2012.

[22] The Tetrad Group. The Tetrad project, 2008. http://www.phil.cmu.edu/projects/tetrad/.

[23] J. D. Ullman. *Principles of database systems*. W. H. Freeman & Co., 2 edition, 1982.

[24] Xiaoxin Yin, Jiawei Han, Jiong Yang, and Philip S. Yu. Crossmine: Efficient classification across multiple database relations. In *ICDE*, pages 399–410. IEEE Computer Society, 2004.