A Hierarchy of Independence Assumptions for Multi-relational Bayes Net Classifiers

Oliver Schulte, Bahareh Bina, Branden Crawford, Derek Bingham, Yi Xiong

***Abstract*—Many databases store data in relational format, with different types of entities and information about their attributes and links between the entities. Link-based classification (LBC) is the problem of predicting the class attribute of a target entity given the attributes of entities linked to it. In this paper we propose a new relational Bayes net classifier method for LBC, which assumes that different links of an object are independently drawn from the same distribution, given attribute information from the linked tables. We show that this assumption allows very fast multi-relational Bayes net learning. We define three more independence assumptions for LBC to unify proposals from different researchers in a single novel hierarchy. Our proposed model is at the top and the well-known multi-relational Naive Bayes classifier is at the bottom of this hierarchy. The model in each level of the hierarchy uses a new independence assumption in addition to the assumptions used in the higher levels. In experiments on four benchmark datasets, our proposed link independence model has the best predictive accuracy compared to the hierarchy models and a variety of relational classifiers.**

Keywords: Link-based Classification, Bayes net classifier, Statistical-relational learning, Knowledge Discovery in Databases

I. Introduction

Many databases store data in relational data tables, with different types of entities and information about their attributes and links between the entities. An important learning task is link-based classification (LBC) which takes advantage of attributes of links and linked entities, in addition to attributes of the target entity, in order to predict the class label [22]. For instance, the intelligence of a student can be predicted given difficulties of courses the student has taken, her grades in the courses, and the ranking of the student. A major approach to LBC is based on directed graphical models derived from Bayes nets (BNs) [26]. We propose a new multi-relational Bayes net classifier based on the *Path Independence Assumption* (PI).

A path is a chain of links that connect the target entity to a related entity. Under Path Independence different paths associating an object to other objects are independent given the attributes of the objects. For example, the assumption entails that two courses that an intelligent student has taken can be chosen independently from the same distribution given the difficulty of the courses. We derive a closed form classification formula. The main advantages of the path independence assumption compared to other multi-relational classifiers are as follows.

1) The Path Independence assumption models all the dependencies among attributes, given the path structures in the relational data, but no dependencies among the paths. Therefore, the Bayes net models are simpler and have fewer parameters.
2) There is no problem with cycles in the instantiated network (grounding) of the directed graph [31], [13].
3) Aggregation functions (e.g. average) [12] or combining rules (e.g., noisy-or [18]) are not required. Thus no information is lost and the computationally expensive search for aggregate features is avoided [5].

We develop a theoretical hierarchical framework for multi-relational Bayes net classifiers that are derived from independence assumptions. The hierarchy allows us to investigate the trade-off between the expressive power of the model on the one hand, and the scalability of learning on the other. We present three other independence assumptions further to path independence.

1) Influence Independence: given the target class label, the attributes of each related object are independent of the attributes of the target entity.
2) Naive Bayes: given the target class label, the attributes of each related object and the attributes of the target entity are all mutually independent.
3) Path-class Independence: the existence of a path from the target object to a related object is independent of the class label.

By combining *Path Independence* with these assumptions cumulatively (i.e. (Path Independence+1), (Path Independence+1+2), etc.), we obtain closed form formulas for three other relational Bayes Net classifiers. These independence assumptions unify, into a single hierarchy, proposals for LBC with directed graphical models from different researchers [23], [13], [5]. Formal statements of independence assumptions are important because they facilitate comparing graphical models with other approaches, and they allow us to analytically derive exact expressions for classification.

A. Evaluation

We compare the performance of classifiers in the hierarchy on four benchmark datasets. Furthermore, we compare our model hierarchy with Markov logic networks [8], and Tilde (Top-Down Induction of First-Order Logical Decision Trees) [3]. Learning with fewer independence assumptions is slower, but achieves more accurate predictions. In our results,

Schulte and Crawford are with the School of Computing Science, Simon Fraser University, Canada (email: {oschulte,bsc7@cs.sfu.ca}), Bingham and Xiong with the Department of Statistics (email:{dbingham,xiong@stat.sfu.ca,}), Simon Fraser University, and Bina is with Microsoft, Seattle.

This research was supported by NSERC discovery grants to OS.

978-1-4673-5895-8/13/$31.00 ©2013 IEEE

the performance of the model only using Path Independence was superior to the remaining three. Furthermore, the BN learned only with the Path Independence Assumption can be used to prune irrelevant tables and attributes. Thus our experiments suggest that Path Independence strikes an attractive balance between statistical and predictive power on the one hand, and efficient and scalable learning on the other.

Contributions. The main contributions of our work are as follows.

1) Developing an efficient multi-relational Bayes Net classifier.
2) Defining a new framework for multi-relational Bayes net classifiers, based on four basic independence assumptions.
3) The formal definition of independence assumptions and derivation of closed form classification formulas.

Paper Organization. We first describe related work, review background material and define our notation. Then we define our new relational Bayes net classifier and formalize a hierarchy of independence assumptions for multi-relational learning. The final section evaluates the classification performance of the different models on four benchmark datasets.

II. Related Work.

LBC has been researched extensively; we selectively review the work most related to ours. Our research combines ideas from data mining and statistical relational learning (SRL): The data mining approach of achieving scalability via independence assumptions with the statistical methods of directed graphical models. Manjunath *et al.* [23] outline the advantages of independence assumptions for LBC, especially scalability. They developed the heterogeneous Naive Bayes classifier (HNBC), which employs a naive Bayes assumption across tables, not within tables. This classifier is derived at the second level of our hierarchy.

Graphical Relational Models. Probabilistic Relational Models (PRMs) [15] upgrade Bayesian networks to deal with relational data. Getoor et al. [13] extend PRMs with link indicator nodes to model link structure between entities, and constrain the PRM such that the probability of a link indicator being true is a function of the attributes of the linked entities. This is a special case of Path Independence assumption for links = paths of length 1. They did not give a structure learning algorithm for this model, and combined it with a Naive Bayes classifier. We entail the classification formula of Naive Bayes+existence of link structure at the third level of our hierarchy.

Unlike directed graphical models which impose an acyclicity constraint, undirected ones do not have a cyclicity problem and are widely used for LBC. Two major formalisms are relational conditional Markov random fields [31] and Markov Logic Networks (MLNs) [8]. Undirected graphical models can represent essentially arbitrary dependencies [31] compared to directed models with the path independence assumption or more. The trade-off for the expressive power of undirected models is higher complexity in learning, especially scalable model learning is a major challenge in the multi-relational setting [16], [29].

Multi-Relational Naive Bayes. There has been extensive research into multi-relational versions of the single-table Naive Bayes Classifier (NBC) [4], [10], [25], [5]. Our NB classification formulas are mathematically derived from explicitly formalized independence assumptions. Inductive Logic Programming approaches, such as 1BC, typically use existential quantification to aggregate information form related objects [10]. Multi-relational Bayesian Classifier [5] is the most recent NBC for multi-relational data; its classification formula is entailed by the conjunction of all our four independence assumptions.

Propositionalization/Feature Generation Approaches. Many approaches to upgrading standard single-table classifiers use a mechanism for *relational feature generation*. Feature generation is done by using aggregate functions to summarize the information in links (e.g., [20], [27]) or learning an existentially quantified logical rule [21]. Several recent systems combine both aggregation and logical conditions (e.g., [27]). Instead of searching for new features, the predictors in our model are the descriptive attributes as defined in the relational database schema. Relational feature generation is by far the most computationally demanding part of such approaches. For example, generating 100,000 features on the CiteSeer dataset, which is smaller than the databases we consider in this paper, can take several CPU days [27, Ch.16.1.2]. So models that use independence assumptions, rather than generated features to combine the information from different tables, are orders of magnitude faster to learn, as our experiments confirm.

In practice one often needs to perform collective classification: predict the class label of several interrelated entities simultaneously. While collective classification has received much attention [31], [17], our theoretical framework for multi-relational classification formulas is already quite rich, so our empirical evaluation focuses only on individual classification and leaves applications to collective classification for future work.

III. Preliminaries and Notation

We review background on Bayes nets and relational databases. We describe a common preprocessing step for link-based classification, which extends the original database tables to join tables that include information about the target instances.

A. Bayes Net Classifiers

Bayes nets are a widely used model class for representing probabilistic knowledge. Pearle's seminal text is a good introduction to Bayes net concepts [26]. We consider Bayes nets for a set of variables $V = \{X_1, \ldots, X_n\}$ where each X_i has a *finite* number of values or states. A **Bayes net structure** $G = (V, \mathbf{E})$ for a set of variables V is a directed acyclic graph (DAG) over node set V. A Bayes net (BN) is a pair (G, θ_G) where θ_G is a set of parameter values

$Student(\underline{student\_id}, intelligence, ranking)$
$Course(\underline{c\_id}, diff, rating)$
$Professor\ (\underline{prof\_id}, teaching\_ability, popularity)$
$Registration\ (\underline{student\_id}, \underline{c\_id}, grade, satisfaction)$
$RA\ (\underline{student\_id}, \underline{prof\_id}, salary, capability)$
$Taughtby(\underline{c\_id}, \underline{prof\_id}, evaluation)$

TABLE I
A RELATIONAL SCHEMA FOR A UNIVERSITY DOMAIN.

that specify the probability distributions of each variable conditioned on instantiations of its parents. A BN (G, θ_G) defines a joint probability distribution over assignments to the variables V. Bayes nets are often used as classifier to predict the probability of a target class label given features [28]. Independence assumptions can be represented in a Bayes net as the absence of links, more generally connections, between independent variables [26]. The fact that Bayes nets can graphically encode many different independence assumptions makes them a good choice of model class for our study of multi-relational independence assumptions.

B. Relational Databases

A standard **relational schema** contains a set of tables. We assume that the schema follows an entity-relationship model (ER model) [32, Ch.2.2], so the tables in the relational schema are divided into *entity tables* $E_1, E_2, \ldots$ and *relationship tables* $R_1, R_2, \ldots$ that link entity tables to each other by foreign key pointers. Both entity tables and relationship tables may feature descriptive attributes. Key fields contain the ids of entities. The **natural join** $\bowtie$ of two tables is the set of tuples from their cross product that agree on the values of fields common to both tables [32]. For simplicity, we assume that relationships are binary, and that the target table is an entity table, but this is not essential for our results. A **database instance** specifies the tuples contained in the tables of a given database schema.

The symbol T is reserved for the target table, the symbol t denotes the target object, and $c(t)$ denotes the class label. The non-class attributes of t are collectively denoted as $\boldsymbol{a}(t)$, such that the target table contains a tuple $(\boldsymbol{a}(t), c(t))$ of attribute values.

Examples. As a running example, we use a university database. The schema is shown in table I. The schema has three entity tables: Student, Course and Professor, and three relationships: $Registration$ records courses taken by each student, $Taughtby$ records courses taught by a professor, and RA records research assistantship of students for professors. The target table is $T = Student$. The class attribute is $Intelligence$ of $Student$. There is only one non-class attribute from the student table, Ranking. If the target entity is $t = jack$, then we write $\boldsymbol{a}(t) = Ranking(jack)$, and $c(t) = Intelligence(jack)$.

C. Pathways and Join Tables.

One of the key challenges in multi-relational classification is the need to consider different pathways or table joins through which the target entity may be linked to other entities. Han et al. [5] proposed a graphical way to structure the space of possible pathways (see also [23]). A **Semantic Relationship Graph** (SRG) for a database $\mathcal{D}$ is a directed acyclic graph (DAG) whose nodes are database tables in $\mathcal{D}$ and whose only source (starting point) is the target table. Self-joins and other repeated occurrences of the same table can be handled by duplicating the table [5]. If an edge links two tables in the Semantic Relationship Graph, then the two tables share at least one primary key. We consider Semantic Relationship Graph paths of the form $T, R_1, \ldots, R_k, E_k$ that end with an entity table. For each such path, there is a corresponding **path join** $T \bowtie R_1 \cdots \bowtie R_k \bowtie E_k$. An **extended database** is a database that contains the entity tables together with path join tables, which we simply refer to as **path join tables**.

Example. Figure 1 shows the Semantic Relationship Graph for the University example. In the extended university database valid joins include the following:

- $J_1 = Student \bowtie RA \bowtie Professor.$
- $J_2 = Student \bowtie Registration \bowtie Course.$
- $J_3 = Student \bowtie Registration \bowtie Course \bowtie Taughtby \bowtie Professor.$

Figure 2 shows an instance of the university schema. Table J_2 is the extended table derived from joining Course, Student and Registration.

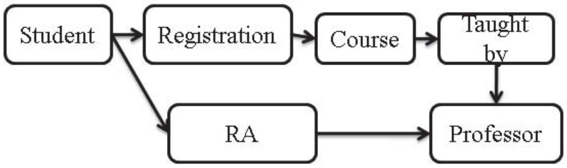
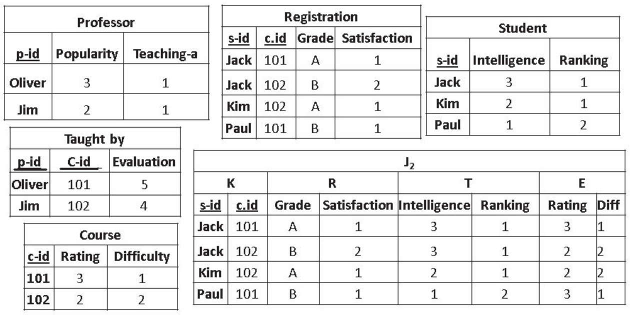


Fig. 1. Semantic relationship graph for the university schema.



Professor		
p-id	Popularity	Teaching-a
Oliver	3	1
Jim	2	1

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

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

Taught by		
p-id	C-id	Evaluation
Oliver	101	5
Jim	102	4

Course		
c-id	Rating	Difficulty
101	3	1
102	2	2

J_2							
K		R		T		E	
s-id	c.id	Grade	Satisfaction	Intelligence	Ranking	Rating	Diff
Jack	101	A	1	3	1	3	1
Jack	102	B	2	3	1	2	2
Kim	102	A	1	2	1	2	2
Paul	101	B	1	1	2	3	1

Fig. 2. A small instance for the university database with an extended table, and without the RA relation. The target table is $Student$, the target entity is Jack, and the class label is $Intelligence$.

To provide rigorous mathematical definitions and analysis, it is necessary to introduce a fairly elaborate notation to refer to tables, columns, rows, and tuples in an extended database. We begin with notation to denote the different kinds of columns in a path join table. The row r of table J_i is denoted by $J_{i,r}$. The values in each table row are partitioned as follows.

- **Key field values** are denoted as $R\_keys_{i,r}$.
- **Link attribute values** are denoted as $R\_atts_{i,r}$.
- **Target table attribute values** are denoted as $T\_atts_{i,r}$.
- **Other entity attribute values** are denoted as $E\_atts_{i,r}$.

Figure 2 shows the $R\_keys, R\_atts, E\_atts, T\_atts$ division of the columns of join table J_2.

Table Row	R_keys	R_atts	E_atts	T_atts
$J_{2,1}$	[jack,101]	[A,1]	[3,1]	[3,1]
$J_{3,1}$	[jack,101,oliver]	[A,1,5]	[3,1,3,1]	[3,1]

TABLE II
EXAMPLES OF NOTATION FOR VALUE TUPLES.

D. Probabilities and Relational Structures.

Most models in statistical-relational learning (SRL) combine probability with relational structures by treating the value of descriptive attributes (table entries) as random variables [15], [8], [18]. These models also allow for uncertainty about the existence of links or paths; in other words, probabilities may be assigned to the existence of a link as well as to the value of an attribute. We view a table J as a conjunction of the information pieces in it. Thus we consider probability assignments for the values of table entries. In our notation above, such probabilities are specified as follows.

- $P(R\_keys_{i,r} = ids)$ denotes the probability that the entities whose IDs are specified in the key fields are connected by the path corresponding to join table J_i.
- $P(R\_atts_{i,r} = values)$ denotes the probability that the attributes of the link or path associated with row r are specified by the tuple $values$.
- $P(T\_atts_{i,r} = values)$ denotes the probability that the target table entity that appears in row r has descriptive attribute features specified by the tuple $values$.
- $P(E\_atts_i = values)$ denotes the probability that the non-target entities that appears in row r have descriptive attribute features specified by the tuple $values$.

Example. We illustrate the definitions using the first row of join table J_2 in Figure 2.

- $P(R\_keys_{2,1} = [jack, 101])$ denotes the probability that Jack has registered in course 101.
- $P(R\_atts_{2,1} = [A, 1])$ denotes the probability that Jack received a grade of A in course 101 and that his satisfaction level in the course was 1.
- $P(T\_atts_{2,1} = [3, 1])$ denotes the probability that Jack's intelligence level is 3 and that his ranking is 1.
- $P(E\_atts_{2,1} = [3, 1])$ denotes the probability that course 101's rating is 3 and that its difficulty level is 1.

IV. THE PATH INDEPENDENCE CLASSIFIER

We define and discuss our main independence assumption, then derive a closed-form classification formula. In what follows fix an extended database with target attributes T, entity tables' attributes $E_1, \ldots, E_k = \mathbf{E}$, and path tables $J_1, \ldots, J_m = \mathbf{J}$.

A. Definition and Classification Formula

For a target entity t, the goal is to find the most probable class label

$$c^* = \operatorname*{argmax}_c P(\mathbf{J}, \mathbf{E}, T) = \operatorname*{argmax}_c P(\mathbf{E})P(T)P(\mathbf{J}|\mathbf{E}, T)$$
$$= \operatorname*{argmax}_c P(c(t)|\boldsymbol{a}(t)) \cdot P(\mathbf{J}|\mathbf{E}, T). \tag{1}$$

To derive Equation (1) we use the fact that $P(\mathbf{E}, T) = P(\mathbf{E}) \cdot P(T)$ since attributes of different entities are independent (without conditioning on the existence of a link). Equation (1) says that LBC can be decomposed into a probabilistic classifier model for the single target table and a probabilistic model of the link structure conditional on the descriptive attributes of entities. The latter can be built using the following independence assumption.

Definition 1: The **Path Independence Assumption** (PI) states that different paths, represented in different rows in join tables, are conditionally independent given the attributes of the entities in the path:

$$P(\mathbf{J}|\mathbf{E}, T) = \prod_{i=1}^{m} \prod_{r=1}^{rows_i} P(R\_keys_{i,r}, R\_atts_{i,r}|E\_atts_{i,r}, T\_atts_{i,r}) \tag{2}$$

The symbol $rows_i$ is the number of rows in join table J_i that contain the target entity.

Example for Independence Assumption 2. Consider the university example. The Path Independence Assumption implies that given the attributes of Jack and the attributes of courses, the fact that Jack takes course 101 and his grade in 101, is independent of Jack's taking course 102 and his grade in 102. Using the values as shown in Figure 2, the computation defined by Equation (2) is as follows, given a possible class label c for class attribute $Intelligence$. Here and below we list the nonclass attributes in the order $Grade, Satisfaction, Ranking, Rating, Diff$. The first factor lists the values of these attributes for course 101, and the second for course 102.

$$P(J_2|\mathbf{E}, T) \propto P(R\_keys = [jack, 101], A, 1|Int = c, \mathit{1}, \mathit{3}, \mathit{1})$$
$$\times P(R\_keys = [jack, 102], B, 2|Int = c, \mathit{1}, \mathit{2}, \mathit{2}).$$

The next proposition derives the **Path Independence classification formula** from Equations (1) and (2).

Proposition 1: Given the Path Independence Assumption 2, the most probable class label c^* can be computed as follows:

$$\operatorname*{argmax}_c P(c(t)|\boldsymbol{a}(t)) \times$$
$$\prod_{i=1}^{m} \prod_{r=1}^{rows_i} \frac{P(c(t)|R\_keys_{i,r}, R\_atts_{i,r}, E\_atts_{i,r}, \boldsymbol{a}(t))}{P(c(t)|\boldsymbol{a}(t))}. \tag{3}$$

The proof is in the appendix. The formula (3) can be read as follows. For each path join table and for each row in it, compute the probability of the class label given the row entries, and divide the result by the class posterior in the target table. Thus the formula measures the additional information gained by considering each path, relative to considering only the target table.

Example for Classification Formula 3. Consider the table with index $i = 2$ in Formula (3), that is, join table J_2 from Figure 2. Considering only the rows with target entity $jack$, the inner product of Formula (3) corresponds to the following computation over the first two rows in the table (courses 101 and 102).

$$\frac{P(Int(jack) = c|R\_keys = [jack, 101], A, 1, 3, 1, 3, 1)}{P(Int(jack) = c|Rank = 1)} \times \frac{P(Int(jack) = c|R\_keys = [jack, 102], B, 2, 3, 1, 2, 2)}{P(Int(jack) = c|Rank = 1)} \tag{4}$$

Most multi-relational classifiers consider information from existing links only. Likewise, we provide classification formulas only for using existing links, as illustrated in the example. It is possible to derive closed-form formulas for absent links as well.

B. Discussion

1) Tabular Interpretation: Join tables have common fields from common entities, which leads to dependencies between rows, so the path independence assumption requires conditioning on this common information. This is an instance of the general principle that *structured objects may become independent if we condition on the components that they share.* For instance, the event that Jack registers in course 101 is assumed to be independent of the event that Jack registers in course 102, given the attributes of the courses (E_atts) and those of Jack (T_atts). These are the common fields in the two rows of the J_2 table of Figure 2.

2) Impact of PI Assumption: We emphasize that we do not claim that the Path Independence Assumption is exactly true in a given database. For example, whether Jack registers in one course is normally correlated with his registration in another course. Part of this correlation is mediated through the attributes of Jack and the courses, and can be adjusted for by conditioning on these variables. However, in general we would expect some correlation between registration events to persist even conditional on the attributes of Jack and the courses. Likewise, links from different tables may be correlated as well. The use of the path independence assumption is therefore best viewed as a simplifying approximation to the actual dependencies in the dataset. Path independence can be compared to the non-relational Naive Bayes assumption as follows.

1) Like the Naive Bayes assumption, Path Independence is often true enough to permit accurate predictions of an entity's attributes [7].
2) One way to view the Naive Bayes assumption is that it defines a choice of which correlations to model: it allows a model to present correlations between features and the class label, but not correlations among the features. Analogously, the path independence assumption allows a model to capture complex correlations between attributes and the class label given the path structure, but not to represent correlations among the paths. This is formally demonstrated by the PI classification formula (3), which incorporates dependencies of the class label on attributes of related entities and links, but not correlations among links.

The next section defines additional independence assumptions to construct a hierarchy of multi-relational Bayes net classifiers.

V. A Hierarchy of Multi-relational Bayes Net Classifiers

A. A Hierarchy of Independence Assumptions

We show that several previous proposals for classifications with independence models result from augmenting the path independence assumption with further principles. Figure 3 provides an overview of the independence assumptions. The graphs shown represent the assumptions as constraints on Bayes net structures (using d-separation [**?**].) It is difficult to represent Path Independence and Path-Class Independence in the same model using graph representation. In the model shown, Path-Class Independence could be encoded in the conditional probability parameters.

Definition 2: (1) The **Influence Independence Assumption** states that given the class label, information from the target table and from each entity-relationship table can be combined independently.

$$\begin{aligned} P(R\_keys_{i,r}, R\_atts_{i,r}, E\_atts_{i,r}, \boldsymbol{a}(t)|c(t)) = \\ P(R\_keys_{i,r}, R\_atts_{i,r}, E\_atts_{i,r}|c(t)) \cdot P(\boldsymbol{a}(t)|c(t)) \end{aligned} \tag{5}$$

(2) For the target table, the **Naive Bayes Assumption** is just the usual single-table principle which says that predictive features are independent given the class label:

$$P(\boldsymbol{a}(t)|c(t)) = \prod_{\boldsymbol{a}_j \in \boldsymbol{a}(t)} P(\boldsymbol{a}_j|c(t)). \tag{6}$$

For attributes from join tables, the **relational Naive Bayes Assumption** says that they are independent given the class label and given the existence of a pathway:

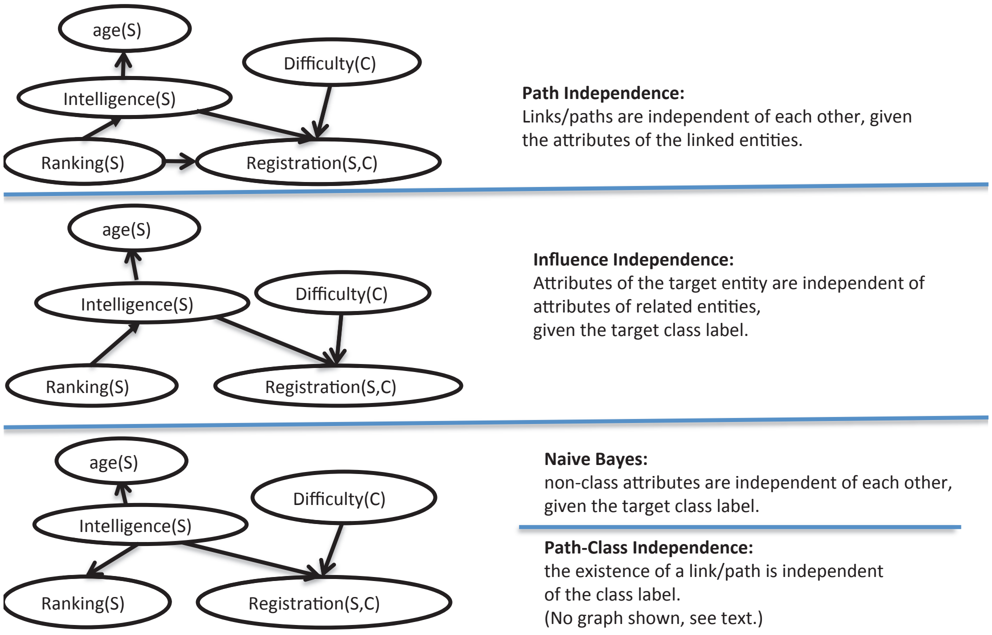
$$\begin{aligned} P(E\_atts_{i,r}, R\_atts_{i,r}|c(t), R\_keys_{i,r}) = \\ \prod_{z_{i,r} \in (E\_atts_{i,r} \cup R\_atts_{i,r})} P(z_{i,r}|c(t), R\_keys_{i,r}) \end{aligned} \tag{7}$$

for each row r containing the target entity in its key fields.

(3) The **Path-class Independence Assumption** says that the mere existence of a pathway, without further specification of attributes, is independent of the class label:

$$P(R\_keys_{i,r}|c(t)) = P(R\_keys_{i,r}). \tag{8}$$

We derive LBC formulas from these independence assumptions given the existing paths. Table III gives three different link-based classification formulas that are obtained by combining path-independence successively with the assumptions in Definition 2. Table IV illustrates the assumptions using the first row of join table J_2.



Path Independence:
Links/paths are independent of each other, given
the attributes of the linked entities.

Influence Independence:
Attributes of the target entity are independent of
attributes of related entities,
given the target class label.

Naive Bayes:
non-class attributes are independent of each other,
given the target class label.

Path-Class Independence:
the existence of a link/path is independent
of the class label.
(No graph shown, see text.)

Fig. 3. The Hierarchy of Independence Assumptions. Different classification models are obtained by successively adding assumptions. We added an age attribute for students for illustration. The Bayes nets represent a template that can be instantiated with the appropriate entities (e.g., each student instantiates S, each course instantiates C).

Classifier	Assumption	Classification: find class label c that maximizes the given expression
HNBC	Influence+PI	$P(c(t)\mid\boldsymbol{a}(t)) \cdot \prod_{i=1}^{m} \prod_{r=1}^{rows_i} \frac{P(c(t)\mid R\_keys_{i,r}, R\_atts_{i,r}, E\_atts_{i,r})}{P(c(t))}$
Exist+Naive	+ Naive Bayes	$P(c(t)) \cdot \prod_{\boldsymbol{a}_j \in \boldsymbol{a}(t)} P(\boldsymbol{a}_j\mid c(t)) \cdot \prod_{i=1}^{m} \prod_{r=1}^{rows_i} \prod_{z_{i,r} \in (E\_atts_{i,r} \cup R\_atts_{i,r})} P(z_{i,r}\mid c(t), R\_keys_{i,r}) \cdot P(R\_keys_{i,r}\mid c(t))$
MRNBC	+ Path-class	$P(c(t)) \cdot \prod_{\boldsymbol{a}_j \in \boldsymbol{a}(t)} P(\boldsymbol{a}_j\mid c(t)) \cdot \prod_{i=1}^{m} \prod_{r=1}^{rows_i} \prod_{z_{i,r} \in (E\_atts_{i,r} \cup R\_atts_{i,r})} P(z_{i,r}\mid c(t), R\_keys_{i,r})$

TABLE III
THEOREM: COMBINATIONS OF THE INDEPENDENCE ASSUMPTIONS AND THE ENTAILED CLASSIFICATION FORMULAS.

B. Discussion and Related Work

1) The Influence Independence Assumption: The Influence Independence Assumption states that given the class label, information from the target table and from each entity-relationship table can be combined independently. In our example university database, this implies that given the intelligence of Jack, the other attributes of Jack are independent of attributes of courses that Jack has taken; see Table IV(1). The classification formula Table III(1) is similar to the Path Independence Formula (3); the difference is that attributes of the target entity are no longer considered in the big product.

Manjunath *et al.* [23] have developed the HNBC classifier using the influence independence assumption as the basis of their model, though they did not define this assumption axiomatically. A difference is that our formula introduces a scaling factor $P(c(t))$ for each path that scales the impact of related entities by the prior class probability. Structured logistic regression also treats the influence on the class label of the target table and that of linked tables as independent [22], without a scaling factor.

2) The Naive Bayes Assumption: The Naive Bayes assumption is as usual for the single-table case, but also applied to the attributes of links and related entities. Getoor *et al.* have proposed the Exists+Naive Bayes PRM graphical model [13], [31] which combines the Naive Bayes assumption with the Path Independence graphical structure (link nodes are children only), for the specific case of the single link relationship nodes. Our formulation extends their model to the multi-relational case.

The Naive Bayes assumption entails independence *within* a DB table, whereas Path and Influence Independence entail independence *across* tables. Since database designers tend to group related features within a single table, independence assumptions across tables receive more support from the database schema design than those within tables [23].

3) The Path-Class Independence Assumption: Other Naive Bayes classification formulas for relational data do not weight information from linked entities by the probability

Independence Assumption	Example Factorization
Influence+Path Independence	$P(R\_keys = [jack, 101], Gr = A, Sat = 1, Rat = 3, Diff = 1, Rank = 1 \mid Int = c) =$ $P(R\_keys = [jack, 101], Gr = A, Sat = 1, Rat = 3, Diff = 1 \mid Int = c) \times P(Rank = 1 \mid Int = c)$
+ Naive Bayes	$P(Gr = A, Sat = 1, Rat = 3, Diff = 1 \mid Int = c, R\_keys = [jack, 101]) =$ $P(Gr = A \mid Int = c, R\_keys = [jack, 101]) \times P(Sat = 1 \mid Int = c, R\_keys = [jack, 101]) \times$ $P(Rat = 3 \mid Int = c, R\_keys = [jack, 101]) \times P(Diff = 1 \mid Int = c, R\_keys = [jack, 101])$
+ Path-class	$P(R\_keys = [jack, 101] \mid Int = c) = P(R\_keys = [jack, 101])$.

TABLE IV
APPLYING THE INDEPENDENCE ASSUMPTIONS TO ROW 1, TABLE J_2 OF FIGURE 2.

that the link exists given the class label. This corresponds to the path-class independence assumption. The classification formula III(3) is equivalent to the state-of-the-art multi-relational Naive Bayes Classifier of Han et al. [5]. The formula can therefore be seen as making explicit the independence assumptions used in this Naive Bayes Classifier. The fact that previous research has (implicitly) made use of our independence assumptions is evidence that they are useful and natural for link-based classification. At the same time, our hierarchical framework shows that previous Naive Bayes models [13], [5] incorporate significant relational assumptions in addition to the standard single-table factoring of descriptive features.

VI. Learning

In terms of computational feasibility, a key aspect of the Path Independence Assumption is that learning can be carried out *joinwise*: Given a set of extended join tables, we can learn a probabilistic classifier on each join table separately, and combine the results using the product formula 3. The same point holds for classification models that add further assumptions to Path Independence.

Base Classifiers. In this paper we use Bayes net classifiers as a base classifier model. Bayes nets provide powerful probabilistic classifiers. In addition, they incorporate minimal assumptions about the distribution to be learned, which makes them especially suitable for investigating our hierarchy of multi-relational classification formulas: Bayes nets do not confound the effects of the relational assumptions with other model-specific assumptions.

Building Join Tables. Researchers have proposed several ways to construct an extended database [5], [23]. We adopted the simplest approach, which is to enumerate all join paths satisfying foreign key constraints in a data pre-processing step. Self-joins and other repeated occurences of the same table can be handled by duplicating the table [5]. Because of foreign-key constraints, the number of valid joins is typically much smaller than the total number of possible joins, and was feasible for our benchmark datasets.

VII. Evaluation

To evaluate our proposed model we compare it with 5 other classifiers on four benchmark datasets on Accuracy (percentage of correctly classified instances), AUC (Area under the ROC Curve), and runtime. AUC varies the acceptance threshold for a class label and is therefore an aggregate measure of the false positive/false negative error rates of the classifiers. Runtimes were measured on a QUAD CPU Q6700 with a 2.66GHz CPU and 8GB of RAM. Our code and datasets are available on-line from ftp://ftp.fas.sfu.ca/pub/cs/oschulte/sdf. We investigate the following issues.

1) Compared to general multi-relational classification methods, whether using the path independence assumption leads to fast learning times with a competitive classfication performance.
2) The negative impact of the influence independence assumption: analyzing the descriptive attributes of the target entity in conjunction with information from links helps classification performance in comparison with using information from links separately.
3) The negative impact of the Naive Bayes assumption: Taking into account correlations among features within the same (join) table helps classification, compared to assuming their independence conditional on the class label.
4) The negative impact of the path-class independence assumption: weighting the attribute information from a link by how probable the link is given the class label is beneficial.

A. Datasets

We used four benchmark relational datasets.

Hepatitis Database. This data is from the PKDD'02 Discovery Challenge database, with the modifications of [11]. $Biopsy$ is the target table with 206 instances of Hepatitis B, and 294 cases of Hepatitis C. The $inhospital$ table was modified such that we put each unique test in a column and all tests for a patient on a given year in a single tuple. Removing tests with null values, 10 different tests as the table descriptive attributes remained.

Financial Database. This data is from the PKDD CUP 1999 database. We followed the modifications of [5], and chose a subset of the target class $Loan$ (status) with 333 positive, and 76 negative instances. The other entity tables we used are: $Account$, $Order$, $Disposition$, and $Transactions$.

MovieLens Database. This dataset is drawn from the UC Irvine machine learning repository. It contains two entity tables: $User$ and $Item$, and one relationship table $Rated$ with 80,000 ratings. The table $Item$ has 17 Boolean attributes that indicate the genres of a given movie. The class label is the user attribute age that we discretized into three bins with equal frequency.

Mondial Database. This dataset contains data from multiple geographical web data sources [24]. We predict the religion of a country as Christian (positive) with 114 instances vs. all other religions with 71 instances. We followed the modification of [30].

B. Experimental Design

We compared the following multi-relational classifiers using 10 fold cross validation.

PIC (Path Independence Classifier) Bayes net learner with only the path independence assumption; see Formula 3.

HNBC (Heterogeneous Naive Bayes Classifier) see Table III(1): Bayes net learner with the additional assumption of independency between attributes across tables .

E-NB (Exists+Naive Bayes) see Table III(2): Relational Naive Bayes Classifier with path-class dependencies.

MRNBC (Multi Relational Naive Bayesian Classifier) Relational Naive Bayes Classifier with the combination of the four independence assumptions.

MLN (Markov Logic network) We used the state-of-the-art learn-and-join (LAJ) structure learning algorithm for MLNs [29], default discriminative parameter learning and MC-SAT inference algorithm implemented in the open-source Alchemy package [19]. We used the LAJ algorithm because its predictive accuracy outperforms other MLN structure learning algorithms.

Tilde (Top-Down Induction of First-Order Logical Decision Trees) Tilde is a multi-relational decision tree classifier [3]. Tilde is implemented in the ACE data mining system [2]. We ran Tilde with the default setting.

To learn the structure under the path independence and influence independence assumptions, for a single table Bayes net learner we apply the GES search algorithm [6]. We use a generative learner because the Exists+NB and MRNBC classifiers are generative models, so our experiments avoid conflating the impact of different independence assumptions with the impact of generative vs. discriminative training. For parameter learning, we use the standard maximum likelihood estimates (empirical frequencies) for generative models.

C. Results

We first consider training time for the models, then their classification performance.

1) Runtimes: Table V reports the combined runtimes of the structure learning and parameter learning of different algorithms. Under the influence independence resp. Naive Bayes assumptions, the structure learning runtime is basically the sum of the runtimes for applying the single-table Bayes net learner resp. Naive Bayes learner to different join tables. For MLN, the runtime is calculated by adding the structure learning time of the learn-and-join algorithm with the parameter learning time of the Alchemy package. NT stands for “Not Terminated“ within 72 hours. For Tilde we report the decision tree induction time. The Path Independence Classifier allows for the most complex cross-table dependencies and therefore takes the most time, but it is still very fast even on fairly complex data sets like *Financial* and *MovieLens*. Overall the *independence-based methods are much faster than the comparison methods*, by 2 or 3 orders of magnitude depending on the method and the dataset.

	Bayes Net Classifiers				Other Methods	
Dataset	PIC	HNBC	E-NB	MRNBC	MLN	Tilde
Hepatitis	7.43	7.01	**2.07**	**2.07**	3902	853
Financial	28.31	23.21	**15.01**	**15.01**	NT	2429
MovieLens	25.32	17.67	**5.31**	**5.31**	960	1100
Mondial	5.41	5.08	1.89	1.89	5.44	**0.3**

TABLE V
TRAINING TIME OF DIFFERENT MODELS IN SECONDS. **MONDIAL NEEDS TO BE CHECKED.**

2) Classification Performance: Our performance measures are accuracy (percentage of correctly classified target instances), f-measure, and Area-Under-Curve (AUC), shown in Table VI. The F-measure is defined as [33, p.146]

$$\frac{2(\text{True Positive})}{2(\text{True Positive}) + (\text{False Positive}) + (\text{False Negative})}.$$

For the multi-class problems, we report only accuracy, since there is no standard way to extend F-measure and AUC to multi-class problems [9], and since the three measures are highly correlated on the binary class problems.

Accuracy	Bayes Net Classifiers				Reference Methods	
Dataset	PIC	HNBC	E-NB	MRNBC	MLN	Tilde
Hepatitis	**0.80**	0.78	0.78	0.74	0.77	0.61
Financial	**0.91**	0.90	0.89	0.81	NT	0.89
MovieLens	**0.66**	0.57	0.53	0.50	0.484	0.48
Mondial	0.85	0.82	0.78	0.82	0.76	0.71

F-measure	Bayes Net Classifiers				Reference Methods	
Dataset	PIC	HNBC	E-NB	MRNBC	MLN	Tilde
Hepatitis	**0.78**	0.76	0.75	0.76	0.68	0.59
Financial	**0.91**	0.88	0.84	0.81	NT	0.88
Mondial	**0.82**	0.77	0.78	0.75	0.75	0.78

AUC	Bayes Net Classifiers				Reference Methods	
Dataset	PIC	HNBC	E-NB	MRNBC	MLN	Tilde
Hepatitis	**0.85**	0.83	0.83	0.83	0.78	0.61
Financial	**0.85**	**0.85**	**0.85**	0.82	NT	0.69
Mondial	**0.9**	0.89	0.83	0.81	0.82	0.75

TABLE VI
PREDICTIVE PERFORMANCE OF DIFFERENT CLASSIFIERS BY EVALUATION METRIC.

We make the following observations from our experimental results.

1) The independence-based methods as a whole outperform the alternative approaches (MLN and Tilde).
2) The general Bayes net classifier with the *Path Independence assumption* achieves the best classification performance compared to the other Bayes net relational classifiers.
3) A consistent improvement in accuracy results from *not using the path-class independence assumption*. This suggests that taking into account correlations between the class label and the existence of links is beneficial.

4) With fewer independence assumptions, classification performance tends to be better. However, once path-class independence is dropped, the effect can be weak (1-2% improvement except for Mondial).

In the MovieLens dataset, the Path Independence Classifier achieves an unusually high performance compared to alternative methods. The reason is that, as Bayes net learning indicates, given the target attributes from the target table $User$, information from related entities (rated movies) is irrelevant. In such a case the irrelevant information hurts classification accuracy. The PI formula (3) correctly cancels out the irrelevant information, whereas the influence independence formula from Table III(2) retains it.

On the other datasets, the improvement from using Path Independence rather than Influence Independence is small and not always statistically significant. This is not surprising because the difference in the corresponding classification formulas is small (cf. Section V-A). Bina *et al.* have shown that learning weights for different extended join tables improves classification with the Influence Independence assumption significantly [1]. Extending the Path Independence Classifier with weights should work even better; we leave this for future work.

Overall, our results provide evidence that the path independence model makes an attractive trade-off between improving classification accuracy with a modest increase in learning time. It is sufficiently strong to guarantee fast multi-relational learning, almost as fast as the much stronger relational Naive Bayes assumptions. Making weaker assumptions permits a model to capture more relevant correlations, which leads to predictive performance that is consistently as good compared to making stronger assumptions, and sometimes much better.

VIII. Conclusion

The large number of different dependencies of different types that are potentially relevant for link-based prediction are a challenge for model selection algorithms. We propose a novel new classifier based on the *path independence assumption* with an efficient Bayes net model learning algorithm. We developed a hierarchical framework for unifying and investigating model classes for relational data based on three further independence assumptions, with path independence at the top and multi-relational Naive Bayes at the bottom. We derived closed-form classification formulas for each hierarchy level. In our experiments on four benchmark datasets fewer independence assumptions led to higher classification accuracy, with only a modest increase in runtime.

References

[1] B. Bina, O. Schulte, B. Crawford, Z. Qian, and Y. Xiong. Simple decision forests for multi-relational classification. *Decision Support Systems*, in press, available on-line at:http://dx.doi.org/10.1016/j.dss.2012.11.017, 2013.

[2] H. Blockeel, L. Dehaspe, J. Ramon, J. Struyf, A. Van Assche, C. Vens, and D. Fierens. *The ACE Data Mining System: Users Manual*. http://dtai.cs.kuleuven.be/ACE/doc/ACEuser-1.2.16.pdf, 2009.

[3] H. Blockeel and L. D. Raedt. Top-down induction of first-order logical decision trees. *Artificial Intelligence*, 101(1-2):285–297, 1998.

[4] M. Ceci, A. Appice, and D. Malerba. Mr-sbc: A multi-relational naive Bayes classifier. In *PKDD*, pages 95–106, 2003.

[5] H. Chen, H. Liu, J. Han, and X. Yin. Exploring optimization of semantic relationship graph for multi-relational Bayesian classification. *Decision Support Systems*, 48:112–121, 2009.

[6] D. M. Chickering and C. Meek. Finding optimal Bayesian networks. In *UAI*, pages 94–102, 2002.

[7] P. Domingos and M. Pazzani. Beyond independence: Conditions for the optimality of the simple Bayesian classifier. In *International Conference on Machine Learning*, pages 105–112. Morgan Kaufmann, 1996.

[8] P. Domingos and M. Richardson. Markov logic: A unifying framework for statistical relational learning. In *Introduction to Statistical Relational Learning* [14].

[9] R. Espíndola and N. Ebecken. On extending f-measure and g-mean metrics to multi-class problems. *Data mining VI: Data mining, text mining and their business applications*, 35:25–34, 2005.

[10] P. A. Flach and N. Lachiche. Naive Bayesian classification of structured data. *Mach. Learn.*, 57(3):233–269, 2004.

[11] R. Frank, F. Moser, and M. Ester. A method for multi-relational classification using single and multi-feature aggregation functions. In *SIGKDD*, 2007.

[12] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *In IJCAI*, pages 1300–1309. Springer-Verlag, 1999.

[13] L. Getoor, E. Segal, B. Taskar, and D. Koller. Probabilistic models of text and link structure for hypertext classification. In *IJCAI01 Workshop on Text Learning: Beyond Supervision*, Seattle, WA., 2001.

[14] L. Getoor and B. Tasker. *Introduction to statistical relational learning*. MIT Press, 2007.

[15] L. G. Getoor, N. Friedman, and B. Taskar. Learning probabilistic models of relational structure. In *ICML*, pages 170–177. Morgan Kaufmann, 2001.

[16] T. N. Huynh and R. J. Mooney. Discriminative structure and parameter learning for markov logic networks. In *ICML*, pages 416–423, 2008.

[17] D. Jensen, J. Neville, and B. Gallagher. Why collective inference improves relational classification. In *SIGKDD*, pages 593–598. 2004.

[18] K. Kersting and L. de Raedt. Bayesian logic programming: Theory and tool. In *Introduction to Statistical Relational Learning* [14], chapter 10, pages 291–318.

[19] S. Kok, M. Summer, M. Richardson, P. Singla, H. Poon, D. Lowd, J. Wang, and P. Domingos. The Alchemy system for statistical relational AI. Technical report, University of Washington., 2009.

[20] W. V. Laer and L. de Raedt. How to upgrade propositional learners to first-order logic: A case study. In *Relational Data Mining*. Springer Verlag, 2001.

[21] N. Landwehr, K. Kersting, and L. D. Raedt. nfoil: Integrating naïve bayes and foil. In *AAAI*, pages 795–800, 2005.

[22] Q. Lu and L. Getoor. Link-based classification. In *ICML*, 2003.

[23] G. Manjunath, M. N. Murty, and D. Sitaram. A practical heterogeneous classifier for relational databases. In *ICPR*, pages 3316–3319, 2010.

[24] W. May. Information extraction and integration with florid: The mondial case study. Technical report, Universitat Freiburg, Institut fur Informatik, 1999.

[25] J. Neville, D. Jensen, B. Gallagher, and R. Fairgrieve. Simple estimators for relational bayesian classifiers. In *Proceedings of the 3rd IEEE international conference on data mining*, pages 609–612. Citeseer, 2003.

[26] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

[27] A. Popescul and L. Ungar. Feature generation and selection in multi-relational learning. In *An Introduction to Statistical Relational Learning* [14], chapter 8.

[28] T. Roos, H. Wettig, P. Grünwald, P. Myllymäki, and H. Tirri. On discriminative bayesian network classifiers and logistic regression. *Mach. Learn.*, 59(3):267–296, 2005.

[29] O. Schulte and H. Khosravi. Learning graphical models for relational data via lattice search. *Machine Learning*, 88:3:331–368, 2012.

[30] R. She, K. Wang, and Y. Xu. Pushing feature selection ahead of join. 2005.

[31] B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *UAI*, pages 485–492, 2002.

[32] J. D. Ullman. *Principles of database systems*. 2. Computer Science Press, 1982.

[33] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2 edition, 2005.

IX. Appendix: Derivation of Classification Formulas in Table III.

a) Proof of Path Independence Classification Formula (3)*:* From the definition of Path Independence (2) the classification problem is equivalent to the maximization

$$\operatorname*{argmax}_{c} P(c(t)|\boldsymbol{a}(t))\cdot \prod_{i=1}^{m}\prod_{r=1}^{rows_i} P(R\_keys_{i,r}, R\_atts_{i,r}|E\_atts_{i,r}, c(t), \boldsymbol{a}(t)) \tag{9}$$

where we have used the fact that rows not containing the target entity t are irrelevant to the maximization, and that in rows that do contain it, and the target table attributes T comprise the vector $(c(t), \boldsymbol{a}(t))$. Using Bayes' theorem we have

$$P(R\_keys_{i,r}, R\_atts_{i,r}|E\_atts_{i,r}, c(t), \boldsymbol{a}(t)) \propto \frac{P(c(t)|R\_keys_{i,r}, R\_atts_{i,r}, E\_atts_{i,r}, \boldsymbol{a}(t))}{P(c(t)|E\_atts_{i,r}, \boldsymbol{a}(t))} \tag{10}$$

where we have omitted a factor that does not depend on $c(t)$.

We now use the independence fact

$$P(c(t)|E\_atts_{i,r}, \boldsymbol{a}(t)) = P(c(t)|\boldsymbol{a}(t)) \tag{11}$$

which holds since attributes of entities distinct from the target entity are independent of the target class (unless we condition on the existence of a relationship). Now substituting Equation (11) into Equation (10) and applying the result to Equation (9) yields Formula (3).

b) Proof of the Formulas in Table III: We begin with a preliminary observation.

Lemma 1: Suppose that P satisfies the Influence Independence Assumption, and consider a row r in join table J_i. Then

$$P(R\_keys_{i,r}, R\_atts_{i,r}|E\_atts_{i,r}, \boldsymbol{a}(t), c(t)) = \frac{P(R\_keys_{i,r}, R\_atts_{i,r}, E\_atts_{i,r}|c(t))}{P(E\_atts_{i,r})}.$$

The lemma follows easily using the independence fact

$$P(c(t), E\_atts_{i,r}, \boldsymbol{a}(t)) = P(c(t), \boldsymbol{a}(t)) \cdot P(E\_atts_{i,r})$$

as in Equation (11).

Table III, Line 1: From Equation (9), Lemma 1, and the fact that $P(E\_atts_{i,r})$ does not depend on $c(t)$, the classification problem is given by

$$\operatorname*{argmax}_{c} P(c(t)|\boldsymbol{a}(t)) \cdot \prod_{i=1}^{m}\prod_{r=1}^{rows_i} P(R\_keys_{i,r}, R\_atts_{i,r}, E\_atts_{i,r}|c(t)). \tag{12}$$

The formula of Line 1 follows by applying Bayes' theorem and omitting the term $P(R\_keys_{i,r}, R\_atts_{i,r}, E\_atts_{i,r})$ which does not depend on the class label.

For Line 2 we use the equality

$$P(R\_keys_{i,r}, R\_atts_{i,r}, E\_atts_{i,r}|c(t)) = P(R\_atts_{i,r}, E\_atts_{i,r}|R\_keys_{i,r}, c(t)) \cdot P(R\_keys_{i,r}|c(t)). \tag{13}$$

Substituting this result into Equation (12) and applying the Naive Bayes independence assumption establishes Line 2. The Path-Class Independence Assumption entails that the term $P(R\_keys_{i,r}|c(t))$ may be omitted from Equation (13) for classification, which establishes Line 3.