# Probabilistic Relational Models with Relational Uncertainty: An Early Study in Web Page Classification

E. Fersini, E. Messina, F. Archetti
*Department of Informatics, Systems and Communication*
*University od Milano-Bicocca, Milan, Italy*
{*fersini, messina, archetti*}*@disco.unimib.it*

***Abstract*—In the last decade, new approaches focused on modelling uncertainty over complex relational data have been developed. In this paper one of the most promising of such approaches, known as Probabilistic Relational Models (PRMs), has been investigated and extended in order to measure and include uncertainty over relationships. Our extension, called PRMs with Relational Uncertainty, has been evaluated on real-data for web document classification purposes. Experimental results shown the potentiality of the proposed methods of capturing the real "strength" of relationships and the capacity of including this information into the probability model.**

## I. Introduction

Traditional machine learning approaches are consistent with the classical statistical inference problem formulation, in which instances are homogeneous, independent, identically distributed (i.i.d. assumption) and consequently represented as a single flat table. However, this kind of problem formulation does not reflect what happens in the real world, where data are represented by several interrelated entities. If we consider the problem of inducing a predictive model starting from relational data, this task can be accomplished by following one of the subsequent directions: (1) by mapping relational data into a propositional form and applying traditional predictive models based on the i.i.d. assumption; (2) by maintaining the relational data structure and by adapting/enriching learning and/or inference algorithms in order to consider interrelated instances. If we choose the first direction, having a propositional representation obtained by flattening relational data, we can induce the simplest predictive model by specifying a complete joint probability distribution. In this case taking a set of random variables, we must specify a probability value for each of the possible set instantiation. For example, in order to specify an arbitrary joint distribution $P(X^1, X^2, ..., X^m)$ for $m$ dichotomous variables, a table with $2^m$ entries is required. This complexity makes the representation of the probability model impractical for any domain of realistic size. A possible solution that tries to overcome this problem is represented by Bayesian Networks (BNs) [18]. The key component of BNs that reduces the probability model complexity, is the assumption that each variable is directly influenced by only few others. This assumption is captured graphically by a dependency structure $D$: a probability distribution is represented as a directed acyclic graph whose nodes represent random variables and whose edges denote direct dependencies (causal relationships) between a node $X$ and its set of parents $Pa(X)$. Formally, a Bayesian Network asserts that each node is conditional independent of its non-descendants given its parents. This conditionally independence assumption allows us to represent concisely the joint probability distribution. If we consider a distribution $P$ over $m$ feature, it can be decomposed as the product of $m$ conditional distributions:

$$\begin{aligned} P(x^1, x^2, ..., x^m) &= \prod_j P(x^j | x^1, ..., x^{j-1}) = \\ &= \prod_j P(x^j | Pa(x^j)) \end{aligned} \quad (1)$$

where $P(x^i|Pa(x^i))$ is described by a conditional probability distribution (CPD). The set of local conditional distributions represents the parameters $\theta_D$. Despite the great success of Bayesian Networks, both in terms of compact representation and predictive power, their applicability to complex relational domains is still relatively limited. This is due to one main reason: learning and inference algorithms are based on independent instances assumption even if real world data are represented as interrelated objects. Although several methods have been developed in order to create propositional representations that account for relationships [14] [17], the exploitation of relationships during learning and inference is not admitted. This causes a lost of information about relationship between instances that could contribute to increase the predictive power of the classification model. In order to overcome these limitations a growing number of relational approaches have been proposed as Stochastic Logic Programs [2], Relational Bayesian Networks [11], Bayesian Logic Programs [12], Relational Dependency Networks [15] [16], Markov Logic Networks [4] [3], etc.... Among them we choose Probabilistic Relational Models (PRMs) - introduced by Koller in [13] and subsequently investigated by Getoor et al. in [9], [10] and [7] - that are able to model the intersection between uncertainty and relational representation through the conjunction of Bayesian Networks and Relational Databases. The advantages of PRMs are their compact representation and the possibility to efficiently

exploit (with proper adaptation) all available existing tools for learning and inference on Bayesian Networks.
In this paper PRMs have been extended in order to deal with uncertainty over the relational structure and experimentally evaluated on real-data for web document classification purposes.
The outline of the paper is the following. In section II we review traditional PRMs. In Section III our extension of PRMs, focused on web document classification purposes, is presented in order to model the uncertainty over the relational structure. In Section IV an experimental evaluation, that includes a comparison with Bayesian Networks and available PRMs, is conducted on a real case-study. Finally, in Section V a summary of the main contributions of the paper is given and in Section VI conclusions are derived.

## II. Probabilistic Relational Models

Probabilistic Relational Models (PRMs) are a rich representation language, used to combine relational representation with probabilistic directed graphical models. PRMs conceptually extend Bayesian Networks in order to incorporate the relational structure during learning and inference processes. From a high level point of view, PRMs specify a probability model over attributes of interrelated objects rather then over features of propositional samples. The simplest form of PRMs was introduced by Friedman in [6]. In order to describe PRMs, we need to introduce their key components: (1) a relational schema $S$; (2) a dependency structure $D_S$ over attributes with the associated parameters $\theta_{D_S}$; (3) a joint probability distribution $P$.

### A. Relational Schema

A relational schema describes the structure of a relational database. In particular, a relational schema $S$ defines a set of $n$ entities (tables) $E = \{E_1, E_2, ..., E_n\}$, the set of descriptive attributes $\mathcal{A}(E_i)$ associated to each entity $E_i$ and a reference slot $r$ every two related entities. Entities are conventionally indicated by capital letter, while their records - usually called *instances* or *objects* - are referred by the corresponding lower case.
An important role into PRMs is played by those components called *relational skeleton* and *schema completion*.

A **relational skeleton** $\sigma_S$ is defined to be a partial specification of the schema, where objects and reference slots are specified, while attribute values are left unknown.

A **schema completion** $I$ specifies the objects $e_{it}$ that instantiate schema entities $E_i$ and the relationships that exist among them. A completion $I$ specifies therefore a value for each attribute $e_{it}.A$ and a value for each reference slot $e_{it}.r$.

### B. Dependency Structure and CPDs

A PRM specifies a probability distribution over entity attributes by using the same assumption at the base of Bayesian Networks: each variable in a PRM, i.e. attribute $E_i.A$ of an entity, is directly influenced by only few others. This assumption is captured graphically by a dependency structure $D_S$: each attribute $E_i.A$ is represented by a node and the causal dependencies between them are denoted by directed edges. Indeed, PRMs define for each attribute $E_i.A$ a set of parents $Pa(E_i.A)$ that provides direct influences on it.
Having a dependency structure that establish causal relationships between entity attributes, a conditional probability distribution (CPD) is associated to each node of the network. More precisely, let $E_i.A$ be an attribute that can assume a set of $U(E_i.A)$ values and let $Pa(E_i.A)$ be the parent set of $E_i.A$. The CPD for $E_i.A$ is defined by counting how many realizations $u_k \in U(E_i.A)$ occur conditioned to the joint parent values $v_{Pa(E_i.A)}$. Except in cases of one "one-to-one" relationships, i.e. when an object $e_{it}$ is related to only one object $e_{jv}$, the problem of modelling multiple related value dependencies needs to be addressed for CPDs computation. In this case, an aggregation function $\gamma$ (as for example mode, median, maximum, etc) is applied in order to model the dependency of a variable $E_i.A$ with respect to its multi valued parent $\gamma(E_i.r.B)$. The CPDs, that defines the set of parameters $\theta_{D_S}$, can be estimated directly on relational data and, since they are defined at entity level, they can be shared by all the objects that instantiate a given entity. For example if an entity $E_i$ contains an attribute $E_i.A$ and $e_{i1}$ and $e_{i2}$ are instances of $E_i$, then $e_{i1}.A$ and $e_{i2}.A$ will share the same conditional probability distribution that will be used during inference processes.

### C. Joint Probability Distribution

A **PRM** $\Phi$ for a relational schema $S$ is defined as a probabilistic graphical model characterized by:

- a set of parents $Pa(E_i.A)$, where each parent has the form $E_i.B$ or $\gamma(E_i.r.B)$;
- a set of parameters $\theta_{D_S}$, where each CPD represents $P(E_i.A|Pa(E_i.A))$

The main goal of PRMs is to define a distribution over relational data, i.e. a distribution over possible completions of the relational schema that are consistent with the relational skeleton.

Given any relational skeleton $\sigma_S$ the joint probability distribution over all completions of the relational schema $S$ can be computed as:

$$\begin{aligned} P(I|\sigma_S, S, \theta_{D_S}) &= \prod_{e_{it} \in \sigma_S} \prod_{A \in \mathcal{A}(e_{it})} P(I_{Pa(e_{it}.A)}, \theta_{D_S}) \\ &= \prod_{E_i \in E} \prod_{A \in \mathcal{A}(E_i)} \prod_{e_{it} \in \sigma_S(E_i)} P(I_{e_{it}.A}|I_{Pa(e_{it}.A)}, \theta_{D_S}) \end{aligned} \quad (2)$$

where

- $I_{e_{it}.A}$ represents the value of $e_{it}.A$ in the completion $I$,
- $E_i$ denotes the entities belonging to the relational schema $S$,

- $D_S$ represents the dependency structure
- $\theta_{D_S}$ are the parameters associate to the dependency structure $D_S$
- $\sigma_S(E_i)$ denotes the set of objects of each entity defined by the relational skeleton $\sigma_S$

The joint probability distribution of PRMs, described in equation (2) needs to be computed by considering a given relational skeleton $\sigma_S$, which specifies those objects that we have to consider in order to infer a given variable value. In order to explicitly use all the related objects of a given variable during the inference process, a PRM induces a Unrolled Bayesian Network.

Given a PRM $\Phi$ an Unrolled Bayesian Network is defined as follows:

- There is a node for every attribute $e_{it}.A$ of every object $e_{it} \in \sigma_S(E_i)$
- Each $e_{it}.A$ is probabilistically influenced by parents of the form $e_{it}.B$ or $(e_{it}.r.B)$ where $r$ is a slot chain
- The CPD for $e_{it}.A$ is defined as $P(E_i.A|Pa(E_i.A))$

The joint probability distribution can therefore be computed over the unrolled Bayesian Network as:

$$P(I|\sigma_S, \Phi) = \prod_{e_{it} \in \sigma_S(E_i)} \prod_{A \in \mathcal{A}(e_{it})} P(e_{it}.A|Pa(e_{it}.A)) \quad (3)$$

In order to derive a coherent probability distribution the dependency structure on the unrolled Bayesian Network is required to be acyclic [7].

An interesting remark about PRMs concerns the relational structure: PRMs assume that the skeleton is a certain input for inferring variables distribution. Thus, equation (3) determines the probabilistic model of object attributes, do not providing a probabilistic model of relationships between objects. This assumption limits the use of PRMs to that domains in which the relationships between objects, i.e. the reference slots, are fixed and certain in advance. These challenges have been addressed in the literature by investigating PRMs with Structural Uncertainty [9]. Two extension of PRMs have been proposed: Reference Uncertainty and Existence Uncertainty.

### D. PRM Extensions: Reference and Existence Uncertainty

**PRMs with Reference Uncertainty** are a form of PRMs able to deal with those domains in which the existence of reference slots are uncertain, but the number of related objects are known. This model specifies a relational skeleton $\sigma_o$ in which each object that instantiates a class is given and creates a probabilistic model for the value of each reference slot $e_{it}.r \in \sigma_o$. For more details refer to [7].

**PRMs with Existence Uncertainty** are a form of PRMs where not only the existence of reference slots is uncertain, but also the number the number of related objects is unknown. In order to manage this kind of uncertainty over the relational structure, all the objects that can *potentially* exist into the model are considered. Then a special binary attribute, called *exists variable*, is introduced in order to model the existence of a relationship between two *potential* objects. For more details about PRMs with Existence Uncertainty refer to [10].

Now consider the web document classification problem, in which documents are objects and hyperlinks are evidence of relationships. Getoor et al. used PRMs with Existence Uncertainty in order to address this problem. They assert that while we know the set of web pages, we may uncertain about which web pages link to each other and this we have uncertainty over the existence of a link. In their model links may or may not exist, and therefore they could be modelled by a binary random variable.

In this paper we want to consider not only the existence of a link, but also if the link is an expression of a semantic coherence between linked pages: in some cases a link can positively contribute to the inference process, while in some other cases a link could only add noise. For example the link from a page speaking about *football* to a page for the *pdf reader download* assumes less importance than a link between two *football* web pages. In this direction, in which a relation between two objects can be represented by a probabilistic connection, we propose a further extension of PRMs nemed PRM with Relational Uncertainty.

## III. Probabilistic Relational Models with Relational Uncertainty

The probability model provided by traditional PRMs, as mentioned above, have a limitation over the relational structure: uncertainty over relationships is not admitted and therefore parent attributes have the same impact on inferring child distribution.

**PRMs with Relational Uncertainty**, called PRMs/RU, are a form of Probabilistic Relational Models able to deal with domains in which the existence of a reference slot is known, but the relationships that they model could assume different degrees of "strength". That is probabilistic relationships, that model the "strength" of relationships between two related objects, are introduced in PRMs.

Consider for instance the inference process for the simplified unrolled citation graph reported in Figure 1. We can state that the attribute $p_7.topic$ depends on its internal parent $p_7.author$ and on its external related parent attribute $p_9.topic$. If we consider the evidence that the cited paper $p_9$ is about $math$, our expectation in traditional PRM about $p_7.topic$ likely tends to the same topic. However, if we consider the "strength" of this relationship our expectation could be change. If we know that $p_9$ is cited for related aspects, and therefore there exists a weak relationship, we could smooth our expectation accordingly. The reasoning about the citation graph is applicable also to the web domain: the topic of a web page could be inferred by considering

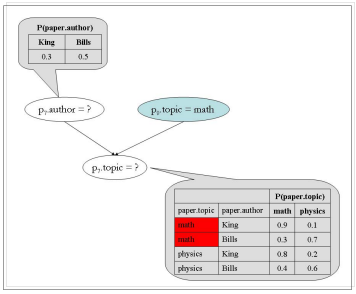


Figure 1. Toy Example of PRMs/RU

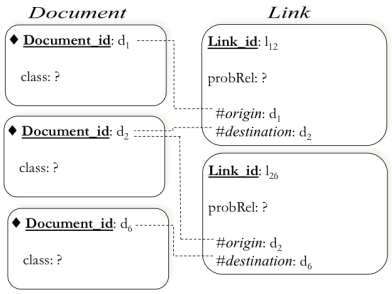


Figure 2. Probabilistic relational skeleton for the web

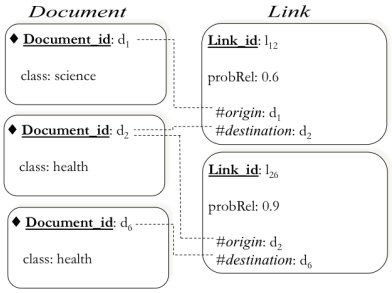


Figure 3. Completion of PRM/RU

the topic of related pages and by taking into account the "strength" of their relationships.

Thus, we can intuitively think at PRMs with Relational Uncertainty as a probabilistic graphical model in which the local distribution of an attribute is influenced not only by attributes of its related parents but also by the "strength" of their relationships. In order to describe PRMs/RU, we need to provide a new definition of (1) the relational schema, (2) the dependency structure over attributes with the associated set of parameters, (3) the joint probability distribution.

### A. Relational Schema

In order to include relational uncertainty in PRMs a new probabilistic skeleton $\sigma_P$ is proposed.

Given a relational schema $S$, a **probabilistic relational skeleton** $\sigma_P$ specifies objects belonging to the following entities:

- **Standard Entities** $E_i$, which refer to existing tables within the relational schema $S$ and are characterized by their own set of attributes $\mathcal{A}(E_i)$
- **Support Entities** $\widetilde{E}_j$, that do not explicit belong to the relational schema, but they are introduced to model probabilistic relationships between couple of objects. Each support entity contains:
  - two **reference slots** $\widetilde{E}_j.r1$ and $\widetilde{E}_j.r2$ for identifying related objects
  - one **artificial relationship attribute** $\widetilde{E}_j.U$, that represents the probabilistic relationship holding between two related objects.

A simple example of probabilistic relational skeleton for the web domain is depicted in Figure 2, where objects $d_i$ belong to the standard entity *Document*, objects $l_{ij}$ belong to the support entity *Link* and the relationship is denoted by the *probRel* attribute.

In PRMs/RU a **schema completion** $I^*$ specifies the value for each attribute $e_{it}.A$ belonging to standard entity $E_i$, the value for each attribute $\widetilde{e}_{jp}.U$ belonging to support entity $\widetilde{E}_j$ and the reference slots $\widetilde{e}_{jp}.r1$ and $\widetilde{e}_{jp}.r1$. A schema completion $I^*$ for the web domain specifies reference slots $l_{ij}.origin$ and $l_{ij}.destination$, the value of each object attribute, i.e. document category $d_i.class$, and the value of probabilistic relationship $l_{ij}.probRel$. (See Figure 3).

*1) Relational Uncertainty Over the Web Structure:* In order to measure the uncertainty over the web, we need to estimate the hyperlinks "strength". For this purpose, we can use the procedure proposed in [5]. In particular, we need to (1) identify which semantic portion of a web page contains a link and (2) evaluate its coherence w.r.t. the origin and the destination pages. The first activity is aimed at partitioning a html document, following the idea that it can be viewed not as an atomic unit but also as a granular composition of sub-entities. A multi-entity document representation, that reflects how people perceive a web page, help us to identify the semantic area of web pages in which a hyperlink is wrapped. For this purpose we can see a web page $d_i$ belonging to a document collection $Q$ as:

$$\begin{array}{ll} d_i = (\Omega_i, \Psi_i) & \text{where} \\ \Omega_i = \{d_{i1}, d_{i2}, .., d_{in}\} \ , \ \Psi_i = \{\psi_{i1}, \psi_{i2}, .., \psi_{im}\} & \end{array} \quad (4)$$

$\Omega_i$ represents the finite set of visual blocks and $\Psi_i$ is the set of horizontal and vertical separators. A visual block $d_{ik} \in \Omega_i$ denotes a semantic part of the web page, does not overlap any other blocks and can be recursively considered as a web page.

Given a multi-entity web page representation, in order to estimate the $l_{ij}.probRel$ related to a link $(i, m)$, we need to compute the coherence of the semantic area where a link is located w.r.t. the origin page $d_i$ (Internal Coherence) and w.r.t. the destination page $d_m$ (External Coherence). The semantic areas containing links are identified by visual blocks

and detected by using spatial and visual cues described by the layout information embedded into the HTML language as reported in [1]. Before defining Internal and External Coherence we introduce the concept of link-block. Given a document $d_i = (\Theta_i, \Phi_i)$, a visual block $d_{ik} \in \Theta_i$ is called **link block**, and denoted by $d^*_{ik}$, if there exists a link $(i, m)$ from $d^*_{ik}$ to a destination page $d_m$.

Now, given a web page of origin $d_i = (\Omega_i, \Psi_i)$, let $\Omega^*_i \subseteq \Omega_i$ be the set of link blocks belonging to $d_i$, and let $W$ be the set of terms contained into $d^*_{ik}$, i.e. $W = \{t_j \in d^*_{ik}\}$. The *Internal Coherence* (IC) of a link block $d^*_{ik}$ w.r.t origin document $d_i$ is defined as:

$$IC(d^*_{ik}, d_i) = \begin{cases} \frac{1}{|W|} \sum_{j=1}^{|W|} \frac{\gamma_i(t_j) - \omega_{ik}(t_j)}{\omega_{ik}(t_j) \times \gamma_i(t_j)} & \text{if } |\Omega_i| > 1 \\ 1 & \text{otherwise} \end{cases} \quad (5)$$

where $\omega_{ik}(t_j)$ represents the number of occurrences of term $t_j$ into $d^*_{ik}$, $\gamma_i(t_j)$ is the number of occurrences of term $t_j$ in $d_i$.

The IC index takes value into the interval $[0, 1]$. The extreme values are taken respectively when no terms are shared between $d^*_{ik}$ and the rest of the document $d_i$ and when $d_i$ consists of a single visual block, i.e. $|\Omega_i| = |\Omega^*_i|$=1. In this last case we set $IC(d^*_{ik}, d_i) = 1$, stating that $d^*_{ik}$ is coherent to itself and consequently to $d_i$. In all other cases, the Internal Coherence, assumes values between 0 and 1. In particular, if a subset of terms belongs to a visual block $d^*_{ik} \in \Omega^*_i$ and these terms are well distributed along the entire document $d_i$, then the IC index tends to 1 stating that $d^*_{ik}$ is likely to be coherent to the document $d_i$. If these terms are not frequent in the rest of the document we may suppose that either $d^*_{ik}$ is not related for the document topic or we have a multi-topic document, determining IC values that tend to 0.

Given a web page of origin $d_i = (\Omega_i, \Psi_i)$ and a destination web page $d_m$, let $\Omega^*_i \subseteq \Omega_i$ be the set of link blocks belonging to $d_i$, and let $W$ be the set of terms contained into $d^*_{ik}$, i.e. $W = \{t_j \in d^*_{ik}\}$. The *External Coherence* (EC) of $d^*_{ik}$ w.r.t. the destination page $d_m$ is defined as:

$$EC(d^*_{ik}, d_m) = \frac{1}{|W|\rho_m} \sum_{j=1}^{|W|} \delta_{mk}(t_j) \quad (6)$$

where $\delta_{mk}(t_j)$ represents the number of occurrences of term $t_j$ in $d_m$ such that $t_j$ appears into the visual block $d^*_{ik}$, $\rho_m$ is the number of occurrences of the most frequent term in $d_m$ out of all the terms in $d_m$.

This metric estimates the External Coherence as the relative frequency of terms $t_j$ that occur into the destination page $d_m$ (and at the same time appears into the link block $d^*_{ik}$), w.r.t the number of occurrences of the most "important" term in $d_m$. If a subset of terms belongs to a visual blocks $d^*_{ik}$ and these terms are important into the destination document $d_m$, then $d^*_{ik}$ is likely to be coherent to $d_m$. The strenght $l_{im}.probRel$ of a link from page $d_i$ to page $d_m$ now can be defined as the arithmetic mean of Internal and External Coherence presented respectively in (5) and (6):

$$l_{im}.probRel = \frac{IC(d^*_{ik}, d_i) + EC(d^*_{ik}, d_m)}{2} \quad (7)$$

When the number of link blocks from an origin page $d_i$ to a destination page $d_m$ is greater than 1, an aggregation funtion can be used. In our investigation the mean aggregation function has been employed. Note that if there is a link from $d_i$ to $d_m$ and vice versa, two different values of link "strength" $l_{im}.probRel$ and $l_{mi}.probRel$ could be obtained.

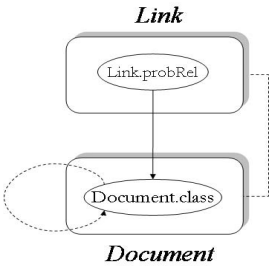


Figure 4. Dependency structure of PRMs/RU

### B. Dependency Structure and CPDs

The **dependency structure** $D_P$ of a PRM/RU is defined as follows:

- each attribute $E_i.A$ and $\widetilde{E}_j.U$ is mapped into a node
- edges between each attributes $E_i.A$ and its parent set $Pa(E_i.A)$ are drawn
- an edge from $\widetilde{E}_j.U$ to at least one attribute of its related entities is established

In the web example, the dependency model is defined as follows: (1) we introduce a node for the descriptive attribute *Document.class*, and a node for the artificial relationship variable *Link.probRel*; (2) we introduce at entity level an edge from *Document.class* to itself in order to model the dependency between the category label of an origin document and the category label of adjoining destination documents. Moreover, an edge from *Link.probRel* to *Document.class* is stated in order to take into account, during learning and inference process, probabilistic relationships between related documents. This dependency model is reported in Figure 4. Having a dependency structure that establishes causal relationships between attributes, conditional probability distributions can be estimaed for each node of the network.

Conditional probability distribution in PRMs/RU can be computed as for traditional PRMs (see subsection 3.2), because probabilistic relationships are modelled as attributes of (support) entities. The set of parameters $\theta_{D_P}$, represented by CPDs, can be estimated directly on relational data. Also in this case, since CPDs are defined at entity level, they can be shared by all objects that instantiate a given entity.

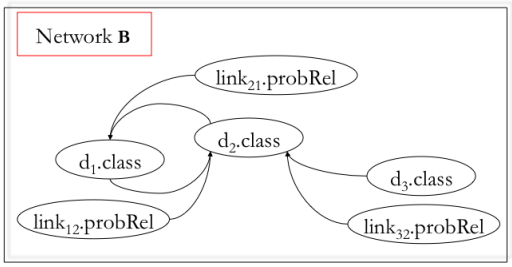


Figure 5. Unrolled Bayesian Network

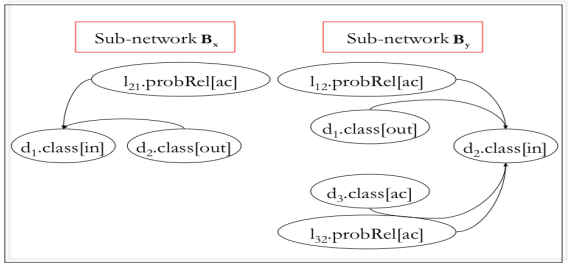


Figure 6. PRMs/RU Model Decomposition

### C. Joint Probability Distribution

A **PRM/RU** $\Phi^*$ for a relational schema $S$ is defined as a probabilistic graphical model characterized by:

- a set of parents $Pa^*(E_i.A)$ that, for each node $E_i.A$, includes all possible parents both from standard and support entities
- a set of parameters $\theta_{D_P} = P(E_i.A|Pa^*(E_i.A))$

PRMs/RU define, as well as traditional PRMs, a joint probability distribution over all completion of the relational schema. The joint probability distribution of PRMs/RU therefore reduces to that one presented in equation (2).

In order to explicitly use specific objects during the inference process, PRM/RU induces an Unrolled Bayesian Network as specified in section II-C. The joint probability distribution over all completions $I^*$ (that specifies both standard and support objects) can be therefore computed as:

$$P(I|\sigma_P, \Phi^*) = \prod_{e_{it} \in \sigma_p} \prod_{A \in A(e_{it})} P(e_{it}.A|Pa^*(e_{it}.A)) \quad (8)$$

In order to obtain a coherent probability model for the distribution presented in equation (8), the Unrolled Bayesian Network must be acyclic. However, this cannot be always ensured.

For example, in the web domain the acyclicity requirement is not guaranteed. If we state the web dependency model presented in Figure 4, in which the category of a web page depends on the category of its adjoining documents, the corresponding Unrolled Bayesian Network could generate cycles as depicted in Figure 5. Moreover, the probabilistic relationship, as pointed out in section III-A, can be asymetric. Indeed the link strength from $d_i$ to $d_j$ can have a different value that from $d_j$ to $d_i$. In order to deal with relational uncertainty and to eliminate graph cycles, a new unrolling procedure based on model decomposition is proposed.

*1) Model Decomposition:* The basic idea of dealing with probabilistic relationships and cyclic dependencies is to create two unrolled acyclic sub-networks stem from the cyclic one. This model decomposition, that starts from an unrolled Bayesian Network $B$, remove cycles through the following procedure:

- each attribute $e_{it}.A$ that is involved into a cycle is splitted into $e_{it}.A[in]$ and $e_{it}.A[out]$, which are placed into the set of nodes $N_{in}$ and $N_{out}$ respectively.
- each attribute $e_{it}.A$ (also comprising that ones belonging to support entities) which is not involved into a cycle is mapped into a node $e_{it}.A[ac]$ and placed into a set of nodes $N_{ac}$
- an edge from $e_{mp}.A[out]$ to $e_{it}.A[in]$ is established only if there exists an edge in $B$ from $e_{mp}.A$ to $e_{it}.A$
- an edge from $e_{mp}.A[ac]$ to $e_{it}.A[in]$ is established only if there exists an edge in $B$ from $e_{mp}.A$ to $e_{it}.A$
- an edge from $e_{mp}.A[ac]$ to $e_{it}.A[ac]$ is established only if there exists an edge in $B$ from $e_{mp}.A$ to $e_{it}.A$

At the end of this procedure we obtain two sub-networks $B_x$ and $B_y$, which are acyclic by construction.

In our web example, from network $B$ shown in Figure 5, we derived the sub-networks $B_x$ and $B_y$ depicted in Figure 6. Given a relational skeleton $\sigma_P$ the joint probability distributions for any relational skeleton $I^*$ can be computed over the sub-networks $B_x$ and $B_y$ as follows:

$$P(I^*|\sigma_P, B_x) = \prod_{e_{it}.A \in B_x} P(e_{it}.A|Pa(e_{it}.A)) \quad (9)$$

$$P(I^*|\sigma_P, B_y) = \prod_{e_{it}.A \in B_y} P(e_{it}.A|Pa(e_{it}.A)) \quad (10)$$

In order to derive a complete joint probability distribution for PRSs/RU, the joint probability distribution defined in equations (9) and (10) are combined as follows:

$$P(I^*|\sigma_P, B_x, B_y) = \frac{P(I^*|\sigma_P, B_x) + P(I^*|\sigma_P, B_y)}{2} \quad (11)$$

## IV. Experimental Investigation

PRM/RU has been evaluated comparing its accuracy to that one obtained by benchmarks models available in the literature: Bayesian Networks and PRM with Existence Uncertainty. Our experimental investigation starts from a dataset construction step, in which about 10000 web pages from popular sites listed in 5 categories of Yahoo! Directories (http://dir.yahoo.com/) are downloaded. See table I. The main goal is to infer the class label (topic) of a document $d_i$ by considering the class label of adjoining documents and

| Category | # of document |
|---|---|
| Art & Humanities | 3280 |
| Science | 2810 |
| Health | 1740 |
| Recreation & Sports | 1250 |
| Society & Culture | 960 |

Table I
DATASET FEATURES

the "strength" of their relationships. In particular, the dependency model depicted in Figure 4 and the corresponding acyclic $B_x$ and $B_y$ unrolled networks are constructed.

The Accuracy metric, which estimates the number of elements correctly classified, is used to evaluate the classification performance. Given the documents belonging to the collection reported in Table I, the accuracy (*Acc*) is estimated as follows:

$$Acc = \frac{\text{number of documents successfully classified}}{\text{total number of documents}} \quad (12)$$

Since **Bayesian Networks** are not suitable for relational data, a "flattening" procedure has been performed in order to reduce the document dataset into a propositional form. A propositional representation for relational data could be obtained by fixing the feature space and by mapping multiple feature values into a single one. For our experimental investigation, the feature space is defined by setting *originClass* and *destinationClass* features. Since a destination document $d_j$ could be linked by several origin documents $d_i$, the value of the *destinationClass* feature could depend on more than one *originClass* value. For this reason, we choose the most frequent *originClass* value with respect to each *destinationClass*. For example if a document is linked by two documents about *Health* and one document about *Science*, the *Health* value is chosen for representing the *originClass*.

The second benchmark predictive model used in this experimental investigation is PRM with Existence Uncertainty. This model, in which the existence or the absence of a relationship is modelled by the $exists$ binary variable, will be called **PRM with existence uncertainty (binary)**.

Since PRMs with Existence Uncertainty are also suitable for managing the strength of a relationship between two related objects, we decided to substitute the binary existence variable with an "aggregated version" of our *probRel*. In this case, if there exists a hyperlink from a document $d_i$ to a document $d_j$ and a hyperlink from a document $d_j$ to a document $d_i$, the truth observation about an existing relationship between $d_i$ and $d_j$ is substituted by a continuous variable $l_{\mathbb{A}(i,j)}.probRel$, whose value is computed as the average of $link_{ij}.probRel$ and $link_{ji}.probRel$. That is, the CPD is learned by considering the continuous values of the $probRel$ attribute instead of the binary values of the $exists$ attribute. This model, in which the existence of a relationship is modelled by the $link_{A(i,j)}.probRel$ continuous variable, will be called **PRM with existence uncertainty (continuous)**.

In Figure 7 the classification performance comparison on web relational data is reported: PRMs with relational uncertainty outperforms, in terms of accuracy, the benchmarks algorithms. We can see that PRMs have in general better performance than BNs. Moreover, we can note that

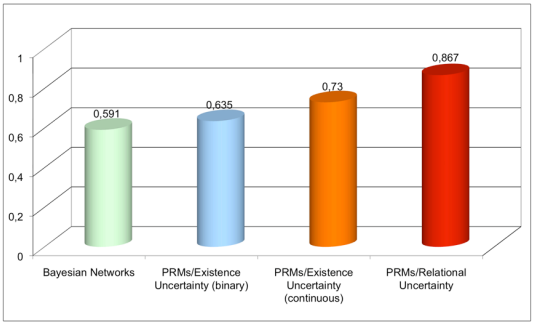


Figure 7. Performance comparison on web relational data

the potential of PRMs/RU and PRMs/EU of capturing the real "strength" of a relationship between two linked documents and the capacity of including this "strength" into the probability model can improve the models predictive power. Finally we can see that the continuous version of PRMs/EU, although able to consider the link strengh, seem to loose some information: the main limitation is the use of a single variable to model structural uncertainty.

## V. SUMMARY OF THE MAIN CONTRIBUTIONS

The main goal of this paper is to extend PRMs in order to include uncertainty over the relational structure and to apply them on a real case study. The main contribution of this paper can be summarized as follows:

1) Modelling uncertainty over the relational structure. PRMs/RU model the uncertainty that an existing relationship is semantically relevant. Therefore, PRMs/RU consider relationships **known to be existent**. This is the main difference with the uncertainty over the relational structure considered by PRMs/EU, which instead estimate the probability that a relationship does exist. More specifically, PRMs/EU model uncertainty by introducing a special existence attribute for each object that instantiate an entity, while PRMs/RU model uncertainty by introducing special probRel attribute only where is necessary, i.e. only if there exists a relationship between two objects.
2) Relational Uncertainty estimation. We provided a method for quantifying the relational uncertainty over related documents by considering web documents not only as an atomic unit, but as a composition of semantic areas. A relational uncertainty estimation that produces probRel=0.6 means that there exist a relation

between two documents and its strength is quoted 0.6. P(probRel), that is the probability distribution of the semantic relevance of the link between the to documents.

3) Modelling ciclicity. Even if a procedure for learning structure and parameters have been refined by Koller at al. [8] in order to ensures graph aciclicity in PRMs instantiation (unrolled bayesian networks), when the dependency structure is given by hand this condition could be not guaranteed. For this purpose, we provided a model decomposition procedure that, starting from a cyclic unrolled bayesian networks, creates two corresponding aciclyc sub-networks characterized by coherent probability distributions. Given these sub-networks a complete joint probability can be obtained by combining their posterior distributions.

## VI. Conclusion

In this paper one of the most promising model able to deal with uncertainty over relational data, known as Probabilistic Relational Model, has been investigated and extended. In particular, Probabilistic Relational Models with Relational Uncertainty have been proposed in order to deal with domains in which the relationships between objects could assume different degrees of "strength". Probabilistic Relational Models with Relational Uncertainty have been experimentally investigated for web document classification purposes. The proposed models have been compared, in terms of classification accuracy, with Bayesian Networks and Probabilistic Relational Models with Existence Uncertainty. The experimental investigation on real data shows that Probabilistic Relational Models with Relational Uncertainty can offer significant improvement with respect to the benchmark models used for prediction in relational domains. Even if the preliminary results obtained are primising, further investigations are required. Moreover, despite the high predictive power of Probabilistic Relational Models, inference procedures are still an open problem. Even if traditional approximate inference algorithms could be applied for posteriors computation, their complexity over large domains make the cost of those algorithms prohibitive.

## References


[1] Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. Extracting content structure for web pages based on visual representation. In *Proc. of the 5th Asia Pacific Web Conference*, pages 406–417, 2003.

[2] James Cussens. Loglinear models for first-order probabilistic reasoning. In *In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 126–133. Morgan Kaufmann, 1999.

[3] Pedro Domingos. Markov logic: a unifying language for knowledge and information management. In *CIKM*, page 519, 2008.

[4] Pedro Domingos and Matthew Richardson. Markov logic: A unifying framework for statistical relational learning. In *Proceedings of the ICML-2004 Workshop on Statistical Relational Learning and its Connections to Other Fields*, pages 49–54, 2004.

[5] Elisabetta Fersini, Enza Messina, and Francesco Archetti. Granular modeling of web documents: Impact on information retrieval systems. In *Proc. of the 10th ACM International Workshop on Web Information and Data Management*, 2008.

[6] Nir Friedman, Lise Getoor, Daphne Koller, and Avi Pfeffer. Learning probabilistic relational models. In *Proc. of the 16th International Joint Conference on Artificial Intelligence*, pages 1300–1309. Morgan Kaufmann Publishers Inc., 1999.

[7] Lisa Getoor, Nir Friedman, Daphne Koller, and Benjamin Taskar. Learning probabilistic models of link structure. *J. Mach. Learn. Res.*, 3:679–707, 2003.

[8] Lise Getoor, Nir Friedman, and Benjamin Taskar. Learning probabilistic models of relational structure. In *In Proc. ICML*, pages 170–177. Morgan Kaufmann, 2001.

[9] Lise Getoor, Daphne Koller, Benjamin Taskar, and Nir. Learning probabilistic relational models with structural uncertainty. In *Proc. of the ICML-2000 Workshop on Attribute-Value and Relational Learning:Crossing the Boundaries*, pages 13–20, 2000.

[10] Lise Getoor, Eran Segal, Ben Taskar, and Daphne Koller. Probabilistic models of text and link structure for hypertext classification. In *IJCAI01 Workshop on Text Learning: Beyond Supervision*, 2001.

[11] Manfred Jaeger. Relational bayesian networks. In *Uncertainty in Artificial Intelligence*, pages 266–273, 1997.

[12] Kristian Kersting and Luc De Raedt. Basic principles of learning bayesian logic programs. Technical report, Institute for Computer Science, University of Freiburg, 2002.

[13] Daphne Koller. Probabilistic relational models. In *Proc. of the 9th International Workshop on Inductive Logic Programming*, pages 3–13. Springer-Verlag, 1999.

[14] Andrew Mccallum. A comparison of event models for naive bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*, pages 41–48. AAAI Press, 1998.

[15] Jennifer Neville and David Jensen. Dependency networks for relational data. In *Proc. of the 4th IEEE International Conference on Data Mining*, pages 170–177, 2004.

[16] Jennifer Neville and David Jensen. Relational dependency networks. *Journal of Machine Learning Research*, 8:653–692, 2007.

[17] Jennifer Neville, David Jensen, Lisa Friedland, and Michael Hay. Learning relational probability trees. In *Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 625–630. ACM, 2003.

[18] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., 1988.