Structure Discovery using Statistical Relational Learning

Lise Getoor
Computer Science Department & UMIACS
University of Maryland
College Park, MD 20912
getoor@cs.umd.edu

Abstract

Statistical relational learning is a newly emerging area of machine learning that combines statistical modeling with relational representations. Here we argue that it provides a unified framework for the discovery of structural information that can be exploited by a data management system. The categories of structure that can be discovered include: instance-level dependencies and correlations, for example intra-table column dependencies and inter-table join dependencies; record linkages and duplicates; and schema matching and schema discovery from unstructured and semi-structured data.

1 Introduction

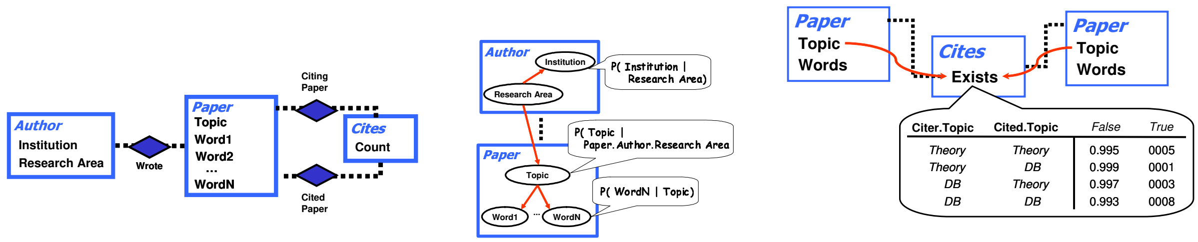
Combining statistical approaches to machine learning with relational learning methods is receiving increasing attention. There have been several recent workshops held, including two on "Learning Statistical Models from Relational Data" [25, 26], two on "Multi-relational Data Mining" [18, 19] and a summer school on relational data mining [20].

Statistical relational learning allows the combination of meta-information given by the domain schema and other background knowledge with principled approaches to model construction and inference based on statistical machine learning theory. A number of different models and formalisms have been proposed. One approach is probabilistic relational models (PRMs) [34]. Other approaches include Bayesian Logic Programs [33], Stochastic Logic Programs [42] and others. Here, we give a brief introduction to PRMs (for more details see [23]). First, we focus on their application to the discovery of instance-level structure. Next, we propose ways that they can be used to discover additional meta-level structure such as mappings and transformations. A key advantage is the ability to have a unified probabilistic framework for reasoning about data and meta-data together.

2 Probabilistic Relational Models

Probabilistic relational models (PRMs) are a recent development [34, 48, 53] that extend the standard attribute-based Bayesian network representation to incorporate a much richer relational structure. These models allow the specification of a probability model for *classes* of objects rather than for simple attributes; they also allow properties of an entity to depend probabilistically on properties of other *related* entities. The probabilistic class

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering



(a) (b) (c)

Figure 1: (a) The relational schema for a simple bibliographic domain. (b) The probabilistic dependencies for a simple bibliographic domain. (c) A probabilistic model for the existence of a citation.

model represents a generic dependence, which is then instantiated for specific circumstances, i.e., for particular sets of entities and relations between them.

As in the case of Bayesian networks, a key advantage of the models is that they provide a compact representation for a joint distribution over a set of random variables. The key that allows this compression is the discovery of independencies, and more importantly, conditional independencies, that hold in the joint distribution. Conditional independence is structure that can be discovered in the data; once discovered, it gives guidance in factoring the distribution. The collection of conditional independencies are constraints (or approximate constraints) that we can discover and exploit.

One component of a PRM is a relational schema for the domain, describing the tables, the attributes (columns) in the tables and the possible foreign-key joins between tables. Figure 1(a) shows a simplified relational schema for a bibliographic domain. The relational schema describes the categories of objects in the domain (here we have Papers, Authors, and Citations), the attributes of the objects (words, topic, count, institution and research area) and the foreign-key joins between the objects (written-by, citing paper and cited paper).

In addition to the relational component, the PRM describes the probabilistic relationships between attributes. Figure 1(b) shows a portion of the probabilistic dependency model. The direct probabilistic dependencies are indicated by arrows in the figure. Attributes are potentially correlated when there is a path between them. In addition, the probabilistic component also describes how the attributes interact by providing a local probability model or conditional probability distribution (CPD) for an attribute given its parents. These are indicated by the callouts in the figure. The local probability model describes the distribution of the child attribute given different values of the parents. In the case where the attributes are categorical, the local probability model is commonly a multinomial, or a tree-structured representation of the conditional distribution [8].

Importantly, PRMs can also model the probability of the existence of links between entities. In [24], we describe several approaches. One of these is to directly model the probability of the existence of a link. However, rather than build a model that uses the identity or key of the individual objects, which would be both unwieldy and would not support generalization, we base our probability model on attributes of the objects. Figure 1(c) shows a simple model for the existence of a citation between two papers. In this case the probability is based on the topics of the papers; papers that are on the same topic are more likely to cite each other.

A relational skeleton σ specifies the objects in our distribution. It defines the set of random variables of interest. A PRM describes the dependency structure $\mathcal{S}$ between the attributes, and the parameters of the CPDs in the model, $\theta_{\mathcal{S}}$. The semantics of a PRM specifies a probability distribution over the possible joint assignments $\mathcal{I}$ to the random variables of σ. As with Bayesian networks, the joint distribution over these assignments can be factored. That is, we take the product, over all $x.A$, of the probability in the CPD of the specific value assigned by the instance to the attribute given the values assigned to its parents. Formally, this is written as follows:

$$P(\mathcal{I} \mid \sigma, \mathcal{S}, \theta_{\mathcal{S}}) = \prod_{X_i} \prod_{A \in \mathcal{A}(X_i)} \prod_{x \in \sigma(X_i)} P(\mathcal{I}_{x.A} \mid \mathcal{I}_{\mathrm{Pa}(x.A)}) \tag{1}$$

where $\mathcal{I}_{x.A}$, and $\mathcal{I}_{\mathrm{Pa}(x.A)}$ are the values of a particular attribute and the values of its parent in $\mathcal{I}$. This expression is very similar to the chain rule for Bayesian networks. There are two primary differences. First, our random variables are the attributes of a set of objects. Second, the set of parents of a random variable can vary according to the relational context of the object — the set of objects to which it is related.

Learning Probabilistic Relational Models In order to learn a PRM from an existing database, we adapt and extend the machinery that has been developed over the years for learning Bayesian networks from data [30, 9, 35, 13] to the task of learning PRMs from structured data [21, 23]. In the learning problem, our input contains a relational schema that specifies the basic vocabulary in the domain — the set of tables, the attributes of each table and the possible foreign-key joins between tuples in the different tables. Our training data consists of a fully specified instance of that schema stored in a database.

There are two components of the learning task: parameter estimation and structure learning. In the parameter estimation task, we assume that the qualitative dependency structure of the PRM is known; i.e., the input consists of the schema and training database (as above), as well as a qualitative dependency structure $\mathcal{S}$. The learning task is only to fill in the parameters that define the CPDs of the attributes. In the structure learning task, we must discover the dependency structure $\mathcal{S}$ as well.

Parameter Estimation We are given the structure $\mathcal{S}$ that determines the set of parents for each attribute, and our task is to learn the parameters $\theta_{\mathcal{S}}$ that define the CPDs for this structure. The key ingredient in parameter estimation is the *likelihood function*, the probability of the data given the model. This function measures the extent to which the parameters provide a good explanation of the data. Intuitively, the higher the probability of the data given the model, the better the ability of the model to predict the data. The likelihood of a parameter set is defined to be the probability of the data given the model: $L(\theta_{\mathcal{S}} \mid \mathcal{I}, \sigma, \mathcal{S}) = P(\mathcal{I} \mid \sigma, \mathcal{S}, \theta_{\mathcal{S}})$. We can perform *maximum likelihood* parameter estimation or we can take a *Bayesian approach* to parameter estimation by incorporating parameter priors. For an appropriate form of the prior and by making standard assumptions, we can get a closed form solution for either of these estimates. The key computation required is the calculation of the sufficient statistics, or counts, for each CPDs in the PRM. The straightforward implementation for this requires a pass over the database; intelligent use of previously computed statistics can have significant impact on efficiency when more than one model must be evaluated.

Structure Learning Next we are faced with the more challenging problem of learning a dependency structure automatically. The main problem here is finding a good dependency structure among the huge number of many possible ones. As in most learning algorithms, there are three important components that need to be defined: the **hypothesis space** which specifies which structures are candidate hypotheses that our learning algorithm can return; a **scoring function** that evaluates the "goodness" of different candidate hypotheses relative to the data; and the **search algorithm**, a procedure that searches the hypothesis space for a structure with a high score. We use a greedy local search; see [21, 22] for more detail. This search is heuristic; the algorithm is not guaranteed to find a global optima. The computational complexity of the algorithm depends of the size of the hypothesis space and the score landscape. In practice, we have found for medium-size data sets (that fit in main memory), the running time for the learning algorithms are quite reasonable, ranging from a few minutes to a half hour.

Making use of a Learned Probabilistic Relational Model A learned PRM provides a statistical model that can uncover and discover many interesting probabilistic dependencies that hold in a domain. Unlike a set of (probabilistic) rules for classification, PRMs specify a joint distribution over a relational domain. Thus, like Bayesian networks, they can be used for answering queries about any aspect of the domain given any set of observations. Furthermore, rather than trying to predict one particular attribute, the PRM learning algorithm attempts to tease out the most significant direct dependencies in the data. The resulting model thus provides a high-level, qualitative picture of the structure of the domain, in addition to the quantitative information provided by the probability distribution. Thus, PRMs are ideally suited to exploratory analysis of a domain and relational data mining.

We have applied these techniques with great success in a variety of domains. One of the interesting domains has been a study of tuberculosis patients, along with their contacts and genetic information about the TB strains [28]. Another very different application of these methods is to the classic problem of selectivity estimation in databases. We have shown that these structured relational models can be constructed and used to give very accurate estimates for a wide class of queries [29].

3 Identity Uncertainty

The above describes how probabilistic models can be used to discover structure in the data; both structure in the distribution of attribute values, and structure in the existence of links, or joins, between elements. Another structure that probabilistic relational models can discover is instance-level dependencies, for example when two records refer to the same individual. In many practical problems, such as information extraction, duplicate elimination and citation matching, objects may not have unique identifiers. The challenge is to determine when two similar-looking items in fact refer to the same object. This problem has been studied in statistics under the umbrella of record linkage [61, 62]; in the database community for the task of duplicate elimination [57]; and in the artificial intelligence community as identity uncertainty [56, 50, 5].

In the statistical relational learning setting, it is important to take into account not just the similarity of objects based on their attributes, but also based on their links. In a bibliographic setting, this means taking into account the citations of a paper. Note that as matches are identified, new matches may become apparent. Pasula et al. [50, 39] have studied the use of probabilistic relational models for citation matching. They are able to explicitly reason about the individuals in the domain, and whether two citations actually refer to the same individual. This becomes a complex inference task, but they are able to solve it using Markov chain Monte Carlo techniques.

4 Schema Matching and Mapping

Interest in applying machine learning to schema matching has been growing [17, 16]. Statistical relational learning can also be used for schema mapping and discovery. By making full use of the data distribution, the meta information, we can perform more accurate mappings. Rahm et al. [55] give a taxonomy of schema matching approaches. Probabilistic approaches can provide the foundation on which to combine element and structure-level schema matching with element and instance-based data content matching. Because they consider instance-level semantics, probablistic approaches also show promise in developing not just syntactic matchings between schemas, but mappings with a well-founded data semantics such as those used in the Clio system [40].

5 Schema Discovery

Applications of statistical relational learning to semi-structured and unstructured data range from information extraction from unstructured text to complex wrapper generation [12]; however the majority of these approaches assume that some schema information is given.

Automatic schema extraction from lists and tables is ideally suited to the use of probabilistic models. Probabilistic models can reason over different segmentations of the data, and can find the appropriate number of elements. Lerman et al. [37] describe a heuristic approach that automatically learns a schema using information from list and detail pages on web sites. A probabilistic approach can also be taken. The probabilistic method has the advantage that it can more easily handle noise (missing fields) and it can combine the evidence from multiple sources (i.e. the list and detail pages) in a principled manner.

6 Statistical Modeling Challenges

Statistical machine learning provides the theory for parameter estimation, model selection, prediction and inference. Regardless of the particular model family, there are some unique challenges to applying statistical modeling techniques to structure discovery in multi-relational domains.

Logical versus Statistical Dependencies: The first challenge in structure discovery is coherently handling two different types of dependence structures:

- **relational or link structure** - the logical relationships between objects
- **probabilistic dependency** - the statistical relationship between attributes of objects.

In learning statistical models for relational data, we must not only search over probabilistic dependencies, as is standard in any type of statistical model selection problem, but potentially we must search over the different possible logical relationships between objects. This search over logical relationships has been a focus of research in inductive logic programming [41, 36] and some of the methods and machinery developed in this community may be applicable here. Methods for inferring functional dependencies and multi-valued dependencies [32, 2, 1, 3] are also important. The search over probabilistic dependencies has been studied, for example in structure learning for Bayesian networks [13, 30] and in the work described earlier on learning probabilistic relational models. There is an opportunity to integrate these more tightly.

Feature Construction: A second challenge is feature construction in the relational setting [14, 52, 51]. The attributes of an object provide a basic description of the object. Traditional classification algorithms are based on these types of object features. In a relational approach, it may also make sense to use attributes of linked objects. Further, if the links themselves have attributes, these may also be used. However, as others have noted, simply flattening the relational neighborhood around an object can be problematic. For example in hypertext domains, simply including words from neighboring pages degrades classification performance [10, 49]. A further issue is how to deal appropriately with relationships that are not one-to-one. In this case, it may be appropriate to compute *aggregate* features over the set of related objects. We have shown that this works well for learning probabilistic relational models [23]; others have examined it in the context of ILP approaches [51]. In [38], we show that some simple link statistics computed based on an object's links can improve predictive accuracy. The feature construction in relational domains can be used by a range of statistical models, including the models described above, first-order decision trees [7, 47], relational neural networks [6] and others.

Collective Classification: A third challenge is statistical inference using a learned statistical relational model. A learned model specifies a distribution over link and content attributes, which may be correlated based on the links between them. Intuitively, for linked objects, making a prediction for the category of one object can influence the inferred categories of its linked neighbors. This requires a more complex classification algorithm than for a traditional non-relational statistical model. The predictions that need to be made are all correlated. This is one of the fundamental reasons why one cannot simply convert a relational learning problem into a collection of independent propositional inference tasks.

This complicates things, but at the same time allows for more accurate inferences to be made. Iterative classification algorithms have been proposed for hypertext categorization [10, 49] and for relational learning

[46, 59, 58]. The general approach of iterative classification has been studied in numerous fields, including relaxation-labeling in computer vision [31], inference in Markov random fields [11] and loopy belief propagation in Bayesian networks [44]. Some approaches make assumptions about the influence of the neighbor's categories (such as that linked objects have similar categories); we believe it is important to *learn* how the link distribution affects the category. Other approaches to collective classification assume a very regular logical structure, for example either a chain, such as HMMs [54] or a dynamic Bayesian network [43], or a grid. Approaches that can handle more general relational structures are needed.

Instances versus Classes: A final modeling challenge is whether our statistical models refer explicitly to individuals, or only to classes or categories of objects [27, 15]. In many cases, we'd like to model that a connection to a particular object or individual is highly predictive; on the other hand, if we'd like to have our models generalize and be applicable to new, unseen objects, we also have to be able to model with and reason about generic collections of objects. The need to reason explicitly about individuals is a fundamental element of link-based object identification. But it also ties back to the feature construction problem; it may be useful to have features that refer to particular individuals. The majority of current approaches either only model classes of objects [21], or only reason about individuals [60, 4]. We have some initial results in the context of probabilistic relational models [22]. This is a general issue that is not specific to any particular model choice.

Computational Complexity and Scaling: The scale and dimensionality of statistical models that can be tackled has increased by several orders of magnitude in recent years. Models that were considered too complex for analysis just a decade ago are now routinely solved using large scale Markov chain Monte Carlo simulations [45].

While there have been great strides made in the size and complexity of models now amenable to computational analysis, the size of data bases and corpora has grown even faster. There is gap, and it is not clear that whether the gap is getting wider or narrowing. Regardless, it is our opinion that the use of coherent complete probabilistic frameworks hold greater promise for scaling than collections of independent ad-hoc procedures. Even in cases where the complete framework proves intractable, analysis of the complete framework will give guidance for the best places to make approximations, and will give tools to do analysis and comparison to see the effect of the approximations.

7 Conclusion

We have given a brief introduction to statistical relational learning, and have argued that it is ideally suited to data source structure discovery. Probabilistic models that support both data-level and meta-level representation and reasoning provide a unified framework for representing instance-level structure, object-level structure (for record consolidation) and schema- level structure (for schema mapping and schema discovery).

References

[1] S. Bell. Discovery and maintenance of functional dependencies by independencies. In *Proceedings of the Workshop on Knowledge Discovery in Databases*, pages 27–32. AAAI Press, 1995.

[2] S. Bell. Dependency mining in relational databases. In Dov M. Gabbay, Rudolf Kruse, Andreas Nonnengart, and Hans Jürgen Ohlbach, editors, *Proceedings of the First International Joint Conference on Qualitative and Quantitative Practical Reasoning*, volume 1244 of *LNAI*, pages 16–29, Berlin, June9–12 1997. Springer.

[3] S. Bell and P. Brockhausen. Discovery of constraints and data dependencies in relational databases. In Nada Lavrač and Stefan Wrobel, editors, *Proceedings of the 8th European Conference on Machine Learning*, volume 912 of *LNAI*, pages 267–270, Berlin, April 1995. Springer.

[4] A. Bernstein, S. Clearwater, and F. Provost. The relational vector-space model and industry classification. In *Proceedings of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data*, pages 8–18, 2003.

[5] M. Bilenko and R. Mooney. On evaluation and training-set construction for duplicate detection. In *Proceedings of Knowledge Discovery and Data Mining*, pages 39–48, 2003.

[6] H. Blockeel and M. Bruynooghe. Aggregation versus selection bias, and relational neural networks. In *Proceedings of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data*, pages 22–23, 2003.

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

[8] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. Context-specific independence in Bayesian networks. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 115–123. Morgan Kaufman, August 1996.

[9] W. Buntine. Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, 2:159–225, 1994.

[10] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 307–318, Seattle, Washington, 1998.

[11] R. Chellappa and A.K. Jain. *Markov random fields: theory and applications*. Academic Press, Boston, 1993.

[12] W. Cohen. Learning and discovering structure in web pages. *Data Engineering Bulletin*, 2003. in this volume.

[13] G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[14] C. Cumby and D. Roth. Feature extraction languages for propositionalized relational learning. In *Proceedings of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data*, pages 24–31, 2003.

[15] J. Cussens. Individuals, relations and structures in probabilistic models. In *Proceedings of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data*, pages 32–26, 2003.

[16] A. Doan, P. Domingos, and A. Halevy. Reconciling schemas of disparate data sources: A machine-learning approach. In *Proc. of ACM SIGMOD Conf. on Management of Data*, pages 509–520, 2001.

[17] A. Doan, P. Domingos, and A. Halevy. Learning to match the schemas of data sources: a multistrategy approach. *Machine Learning Journal*, 50(3):279–301, 2003.

[18] S. Dzeroski, L. De Raedt, and S. Wrobel. *Proc. KDD-2002 Workshop on Multi-Relational Data Mining*. ACM Press, 2002.

[19] S. Dzeroski, L. De Raedt, and S. Wrobel. *Proc. KDD-2003 Workshop on Multi-Relational Data Mining*. ACM Press, 2003.

[20] S. Dzeroski and B. Zenko. A report on the summer school on relational data mining. *SIGKDD Explorations*, 5:100–102, 2003.

[21] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1300–1307, Stockholm, Sweden, 1999. Morgan Kaufman.

[22] L. Getoor. *Learning Statistical Models of Relational Data*. PhD thesis, Stanford University, 2001.

[23] L. Getoor, N. Friedman, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In S. Dzeroski and N. Lavrac, editors, *Relational Data Mining*, pages 307–335. Kluwer, 2001.

[24] L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models of link structure. *Journal of Machine Learning Research*, 3:679–707, 2002.

[25] L. Getoor and D. Jensen. *Proc. AAAI-2000 Workshop on Learning Statistical Models from Relational Data*. AAAI Press, 2000.

[26] L. Getoor and D. Jensen. *Proc. IJCAI-2003 Workshop on Learning Statistical Models from Relational Data*. AAAI Press, 2003.

[27] L. Getoor, D. Koller, and N. Friedman. From instances to classes in probabilistic relational models. In *Proceedings of the ICML-2000 Workshop on Attribute-Value and Relational Learning: Crossing the Boundaries*, pages 25–34, 2000.

[28] L. Getoor, J. Rhee, D. Koller, and P. Small. Understanding tuberculosis epidemiology using probabilistic relational models. *Artificial Intelligence in Medicine Journal*, 2003. to appear.

[29] L. Getoor, B. Taskar, and D. Koller. Using probabilistic models for selectivity estimation. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 461–472. ACM Press, 2001.

[30] D. Heckerman. A tutorial on learning with Bayesian networks. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 301–354. MIT Press, Cambridge, MA, 1998.

[31] R. Hummel and S. Zucker. On the foundations of relaxation labeling processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):267–287, 1983.

[32] M. Kantola, H. Mannila, K.-J. Rih, and H. Siirtola. Discovering Functional and Inclusion Dependencies in Relational Databases. *International Journal of Intelligent Systems*, 7(7):591–607, September 1992.

[33] K. Kersting, L. de Raedt, and S. Kramer. Interpreting Bayesian logic programs. In *Proceedings of the AAAI-2000 Workshop on Learning Statistical Models from Relational Data*, pages 29–35. AAAI Press, 2000.

[34] D. Koller and A. Pfeffer. Probabilistic frame-based systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 580–587, Madison, WI, 1998. AAAI Press.

[35] W. Lam and F. Bacchus. Learning Bayesian belief networks: An approach based on the MDL principle. *Computational Intelligence*, 10:269–293, 1994.

[36] N. Lavrac and S. Dzeroski. *Inductive Logic Programming: Techniques and Applications*. Ellis Horwood, New York, NY, 1994.

[37] K. Lerman, C. Knoblock, and S. Minton. Automatic data extraction from lists and tables in web sources. In *IJCAI Workshop on Automatic Text Extraction and Mining*, 2001.

[38] Q. Lu and L. Getoor. Link-based classification. In *Proceedings of the Twentieth International Conference on Machine Learning*, Washington, DC, 2003.

[39] B. Marthi, B. Milch, and Stuart Russell. First-order probabilistic models for information extraction. In *Proceedings of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data*, 2003.

[40] R. J. Miller, L. M. Haas, and M. Hernández. Schema Mapping as Query Discovery. In *Proc. of the Int'l Conf. on Very Large Data Bases (VLDB)*, pages 77–88, Cairo, Egypt, September 2000.

[41] S. Muggleton, editor. *Inductive Logic Programming*. Academic Press, London, UK, 1992.

[42] S.H. Muggleton. Learning stochastic logic programs. In *Proceedings of the AAAI-2000 Workshop on Learning Statistical Models from Relational Data*, pages 36–41. AAAI Press, 2000.

[43] K. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, Computer Science Dept, UC Berkeley, 2002.

[44] K. Murphy and Y. Weiss. Loopy belief propagation for approximate inference: an empirical study. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 467–475. Morgan Kaufman, 1999.

[45] R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, University of Toronto, 1993.

[46] J. Neville and D. Jensen. Iterative classification in relational data. In *Proc. AAAI-2000 Workshop on Learning Statistical Models from Relational Data*, pages 13–20. AAAI Press, 2000.

[47] J. Neville, D. Jensen, L. Friedland, and M. Hay. Learning relational probability trees. Technical Report 02-55, University of Massachusetts Amherst, 2002.

[48] L. Ngo and P. Haddawy. Answering queries from context-sensitive probabilistic knowledge bases. *Theoretical Computer Science*, 171:147–177, 1996.

[49] H. Oh, S. Myaeng, and M. Lee. A practical hypertext categorization method using links and incrementally available class information. In *Proc. of SIGIR-00*, pages 264–271, 2000.

[50] H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. Identity uncertainty and citation matching. In *Advances in Neural Information Processing Systems 15 (NIPS 2002)*. MIT Press, 2003.

[51] C. Perlich and F. Provost. Aggregation-based feature invention and relational concept classes. In *Proceedings of Knowledge Discovery and Data Mining*, pages 167–176, 2003.

[52] S. Della Pietra, V. Della Pietra, and J. Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393, 1997.

[53] D. Poole. Probabilistic Horn abduction and Bayesian networks. *Artificial Intelligence*, 64:81–129, 1993.

[54] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[55] Erhard Rahm and Philip A. Bernstein. A survey of approaches to automatic schema matching. *VLDB Journal: Very Large Data Bases*, 10(4):334–350, 2001.

[56] S. Russell. Identity uncertainty. In *Proc. of IFSA-01*, Vancouver, 2001.

[57] S. Sarawagi and A. Bhamidipaty. Interactive deduplication using active learning. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)*, pages 269–278, 2002.

[58] B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *Proc. of UAI-02*, pages 485–492, Edmonton, Canada, 2002.

[59] B. Taskar, E. Segal, and D. Koller. Probabilistic classification and clustering in relational data. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 870–876, Seattle, Washington, 2001. Morgan Kaufman.

[60] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[61] W. E. Winkler. Advanced methods for record linkage. Technical report, Statistical Research Division, U.S. Census Bureau, 1994.

[62] W. E. Winkler. Methods for record linkage and bayesian networks. Technical report, Statistical Research Division, U.S. Census Bureau, 1994.