General Background

My research has investigated how machine learning can be applied to discover knowledge that supports artificial intelligence (AI) systems. A key aspect of an AI system is how information about its environment is represented; more informative representations increase both the performance and the complexity of the system [Russell and Norvig 2010]. AI employs three types of representations [Russell and Norvig 2010: Ch. 2.4.7]: 1) *Atomic*: an environment state is a black box with no internal structure. 2) *Factored*: a state consists of a vector of attribute values. This is the traditional representation of statistics and machine learning. 3) *Structured*: a state includes objects, each of which has relationships to other objects. Both objects and their relationships may have attributes. As shown in Figure 1, structured states may be visualized as abstract heterogeneous *information networks* [Sun and Han 2012] where objects are nodes and relationships are links. *First-order logic* is a prominent syntax for describing information network data. A factored representation can be viewed as the limiting case of an information network with 0 links [Nickel et al. 2016].

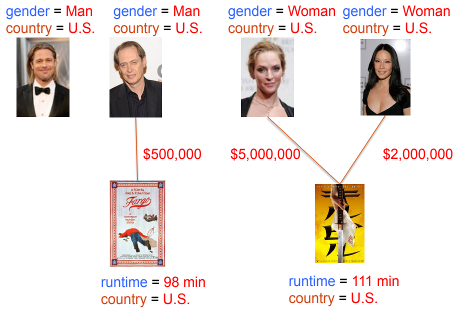


Figure 1. Small excerpt from the IMDb database represented as an information network. Links represent the AppearsIn relationship.

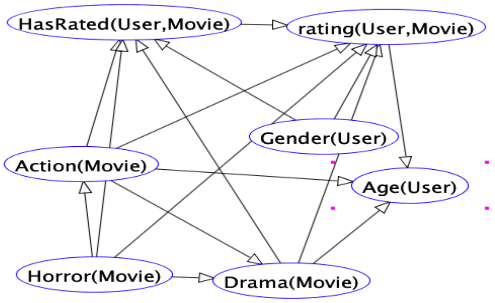


Figure 2. A first-order Bayes net for the IMDb database. The links show that movie ratings depend on both user gender and movie genre. The net specifies the probability of a node value given parent values (not shown). The learned probabilities show that men are more likely to rate action movies, but do not give higher ratings to them.

Recent Progress attributable to previous Discovery Grant. The proposed research will build on previous advances by my group in *learning first-order Bayes net models*, which marry graphical model concepts with first-order logic [Russell 2015; see Figure 2].

Algorithmic. Together with my graduate students **Khosravi** (now U of Queensland) and **Qian** (now BlackDuck Research Vancouver), I developed a *new model discovery algorithm* that has pushed the scalability boundary by orders of magnitude. Our learn-and-join algorithm constructs a model 100-1000 times faster and scales to millions of data points, orders of magnitude improvements over the previous state-of-the-art. (More details below).

Theoretical. For learning a Bayes net structure, the most common approach is search+score: define and maximize a *model score* that measures how well the structure models the data. The fundamental issue for a model score is how to balance data fit (measured by log-likelihood) with model complexity (measured as a function of the number of model parameters). Together with my MSc student **Gholami**, we showed how to strike this balance for information network data [Schulte and Gholami 2017]. Our empirical evaluation showed that getting the right balance is crucial for learning good models. Our main theorem guarantees that the first-order Bayes nets selected by our scoring method converge to an optimal model of event frequencies in the information network, as the network size increases. This is the first large-sample performance guarantee for first-order model selection. It cannot be achieved using a traditional single model score, but only with a *gain function* that compares two candidate models.

Exception Mining. A major topic in network analysis is finding nodes that are exceptional or anomalous [Akoglu et al. 2015]. The network neighbourhood of a node has been called its *egonet.* The similarity of two nodes can be measured by the similarity of their egonets. My PhD student Riahi and I developed a novel *exceptionality metric* that compares a node's egonet to the entire network [**Riahi** and Schulte 2015; best student paper award]. The metric can be efficiently computed using the statistical patterns represented in a first-order Bayes net. We applied the metric to movie and soccer data and found that actors and players rated as exceptional by our metric were highly likely to be successful in their domains.

Objectives

The overarching goal for my research program is to advance machine learning for information network data. Short-term (5 year) objectives are to advance first-order Bayes net learning in several directions:

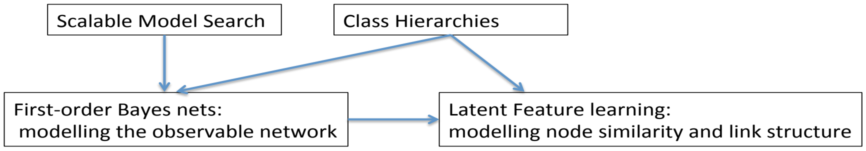
1. Develop a Bayes net model discovery for *large numbers* of objects, links, and attributes.
2. *Leverage class hierarchies* for more scalable and statistically valid model construction.
3. *Combine Bayes net models without latent features with neural net models for latent features.* Latent features represent types, clusters, or embeddings.

A long-term objective of my research program is to develop machine learning for *applications where relational information must be extracted from data*, such as text and images.

Methodology (Evaluation)

The key performance metrics for our developed algorithms will be 1) *learning time* and 2) *predictive accuracy*. For predictive accuracy, my students and I will use standard metrics on benchmark datasets for test set prediction (e.g. conditional log-likelihood and Area-Under-Curve as in [Schulte 2016]), and for knowledge graph completion (e.g., recall@k [Garcia-Duran and Niepert 2017]). For learning time, success means that the new algorithms should learn the same model on previous datasets, but orders of magnitude faster. Large complex datasets like Freebase are beyond the reach of any current graphical structure learning method. Success would allow us to construct a model for Freebase in hours.

Below, the three research objectives are discussed individually, with each section divided into Literature Review, my Recent Progress, Methodology, and Anticipated Significance.


Scalable Model Search
Class Hierarchies
First-order Bayes nets: modelling the observable network
Latent Feature learning: modelling node similarity and link structure

Objective 1. (MSc 1, PhD 1).

The network scalability parameters are (1) the number of nodes, (2) the number of facts, meaning attribute-value combinations for nodes and links, and (3) the number of node and link attributes [Nickel et al. 2016]. *My objective is to scale our current Bayes net structure learning system to 100M nodes, 500M facts, and 30K attributes.* For comparison with large public structured datasets, this is more than sufficient for DBPedia, Yago2, and Wikidata (the largest). Wikidata contains 18M nodes, 66M facts, 1,632 attributes. The objective is almost sufficient for Freebase (40M nodes, 637M facts, 35K attributes), but much less than the Google Knowledge Graph (570M nodes, 1.8B facts, 35K attributes).

Literature Review. There has been a considerable amount of research on learning the structure of a first-order model graphical model from network data [Kimmig et al. 2014], but scalability has remained a challenge. For example, the state-of-the-art boosting method for learning Markov Logic Networks takes days to build a model for an IMDb dataset with about 15M facts [Schulte et al. 2016]. For this reason, complex applications like information extraction have not used graphical model learning (e.g., Zhu et al. used a small number of hand-crafted rules [2015]). The dominant time cost is *computing sufficient statistics* [Lv et al. 2012]: For a given set of attributes, how often does each combination of attribute values occur in the data? Efficient algorithms exist for counting in information networks (eg, [Venugopal et al. 2015]); but they apply only to attribute values that involve the presence, not absence of links.

Recent progress. We have built the most scalable first-order structure learning system, the open-source FactorBase [FactorBase 2017]. The key innovations are the following: 1) A new *model search strategy* [Schulte and Khosravi 2012a] described in our contributions section; 2) A new scalable exact *counting algorithm* that sufficient statistics, that involve *both* present and absent links, occur in the network. The algorithm provides the capability to learn correlations among different link types, which are useful for a number of data mining and learning tasks [**Qian** et al. 2014]; and 3) Leveraging existing *systems*: We discovered that the Structured Query Language (SQL) can be used as a high-level scripting language for building machine learning programs for structured data, that are portable, modular, compact and whose execution is highly optimized by a database system (DBMS) [**Qian** and Schulte 2015]. FactorBase stores sufficient statistics in a special kind of DBMS table called *contingency table*.

Methodology. Our current FactorBase system *scales well in the number of facts*, by exploiting natural groups of attributes. For instance, it performs a single joint analysis of all attributes of a node type (e.g., actor). The scalability bottleneck is *the number of attributes*, because we pre-compute the occurrence counts of all possible different attribute value combinations, which can number in the millions.
To solve this problem, our innovation will be to replace pre-counting by *on-demand counting* [Lv et al. 2012]. At any given point, a graphical model search considers only a small local set of attributes, with a manageable number of sufficient statistics. However, the attribute set changes dynamically during model search. Instead of pre-computing an exponential number of counts once, on-demand computes a constant number of counts many times. On-demand computing requires re-use of computed counts through efficient caching. We will research count caches for information networks. A promising direction is combining cache structures (e.g., ADtrees [Moore and Lee 1998]) with our contingency tables for storing counts. We will also leverage powerful data analysis systems such as Spark and Hadoop that support large memory resources from clusters. The programming support for these systems includes SQL so porting our scripts should be feasible if not straightforward.

Significance. Scaling model construction to large numbers of facts and attributes is a key step for fully realizing the potential of relational learning in high-impact applications.

Objective 2. (PhD 1) In many structured domains, objects fall into a given hierarchy of classes. My objective is to extend *first-order Bayes net structure learning to leverage domain knowledge that specifies a class hierarchy*. An example of a class hierarchy in ice hockey would be that a right-wing player is a forward, who is an NHL player, who is a hockey player.

Literature Review. Class hierarchies are a major part of domain ontologies and have been much studied in statistics, machine learning, and data mining. A prominent example is the SNOMED CT medical ontology with over 300,000 class definitions [Wang et al. 2002]. Structured data in the Semantic Web come with a class hierarchy [Rettinger et al. 2009]. In statistics, a class hierarchy is represented in the structure of a hierarchical model (aka multi-level model). Random effects and shrinkage leverage the class hierarchy for parameter estimation [Krushke 2014]. The basic idea is that parameter values for related classes are smoothed towards a common value. For example, parameter values for left-wingers and right-wingers would tend to be closer to each other than to defensemens'. Shrinkage can be achieved through a common prior on parameter values for related classes.

Recent Progress. My student Gholami and I introduced the gain function concept for extending Bayes net scores defined for factored data, to network data [Schulte and **Gholami** 2017; see above].

Methodology. *Statistical:* The goal is to learn a multi-net [Geiger 1996], a collection of related Bayes nets, one for each class in the hierarchy. *We propose a new shrinkage approach to hierarchical model selection based on priors over first-order graphical models*. For factored data, many structure scores use a Bayesian prior, that penalizes models with many edges between attributes. A promising approach is to start with a structure prior for factored data, and use our gain function method to extend it to information networks. In this way, shrinkage encourages related classes to share the same Bayes net edges.
Algorithmic: We will investigate how the algorithms for Objective 1 can be extended for class hierarchies. In preliminary research, we introduced a new search heuristic: propagate learned Bayes net edges from more general to more specific classes [**Riahi** and Schulte 2013].
Evaluation: Class hierarchies are available for the large structured datasets from the Semantic Web community. We can compare learning with and without class hierarchies.

Significance. I expect hierarchy information to improve model discovery in several dimensions. *Accuracy:* Shrinkage is known to lead to more accurate learning and better predictions. *Scalability:* Hierarchical structure supports localization and re-use of computations. *Interpretability:* Incorporating a domain ontology makes the learned rules more meaningful to users.

Objective 3 (PhD 2, MSc 2, 3). Learning latent (unobserved) features is a highly effective approach to analyzing information networks. My objective is to *combine first-order Bayes net structure learning with latent feature discovery via deep learning.* Leveraging the methods described for Objectives 1 and 2, which analyze observable attributes only, will make latent feature learning more scalable and reliable.

Literature Review. A factorization model assumes that there exist latent features for each node, such that the observable attributes of nodes are conditionally independent given the latent features. *This implies that the existence of a link between two nodes can be explained by the similarity of their latent features.* Factorized models achieve accurate predictions for tasks such as link prediction and recommending items to users. For instance, matrix factorization won the $1M Netflix challenge for movie recommendation [Koren 2009]. A growing body of research has investigated how to scale factorization to large information networks [Nickel et al. 2016]. This includes embedding methods from deep learning to map nodes to feature vectors [Dong et al. 2017]. Factorization requires setting parameter values whose number is proportional to the number of nodes, often in the millions. *The large parameter scale raises challenges for computational feasibility and statistical identifiability* (i.e., the possible solutions are highly underdetermined no matter how large the dataset). A recent direction is *combining the strengths of models for observable attributes and for latent features* [Nickel et al. 2016]. One possibility is a log-linear model to combine evidence in favour of an attribute value [Garcia-Duran and Niepert 2017]. For example, if knowledge of a person's workplace allows the model to infer her residence, latent features are unnecesary for predicting her residence and can be used instead to model other parts of the network. Class hierarchies have been applied to constrain latent features [Rettinger et al. 2009].

Recent Progress. We introduced an accurate novel *random regression method for deriving attribute value predictions* from a first-order Bayes net with no latent features [Schulte, **Qian**, et al. 2016]. **Riahi** and I developed a novel *probability-based similarity metric* that uses the complex features learned by a Bayes net to evaluate the similarity of two nodes by comparing their egonets [**Riahi** and Schulte 2015].

Methodology. We propose three innovations for latent feature learning: 1) *Leverage Bayes net structure learning* to construct a powerful model from data. 2) *Combine our random regression method with latent feature prediction* in a log-linear formalism. 3) Use our Bayes net based-similarity measure to *regularize neural network embeddings so that similar nodes tend to have similar embeddings*.

Significance. The cutting-edge approach to information network analysis combines the speed and statistical reliability of learning from observable attributes only, with the expressive power of latent features. Current implementations have not used structure learning methods because they do not scale to the benchmark datasets. The research proposed will close this gap and develop the full power of network learning with all feature types, latent and observed.

Objective 4 (MSc 4,5,6). A long-term goal is *applying relational model discovery to support extracting relationships from text and image data.* Relationship extraction for text is an active research area. For example, Stanford's Deep Dive [Zhang 2015] has been applied to several large collections of scientific articles, for example in medical genetics and pharmacogenomics. Google researchers have extracted a very large information network from the worldwide web. Detecting relationships in images is a recent new computer vision task, supported by the Visual Genome dataset [Lu et al. 2016]. Probabilistic models are an important component of such systems: they can be used to grow knowledge, through *knowledge graph completion*, which uses the model to infer unobserved facts from observed facts.

Impact: Overall, the proposed research has significance for several areas:
1) *Data science*. Studies show that industry data scientists spend 60% of their time *preparing data* for analysis, but only 9% of their time on the actual analysis. Most enterprises store their data in a *relational database*. Because traditional machine learning tools work only with a factored representation, they require the analyst to "extract, transform, load" [ETL 2017] relational data, which loses information and contributes greatly to the data preparation overhead. The philosophical principle of my work has been that instead of requiring the user to transform relational data, machine learning research should develop methods that work directly with it.

2) *Automated Knowledge Base Construction*. A first-order Bayes net represents a *probabilistic knowledge base*: first-order rules with probabilistic parameters that quantify uncertainty [Domingos and Richardson 2007]. Probabilistic knowledge bases have received extensive study in AI [Bacchus 1990]. Relational learning, including the methods of my research program, provides the new capability of constructing probabilistic knowledge bases from data. These learned probabilistic knowledge bases have shown their usefulness in several application domains, such as link-based classification, link-based clustering, query optimization, entity resolution, information extraction, representing uncertainty in databases etc. [Domingos and Richardson 2007; Niu et al. 2011; Wang et al. 2008].

3) *Information Extraction*. A major goal of AI is to extract structured information from unstructured sources such as images and text. Gary Marcus identifies machine reading as "a key to building any truly intelligent system…" [Marcus 2017]. Deep Dive's commercial version Lattice Data was recently acquired by Apple for $200M. An important component of information extraction systems are probabilistic models of objects and relationships. The proposed research develops new methods for constructing probabilistic relational models from data that support information extraction.

Conclusion: Machine learning for AI is a key technology for the future, which has seen strong investment from Canadian governments and businesses. The most advanced AI systems use the most informative data. A major international effort is under way to develop machine learning that harnesses the power of information networks. This fundamental technology will enable next-generation AI applications, such as extracting structured information from massive text and visual data. The proposed research will contribute an important component technology, graphical model structure learning. My research program investigates how to combine the power of symbolic logic with statistical learning; arguably, this is the most challenging, and the most promising, direction to achieve artificial intelligence.