

# Toward Interactive Relational Learning

Ryan Rossi and Rong Zhou

Palo Alto Research Center

## Abstract

This paper introduces the Interactive Relational Machine Learning (*i*RML) paradigm in which users interactively design relational models by specifying the various components, constraints, and relational data representation, as well as perform evaluation, analyze errors, and make adjustments and refinements in a closed-loop. *i*RML requires fast real-time learning and inference methods capable of interactive rates. Methods are investigated that enable direct manipulation of the various components of the RML method. Visual representation and interaction techniques are also developed for exploring the space of relational models and the trade-offs of the various components and design choices.

## Discussion

Relational Machine Learning (RML) (Getoor and Taskar 2007) methods exploit the relational dependencies between nodes to improve predictive performance (Macskassy and Provost 2003). However, these approaches often fail in practice due to low relational autocorrelation, noisy links, sparsely labeled graphs, and data representation (Rossi et al. 2012).

To overcome these problems, we envision an Interactive Relational Machine Learning (*i*RML) paradigm in which users interactively specify relational models and data representation (via transformation techniques for the graph structure and features), as well as perform evaluation, analyze errors, and make adjustments and refinements in a closed-loop. In this work, humans interact with relational learning algorithms by providing input (in the form of labels, similarity/kernel function, hyper-parameters, priors, confidence/uncertainty about particular instances, learning rate, corrections, rankings, probabilities, evaluation) while observing the output (in the form of predictions, uncertainty, feedback, and any useful visual representation of the data).

*i*RML combines fast, flexible, and powerful relational learning techniques with interactive visualization to aid in evaluation, tuning, analysis of errors, model selection, regularization, semi-supervised information, etc. It is designed for rapid interactive visual exploration and learning through visual representations and interaction techniques at each stage. *i*RML enables users to trade off competing goals, encode

prior knowledge, understand the model, and analyze errors and uncertainty.

Existing RML methods use exhaustive grid-based search, which is computationally expensive, suffers from local optimum problems, and is impractical in many situations (Rossi and Neville 2010). However, since *i*RML leverages human capabilities (Figure 1), the space of relational models can be searched more effectively and seamlessly, while also being guided by prior/domain knowledge and experience (which may be difficult to encode into the model directly).

Visual analytic techniques for rich multivariate networks are introduced that combine state-of-the-art visualization and interaction techniques (Ahmed and Rossi 2015) to understand, explore, and interactively analyze RML methods. These techniques facilitate the design and selection of a relational model (from the space of potential models), their evaluation, error and uncertainty analysis, as well as give users the ability to interactively refine them in a closed-loop. In addition, users can quickly understand the interplay between properties of the network and its associated node and edge data, e.g., understanding the relationship between the formation of large stars (or cliques) and how these correlate with attributes of nodes.

Desiderata for the *i*RML paradigm are as follows:

- **Immediate visual feedback.** *i*RML methods should be optimized for the way humans learn (Ahlberg, Williamson, and Shneiderman 1992; Thomas and Cook 2005; Ahmed and Rossi 2015). Thus, they should provide *immediate* and *continuous visual feedback* upon every interaction (e.g., change of a slider for filtering uncertain or misclassified nodes, selection of a subgraph for modeling, or correcting the class label of a node). Further, interactive queries need to be rapid, incremental and reversible with immediate visual feedback.
- **Flexibility and generality.** Methods should be useful for a wide variety of data, constraints, and learning scenarios. They need to also be robust for learning sparsely labeled graphs, noisy relational data, low and varying levels of relational autocorrelation, and other problems that frequently arise in practice.
- **Effectiveness.** Models must have good predictive quality with low error, variance, and bias.
- **Scalable methods.** Fast time- and space-efficient learning methods capable of interactive rates is an important and key requirement. The requirement of rapid and interactive model updates often dictates trading off accuracy with speed. Thus, network sampling methods may be used to balance speed and accuracy.

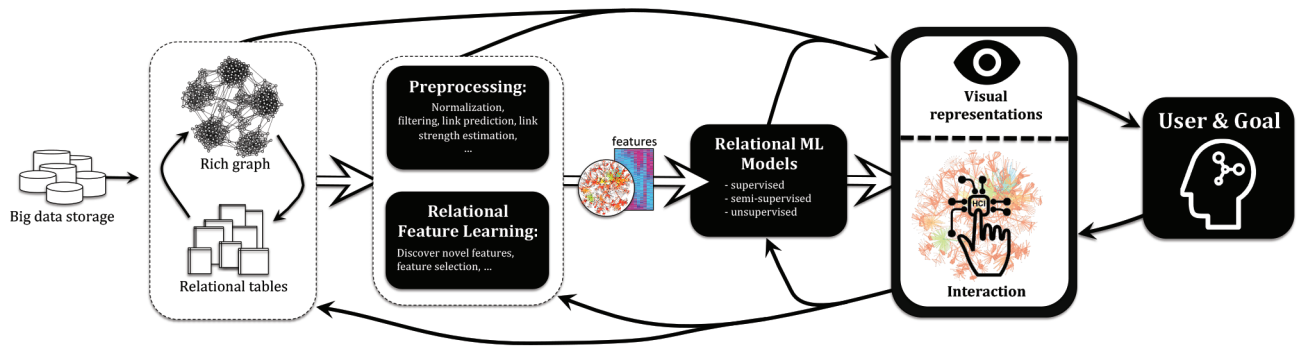


Figure 1: Interactive Relational Machine Learning (*iRML*) Paradigm. Visual encoding of the results from the various components are indicated via the top arrows, whereas user interactions are represented by the arrows at the bottom.

- **Accessibility and simplicity.** To be accessible to domain but non-ML experts, *iRML* methods must be carefully designed to be simple, intuitive, and easy-to-use. Whenever possible, assistance and guidance from the system is desired.
- **Principled models.** Another challenge is the design of intuitive learning and inference methods to facilitate interactive reasoning, understanding, and derivation of theoretical behavior and guarantees. This enables quick understanding and refinement by the user, while also providing a means to backtrack if warranted to understand a specific outcome or anomaly.
- **Unified & expressive models.** A unifying family of relational learning methods that express a large and multi-faceted space of relational models. These models must perform well across a variety of different data, characteristics, and assumptions. They must also generalize to a variety of learning settings (e.g., relational active learning, online/incremental learning, ...).

those nodes of interest. Our *iRML* system also supports interactive visual graph filtering (e.g., remove all uncertain nodes above a user-specified threshold.). In addition, all visualizations are interactive and support brushing, linking, zooming, panning, tooltips, etc. Efficient update rules are also derived to avoid relearning the model (after each user interaction/visual query). For example, after the deletion (or insertion) of a node, we can update the global relational model via a fast localized update. These local updates enable real-time exploration capabilities by leveraging fast exact or approximate solutions.

Many of the components in our *iRML* system may be explored using interactive visualization and analytic techniques, including the attribute to predict, initial features to use (non-relational and graph-based features), local model for estimation, kernel function (RBF, linear, polynomial, etc.), hyper-parameters (for selected kernel), node- and feature-wise normalization scheme (L1, min-max, etc.), as well as whether to use semi-supervised learning (SSL), and meta-features (based on current estimates). Interactive link prediction methods and many other important learning components are also included in our *iRML* system.

## References

- Ahlberg, C.; Williamson, C.; and Shneiderman, B. 1992. Dynamic queries for information exploration: An implementation and evaluation. In *Proc. of SIGCHI*, 619–626.
- Ahmed, N. K., and Rossi, R. A. 2015. Interactive visual graph analytics on the web. In *ICWSM*, 566–569.
- Getoor, L., and Taskar, B., eds. 2007. *Introduction to Statistical Relational Learning*. MIT Press.
- Macskassy, S., and Provost, F. 2003. A simple relational classifier. In *SIGKDD MRDM*, 64–76.
- Rossi, R., and Neville, J. 2010. Modeling the evolution of discussion topics and communication to improve relational classification. In *SIGKDD SOMA*, 89–97.
- Rossi, R. A.; McDowell, L. K.; Aha, D. W.; and Neville, J. 2012. Transforming graph data for statistical relational learning. *JAIR* 45:363–441.
- Thomas, J. J., and Cook, K. A. 2005. *Illuminating the Path: the research and development agenda for visual analytics*. IEEE.

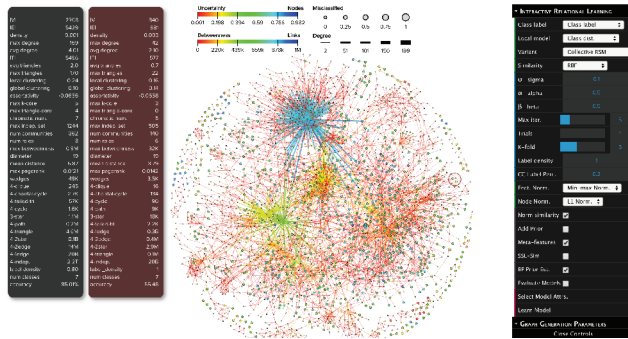


Figure 2: *iRML* overview. This screenshot is from *cora* – a common RML benchmark data set (Macskassy *et al.*).

An overview of our *iRML* system is shown in Figure 2. In that example, we first interactively learn a model, then select the misclassified nodes for further analysis. The global statistics of the selected subgraph are shown in the right-most panel. Node color represents the model’s uncertainty using an entropy-based measure, whereas the size of the node indicates whether it was correctly classified or not. In Figure 2, misclassified nodes are given a larger size so that they can easily be identified for further exploration. Uncertainty (and learned class prob. distribution, statistics, etc.) of a node or set of nodes may also be displayed by selecting or mousing over