Research Statement

Joseph J. Pfeiffer III

Overview

We can cast many domains in machine learning into a network, or relational, framework. Examples include domains such social networks, the web, internet topologies, bioinformatics, and fraud detection, where the entities are interconnect through relationships such as friendships or messages, hyperlinks, packet transfer, interactions and phone calls or emails. Common machine learning tasks in relational networks include:

- Identifying securities brokers engaged in fraud using their email or communications network.
- Predicting whether users of large scale social networks are interested in purchasing a product or viewing a story.
- Recommending friendships between large scale social network users with similar interests and social circles.

Machine learning in relational domains centers around modeling dependencies between item traits and relational structure. My research interests lie in the development of scalable algorithms to handle uncertainty within relational machine learning, including scalable models of network structure. The key insight to my work is that, in large scale and partially observed network domains, missing labels and edges can significantly impact the performance of standard relational machine learning approaches, introducing bias and error. This is a significant difficulty in real world applications, as the majority of network domains have missing relationships or biased observations. **As the majority of relational learning problems occur within a single large, partially observed network (e.g, Facebook and Twitter), I work on methods to model these domains, including correcting uncertainty from missing data, identifying biases in learning and inference in order to improve prediction accuracy, and scaling to large real world networks.** More specifically, I focus on semi-supervised relational machine learning methods that further exploit the (unlabeled) data. Generally speaking, my work can be characterized into three connected areas:

- Defining scalable network models of the conditional distributions of relationships.
- Incorporating structural uncertainty into learning and inference to improve prediction accuracy in partially observed networks.
- Overcoming learning biases during semi-supervised relational machine learning.

My work generally leverages estimates of the unobserved data, such as modeling the impact of missing relationships or incorporating predictions of unknown labels into learning. This approach requires background knowledge from a variety of areas, and I focus on expanding knowledge both in machine learning and network science. Further, my work focuses on methods that are theoretically scalable (subquadratic) while emphasizing demonstrations on large scale relational networks. I will discuss my research in each of these areas, then conclude with future work.

Scalable Generative Network Models

Recent interest in networks has focused on (a) understanding the processes that formed a network and (b) defining a distribution of graphs with similar structural characteristics to real world networks. I have focused my work on understanding the theoretical underpinnings of their corresponding scalable sampling algorithms, extending the resulting framework to allow for modeling additional characteristics such as transitivity and attributes. **My resulting framework is the only one that provides algorithms for scalable sampling of networks with correlated attributes, transitivity, and modeling of higher order graph statistics such as the joint degree distribution.** We can use these samples for a variety of applications; for example, a null empirical distribution for hypothesis tests.

More specifically, I began by developing a novel scalable generative graph model, the *Transitive Chung-Lu* (TCL) generative network model [4]. This model augments the traditional Chung Lu graph model to incorporate transitive closures, a common feature of real world networks. More precisely, this model formulates edge probabilities as a mixture model, combining both the random probability two vertices will close an edge (proportional to the product of their respective degrees) coupled with the probability that a random walk beginning at one edge point would land at the other within two hops. I gave a scalable learning algorithm, allowing us to learn the mixture parameter from real data with varying levels of transitivity. Further, the sampling algorithm is also scalable, generating networks with millions of vertices and edges in subquadratic time.

My work on TCL was one of the first to note that traditional generative network sampling algorithms did not sample from the defined distribution (as stated in preceding works). Rather, it found that the scalable *edge-by-edge* sampling algorithms for the original Chung Lu algorithms (and others) used the Binomial approximation during the sampling procedure (an approximation that TCL also exploits). After initially analyzing existing scalable samplers, I unified and generalized them to develop the *Attributed Graph Model* (AGM) [5]. AGM is a scalable model that allows for scalable sampling of networks with correlated vertex attributes, allowing us to sample large scale networks from the conditional distributions given the attributes on the corresponding vertices. These graph models (TCL and AGM) maintained important characteristics of the original distributions; for example, they provably maintain the expected degree distribution despite their incorporation of additional features (transitivity and attribute dependence). More recent work with collaborators has focused on modeling higher order graph structures [3] or developing exact samplers [1, 2].

Relational Learning and Inference in Partially Observed and Biased Networks

Considerable amounts of relational learning research centers on assuming the network is completely observable; that is, all relationships between the individuals are known. However, this assumption is rarely true in practice: individuals communicate across different mediums (e.g., in person) that aren't observed within the network. Thus, relational machine learning should incorporate the uncertainty of the observed network in order to improve the prediction performance. **My work has focused on scalable learning and inference in partially observed domains, employing probabilistic edges and semi-supervised learning to greatly improve predictive performance for real-world settings**. More precisely, this work targets common real world domains characterized by incomplete or biased data observations.

Recent work of mine has centered around iterative identification of fraudulent individuals in networks [8]. This application proceeds by investigating items in a relational network (e.g, fraudulent brokers in a communication network). At each iteration, a small number of items (e.g., brokers) are investigated: this investigation reveals new relationships to other items (e.g., other brokers) that can be used for learning and prediction in subsequent iterations. As each iteration targets individuals that are involved with fraud, the sample drawn from the population distribution can lie far from the true distribution. For this *Active Exploration* task, our aim is to learn where to query the network in order to maximize the identification of positive nodes. This application is useful for improving relational learning and inference in a variety of directions due to the extreme sampling biases and partial information it must overcome.

Initial work began by understanding the effects of the missing *relational information*, discovering that by operating on a squared network representation I could develop a model that significantly outperformed simple baselines [8]. Next, I centered my research on the effects of the sampling process: I worked to understand the effect of the bias that resulted from the targeted sampling approach [9]. My initial solution to this bias focused on reweighting the training data based on the observed data such that the generalized loss was minimized.

My more recent solutions to this problem account and improve on each of the initial findings. In particular, in [6] I improve on the squared network representation by reducing the runtime complexity of squared inference on a network to *linear* time (in the number of edges). This algorithm is exact, not an approximation, allowing for implementation on considerably larger datasets (by orders of magnitude). Further, it can extend outside the Active Exploration domain, finding potential applications to other domains with large amounts of missing relational information. In [6], I also improve on eliminating the biases from [9] by incorporating *semi-supervised learning* into the learning and inference steps. In doing so, rather than simply reweighting the information I utilize estimates of the unlabeled information to learn from the unlabeled data. By incorporating lessons learned in prior work, I was able to greatly outperform all competing methods [6].

Overcoming Relational Learning Biases during Semi Supervised Learning

My work on the Active Exploration task led me to incorporate semi supervised learning approaches into relational learning. However, in doing so I learned that previous semi supervised learning approaches remained somewhat limited. Relational semi supervised learning approaches, such as *Relational Expectation Maximization* (Relational EM), must utilize approximations during the learning stage of the EM algorithm in order to scale to large network domains. These approximations result in *over-propagation* errors in sparse networks, with the majority of unlabeled items converging to a single label prediction, rather than a variety of values that match the training data. For this work, **I focused on identifying learning biases and correcting these biases during inference in large scale network domains, as well as developing better learning approximations that incorporate more information.** These approaches allow us to perform semi-supervised learning in large relational domains, with empirical demonstrations on networks with millions of edges, far past other relational learning methods.

My initial work in this domain presented stochastic variants to the deterministic Relational EM algorithm [7]. As part of this paper, we discussed how the typical relational EM learning approximation resulted in considerable over propagation error when applied to predict the remaining network instances, frequently predicting all the labels to be a single value. When used as part of the iterative EM algorithm, this resulted in wild variations between the predictions from one round to the next. As a solution to this problem, I proposed the *Relational Data Augmentation* (Relational DA) approach for semi supervised learning in relational domains [7]. Relational DA is similar to Relational EM; however, rather than performing fixed point iterative updates, it integrates over a distribution of possible parameter values. Relational DA avoids the extreme variances created by the Relational EM algorithm, providing considerably more accurate estimates over unlabeled data when a network is sparsely labeled.

Relational EM and Relational DA utilize the known instances and their (possibly uncertain) attributes for learning. A more general approach, that is more in line with typical semi-supervised methods, would utilize the probabilities of all unlabeled instances as (weighted) training examples. However, the bias from the learning approximation kept this more general approach from being utilized as it results in all label probabilities collapse to a single value. To correct for this error, I recently proposed *Maximum Entropy Inference* for collective classification [10]. This method constrains the predicted labels such that the distribution of predictions matches the distribution of the training examples. Further, I proved that it can work with any probabilistic classifier and has a provable constant overhead and that it can be implemented as part of an asynchronous massively parallel collective inference algorithm, which can scale to networks with millions of edges. This is orders of magnitude larger than previous relational machine learning algorithms, which generally operate on networks with tens of thousands of edges. Our approach outperformed both independent learning algorithms and simple relational inference algorithms.

Future Work

Due to the size and complexities of working with large scale relational networks, many previous machine learning and analysis methods are too computationally intensive for practical use. I view my general approach to overcome these limitations as:

- Limiting the search over the solution space to likely possibilities
- Correcting the errors the can result of the search approximation

For example, scalable sampling algorithms typically target highly probable edges, then downsample the edges that fail to meet some required criteria [8, 5, 3]. Similarly, for semi-supervised relational learning we utilizes approximations for the learning step, which we subsequently correct during the inference step [7, 10]. Although these methods require approximations and corrections, they remain scalable while providing better predictive performance than other methods. Further, these methods have theoretical justifications for either the approximation or correction, such as proving that they preserve some model characteristics or place logical constraints on the output space. Hence, by understanding the constraints of the domain and any corresponding assumptions, we can fully eliminate the resulting uncertainty and biases to make more accurate predictions.

In future work, I intend to continue analyzing the impact of biases and uncertainty, due to observational limitations and approximations during learning and inference, towards improving relational machine learning in large scale domains. For example, *active learning* involves iteratively acquiring labels believed to improve subsequent predictions. However, relational active learning is unique as the subsequent labelings impact both the classifier and the inference process. In particular, any classifier learned from the sample is exposed to biased labelings, as the labels are selected to reduce uncertainty and are not a random sample. My development of large scale, relational semi-supervised methods are ideal for application in this domain. First, these semi-supervised approach allows for a largely accurate initial classifier that overcomes the biased observations [6, 10], allowing us to focus on identifying instances in the network that reduce inference error. Second, their scalability allows them to be applied to large networks.

In addition, I will continue to explore avenues where patterns in network structure can help to inform relational classification of node and link attributes. For example, for several network domains we've found that the degree of a vertex is heavily correlated with particular labels (e.g., Amazon sales rank and copurchases), meaning standard assumptions about whether the items are exchangeable are false. Thus, relational learners have a bias when they fail to model this dependency. By continuing to evolve our understanding of the generative network modeling, we can incorporate them into our learning algorithms. As there can be many possible structural characteristics to search over, even considering all possible triangles in networks is too expensive for this task. Thus, applying our scalable learning and representations for structural graph models is a natural step, with possibilities available due to the ability to model edge-attribute dependencies [5], transitivity [8] or higher order graph statistics such as the joint degree distribution [3].

Lastly, to date the majority of relational classification models are conditional models, in that they condition on a single distribution given a sample of network structure. It is still an open question as to extent a model learned from one network can be *transferred* to another network, where the distributions are not necessarily the same. My work on generative models of network structure will help to investigate not only what types of network structures are drawn from the same distribution [8, 5, 3], but also to explore mechanisms to adapt learned models to networks drawn from similar but different domains. By utilizing partial information from across multiple domains, we can improve our predictions across all relational domains. As a result, my established approaches for modeling network information coupled with scalable semi-supervised learning techniques can greatly improve prediction accuracy across a range of practical, real-world domains.

References

[1] S. Moreno, J. J. Pfeiffer III, and J. Neville. A scalable and exact sampling method from probabilistic generative graph models. In *Under Submission: 24th International World Wide Web Conference (WWW 2015)*, 2015.

[2] S. Moreno, J. J. Pfeiffer III, J. Neville, and S. Kirshner. A scalable method for exact sampling from kronecker models. In *Proceedings of the 14th IEEE International Conference on Data Mining (ICDM 2014)*, 2014.

[3] S. Mussmann, J. Moore, J. J. Pfeiffer III, and J. Neville. Incorporating assortativity and degree dependence into scalable network models. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI 2015)*, 2015.

[4] J. J. Pfeiffer III, T. La Fond, S. Moreno, and J. Neville. Fast generation of large scale social networks while incorporating transitive closures. In *Fourth ASE/IEEE International Conference on Social Computing (SocialCom 2012)*, 2012.

[5] J. J. Pfeiffer III, S. Moreno, T. La Fond, J. Neville, and B. Gallagher. Attributed graph models: Modeling network structure with correlated attributes. In *Proceedings of the 23rd International World Wide Web Conference (WWW 2014)*, 2014.

[6] J. J. Pfeiffer III, J. Neville, and P. Bennett. Active exploration in networks: Using probabilistic relationships for learning and inference. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM 2014)*, 2014.

[7] J. J. Pfeiffer III, J. Neville, and P. Bennett. Composite likelihood data augmentation for within-network statistical relational learning. In *Proceedings of the 14th IEEE International Conference on Data Mining (ICDM 2014)*, 2014.

[8] J. J. Pfeiffer III, J. Neville, and P. N. Bennett. Active sampling of networks. In *10th Workshop on Mining and Learning with Graphs (MLG)*, 2012.

[9] J. J. Pfeiffer III, J. Neville, and P. N. Bennett. Combining active sampling with parameter estimation and prediction in single networks. In *Proceedings of the ICML Structured Learning Workshop*, 2013.

[10] J. J. Pfeiffer III, J. Neville, and P. N. Bennett. Overcoming relational learning biases to accurately predict preferences in large scale networks. In *Under Submission: 24th International World Wide Web Conference (WWW 2015)*, 2015.