

The Relational Push-Pull Model: A Generative Model for Relational Data Clustering *

Adam Anthony

Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
1000 Hilltop Circle, Baltimore, MD 21250
(410) 455-8894
aanthon2@cs.umbc.edu

Relational data clustering is a form of relational learning that clusters data using the relational structure of data sets to guide the clustering. Many approaches for relational clustering have been proposed recently. The common assumption in much of this research is that *relations have a binding tendency* (Neville, Adler, & Jensen 2003; Bhattacharya & Getoor 2007; Taskar, Segal, & Koller 2001; Neville & Jensen 2006). That is, edges are assumed to appear more frequently within clusters than between clusters.

This binding quality may be too strong an assumption. Bhattacharya & Getoor (2007) acknowledge that it is possible for a relation to provide “negative evidence,” where the presence of a relation between two objects implies that the objects belong in different clusters. If most of the edges in a relational set provide negative evidence, then this set should be considered to have a *separating*, rather than a binding, tendency. Consider a social network of university students containing both men and women. If the social network is partitioned by gender, edges for a dating relation will appear most frequently between the clusters. This can be visualized as an approximately bipartite structure between the two clusters that “pushes” them apart. Edges for a roommate relation, on the other hand, will appear most frequently within the clusters. This binding relation can be visualized as a net that “pulls” the objects in a cluster closer together, hence the name “Relational Push-Pull Model.”

The prevailing approach to clustering when the relational tendency is not known *a priori* is *stochastic block modeling*. Nowicki & Snijders (2001) presented a method for automated learning of stochastic block models in which only object-object relations are observed. Recently, researchers (Handcock, Raftery, & Tantrum 2007; Tallberg 2005; Anthony & desJardins 2007) have acknowledged that Nowicki & Snijders’ approach ignores a valuable source of information: the attributes associated with the data objects in the relational data set. I propose an extension to Nowicki & Snijders’s model in which relations exist probabilistically as a function of the connected objects’ features.

My research has taken a generative approach to the clustering problem: I assume that each object and each relation

exists as a sample from some underlying mixture of probability distributions. Specifically my model assumes that individual object features are drawn from a cluster-specific distribution. Once the features are drawn, relations between objects are drawn such that they have a co-dependence on both the feature values of the connected objects and the cluster membership of the two objects. Similar approaches include the work of Tallberg (2005), who assumed that object features are drawn from a cluster-specific distribution, and Handcock, Raftery, & Tantrum (2007), who assumed that both object features and cluster membership can have an impact on edge existence. My work differs from theirs in that I am considering the impact of both assumptions in my model; also, the underlying distributions I use are different. An advantage of my approach is that it is applicable to a variety of real-world relational data sets, including the Internet Movie Data Base¹ and a biological interaction (food-web) data set. My current research efforts are applied to the Dimensionality of Nations project (Rummel 1999) and a friendship interaction data set.

The following tasks were proposed for completion of the dissertation, with completed tasks in **bold font**.

- **Survey existing approaches to learning with relational data, with a particular focus on clustering relational data.**
- **Develop and justify a probabilistic model for relational data.**
- **Implement a simulated annealing approach to estimating parameter values and finding a clustering.**
- Apply theoretical work to mining web data, particularly the blogosphere.
- Apply theoretical work to multi-agent team formation.
- Develop an original discrete relational clustering algorithm that approximates my model, or implement a more thorough inference technique for learning model parameters.

To date, I have expanded my coverage of relational learning by reading various papers, as well as attending and participating in various conferences and workshops such as the 2006 ICML workshop on Statistical Relational Learning, the

*This material is based upon work supported by the National Science Foundation under Grant No. #0545726.
Copyright © 2008, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹<http://www.imdb.com>

2007 ICDM workshop on Optimization-Based Approaches to Data Mining, and the up coming AAAI Spring Symposium on Social Information Processing. Additionally, I focused my coursework on machine learning, probability theory, and multi-agent systems so that I would have the background knowledge necessary to complete future research tasks.

In Spring 2007, I defended a Master's thesis in which I originally proposed the Relational Push-Pull Model and presented a learning approach using simulated annealing and maximum likelihood estimators for parameter values to find a clustering. Results from that work were presented at the ICDM workshop on Optimization-Based Approaches to Data Mining (Anthony & desJardins 2007), mentioned above. In Fall 2007, I spent some time improving the relational edge model to be more adaptive to the input data by incorporating logistic regression. I also re-implemented the software to make it more memory and run-time efficient to enable the application of the model to large data sets. I will submit a paper describing these improvements, as well as further experimental justification for the model, to the 2008 International Conference on Machine Learning.

In terms of tasks yet to be completed, I have begun a collaboration with a fellow graduate student, Akshay Java, to collect and analyze a web crawl of the political blogosphere from the month of January 2008. Our intent is to use my model to discover underlying communities of politicians that might represent sub-factions within the larger party affiliation groups. I also believe that my model could be used in a backwards fashion to affect the formation of teams in a mobile networked multi-agent system setting where the ability to maintain a network connection between two agents is measured as a joint probability distribution of the agents' capabilities and their locations in a two- or three-dimensional space.

A final task, which is the most ambitious portion of my research, is to use the properties of my model to develop a discrete clustering approach that would likely be faster and more space-efficient than any inference technique that can be used to estimate a probabilistic model. Such an algorithm would be a significant contribution to the field, since there are few discrete clustering techniques that incorporate object features and object relations simultaneously and would be applicable to many different types of data sets. A discrete algorithm would also be a more effective data mining tool than an inference-based approach, since it would likely involve fewer experimental parameters that vary between data sets (e.g. the cooling schedule in Simulated Annealing or the prior probability hyperparameters in Gibbs sampling). In the event of a major road block to this accomplishment, I intend to instead implement a more powerful inference technique for learning the model, such as expectation maximization or Gibbs sampling, which would likely out-perform my maximum likelihood approach to learning.

References

- Anthony, A., and desJardins, M. 2007. Data clustering with a relational push-pull model. In *Proceedings of the ICDM Workshop on Optimization-based Data Mining Techniques with Applications*.
- Bhattacharya, I., and Getoor, L. 2007. Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data* 1(1):1–36.
- Handcock, M. S.; Rafferty, A. E.; and Tantrum, J. M. 2007. Model-based clustering for social networks. *Journal of the Royal Statistical Society* 170(2):1 – 22.
- Neville, J., and Jensen, D. 2006. Leveraging relational autocorrelation with latent group models. In *Proceedings of the Fifth IEEE International Conference on Data Mining*.
- Neville, J.; Adler, M.; and Jensen, D. 2003. Clustering relational data using attribute and link information. In *Proceedings of the Text Mining and Link Analysis Workshop*. 18th International Joint Conference on Artificial Intelligence.
- Nowicki, K., and Snijders, T. A. B. 2001. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association* 96(455):1077 – 1087.
- Rummel, R. J. 1999. Dimensionality of Nations Project: Attributes of Nations and Behavior of Nation Dyads, 1950-1965. Inter-university Consortium for Political and Social Research.
- Tallberg, C. 2005. A Bayesian approach to modeling stochastic blockstructures with covariates. *Journal of Mathematical Sociology* 29(1):1 – 23.
- Taskar, B.; Segal, E.; and Koller, D. 2001. Probabilistic classification and clustering in relational data. In Nebel, B., ed., *Proceeding of IJCAI-01, 17th International Joint Conference on Artificial Intelligence*, 870–878.