

# Statistical Modeling of Graph and Network Data

Padhraic Smyth

Information and Computer Science

University of California, Irvine

CA 92697-3425

smyth@ics.uci.edu

## 1 Introduction

This brief paper reviews a number of ideas from the statistical and social networks literature that are potentially of interest and relevance to computer scientists working in relational learning. Pointers to references are provided for further reading.

## 2 Statistical Models for Social Network Data

The statistical literature on social networks typically assumes that we are modeling a set of  $n$  entities or “actors” and their binary relationships. The data are often represented in the form of an  $n \times n$  adjacency matrix  $Y$  where entry  $y_{ij} = 1$  (or 0) indicates the presence (or absence) of some form of directed relationship between entities  $i$  and  $j$ , e.g., “ $i$  considers  $j$  to be a friend.” Undirected graphs, with  $y_{ij} = y_{ji}$ , are obviously also of interest. More generally  $y_{ij}$  can measure the “value of the relation” from entity  $i$  to entity  $j$  on some suitably defined scale. In addition, each entity can have a set of covariates, denoted  $\mathbf{x}_i$ , e.g., a vector of demographic measurements, with  $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  being the full set of observed covariate data.

There is a long tradition of developing statistical models for such data in the social networks literature (a comprehensive survey of early work in the field is provided by Wasserman and Faust, 1994). Central to these modeling approaches is the treatment of the edge data measurements,  $y_{ij}$ , as observations from an underlying distribution for a set of binary random variables defined on each of the ordered  $n(n-1)$  pairs  $i$  and  $j$ . The rationale behind this approach is that the observed  $n(n-1)$  relations  $Y$  are noisy indicators as to whether a link truly exists or not. The goal of statistical modeling in this context is to infer a parsimonious model for  $P(Y|X)$  that requires a relatively small number of parameters to explain the pattern of observed relations (and non-relations), as a function of both local network properties (such as the indegree and outdegree of individual nodes) as well as the covariates  $X$ .

Various forms of Markov random fields (MRFs) (Frank and Strauss, 1986) and exponential graph models (sometimes referred to as  $p^*$  models in the social networks literature; Wasserman and Pattison, 1996) have been used to model  $P(Y|X)$ . Much of this work builds on the earlier classic work of Besag in spatial statistics (1974).

These generative modeling frameworks provide all the usual advantages of statistical inference, such as:

- a language for modeling of specific network characteristics, such as reciprocity and transitivity of relations (Wasserman and Faust, 1994);
- modeling techniques for incorporating covariates  $X$ , e.g., via suitably-defined logistic regression models;
- inference methods for handling systematic errors in the measurement of links (Butts, 2003);
- hierarchical Bayes and random effect frameworks that allow individual-level variation to be modeled (Hoff, 2003);
- general classes of methods for parameter estimation and model comparison, such as Markov chain Monte Carlo methods (Snijders, 2002);
- incorporation of clusters of nodes in the graph whose statistical properties are equivalent, such as the block models of Wang and Wong (1987);
- methodologies for incorporating specific prior information such as desired functional forms on degree distributions (Snijders, 2003); and
- interpretability of the resulting model (although the validity of such interpretation for certain types of models is debatable: see comments below).

## 3 The Limitations of Current Statistical Network Models

Unfortunately these models are far from a panacea for all statistical modeling involving network and graph data. Computational issues are a major concern. Parameter estimation in general for Markov random fields is a well-known problem due the intractability of computing the normalization constant in such distributions (which requires, in this context, a sum over all possible graphs with  $n$  nodes).

Perhaps even more troubling is the fact that there are essential and fundamental identifiability problems in the estimation of parameters in many of these models—these estimation problems have only become apparent in relatively recent times. To quote Hoff, Raftery, and Handcock (2002):

“...commonly used models are more global than local in structure and this contributes to model degeneracy and instability problems .... These issues are not resolved by alternative forms of estimation but represent defects in the models themselves....”

Elsewhere, Besag (2002) comments that:

“A particularly blatant use of MRFs occurs in the analysis of social networks, where the parameters in Markov random graphs are often ascribed substantive interpretations that are meaningless....”

## 4 Latent Variable Network Models

In this general context, recent work in statistical modeling of network data has begun exploring newer functional forms for  $P(Y|X)$  that promise to bypass some of the less desirable features of MRFs. Of particular note, and of potential direct interest to machine learning researchers, is the use of latent (hidden) variable models.

Hoff, Raftery, and Handcock (2002) propose an interesting probabilistic model of this form where the probability of an edge between entities  $i$  and  $j$  is a function how far apart they are in a  $k$ -dimensional latent space. The location vector  $\mathbf{z}_i$  for each entity  $i$  is estimated from the data, along with parameters that modulate how covariate information (if any) affects the likelihood of an edge between any pair  $i$  and  $j$ . This leads to models of the form

$$\begin{aligned} \text{LogOdds } P(y_{ij} = 1 | \mathbf{z}_i, \mathbf{z}_j, \mathbf{x}_i, \mathbf{x}_j, \theta) \\ = \alpha + f(\mathbf{x}_i, \mathbf{x}_j; \beta) + d(\mathbf{z}_i, \mathbf{z}_j) \end{aligned}$$

where  $f(\mathbf{x}_i, \mathbf{x}_j; \beta)$  is a scalar-valued function dependent on a set of parameters  $\beta$  governing pairwise covariate effects (such as the similarity of the characteristics of individual entities) and  $d(\mathbf{z}_i, \mathbf{z}_j)$  is a distance in the latent  $k$ -dimensional space. A variety of different distance measures (such as Euclidean and absolute) can be used. Hoff and colleagues illustrate how standard maximum likelihood and Bayesian techniques can be applied to this model for parameter estimation. They then apply variations of this model to relatively small social network data sets (the largest data set had 27 entities (nodes)) and achieve relatively interpretable and robust results.

A major limitation of this type of model, however, is the relative lack of scalability. The likelihood is by definition a product over all pairs of nodes, whether an edge was observed or not, leading to an inherently  $O(n^2)$  algorithm. While this may be practical for relatively small social networks, the algorithm is not directly scalable to many of the large networks (e.g., with  $n = 10^5$ ) that are often of interest to computer science researchers.

This model is also reminiscent of multi-dimensional scaling (MDS), a well-known technique for “projecting” pairwise similarity data into a multi-dimensional vector space. However, this latent-variable graph model is more powerful than MDS in that the full spectrum of techniques for probabilistic modeling (such as incorporation of covariates) can be brought to bear.

## 5 Conclusion

It is not yet clear how the classes of statistical models mentioned above are related to other types of models and learning algorithms that have been proposed in the relational learning literature—probabilistic relational models are clearly of particular relevance. In principle, the intersection of statistical modeling and machine learning techniques appears to be a useful area for further exploration. Leveraging the strengths of each approach should produce new classes of models and applications for rich relational domains.

## Acknowledgements

The research in this paper was supported by the National Science Foundation under Grant NSF-IIS-0083489.

## References

- Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, ser B., 36, 192–225.
- Butts, C. (2003) Network inference, error, and informant (in)accuracy: a Bayesian approach, *Social Networks*, in press.
- Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002) Latent space approaches to social network analysis, *Journal of the American Statistical Association*, 97, 1090–1098.
- Hoff, P. D. (2003) Random effects models for network data, *Proceedings of the National Academy of Sciences*, in press.
- Frank, O. and Strauss, D. (1986) Markov graphs, *Journal of the American Statistical Association*, 81, 832–842.
- Snijders, T. A. B (2003) Accounting for degree distributions in empirical analysis of network dynamics, *Proceedings of the National Academy of Sciences*, in press.
- Snijders, T. A. B (2002) Markov Chain Monte Carlo estimation of exponential random graph models, *Journal of Social Structure*, Vol. 3, No. 2.
- Wang, Y. J., and Wong, G. Y. (1987) Stochastic block models for directed graphs, *Journal of the American Statistical Association*, 82, 8–19.
- Wasserman, S. and Faust, K. (1994) *Social Network Analysis: Methods and Applications*, Cambridge: Cambridge University Press.
- Wasserman, S., and Pattison, P. (1996) Logit models and logistic regression for social networks: I. An introduction to Markov graphs and  $p^*$ . *Psychometrika*, 61, 401–425.