Hierarchical Latent Space Models for Social Network Analysis

Alex Loewi[1,4], Francisco Ralston[2], Octavio Mesner[2,3,4]

[1]Public Policy, [2]Engineering & Public Policy, [3]Statistics & Data Science, [3]Machine Learning

Carnegie Mellon University

Introduction

Problem: Multigraphs are a complex, but naturally occurring, and highly valuable, form for data. While several recent papers attest to an agreement that models for multigraphs are valuable, these models all have several important problems when it comes to using them in one key area – Social Network Analysis (SNA). Social science and network analysis in particular rely heavily on intuitive and interpretable models, which becomes an increasingly large problem as the data they analyze becomes increasingly complex. Here, we present a model that solves several of the major problems of interpretability, while retaining rigorous statistical underpinnings.

Data

The motivating data for this problem comes from a recently conducted study at a charter school in West Baltimore. The study collected, among other sources, a five-layer valued network between the approximately eighty employees of the school. These layers were comprised of the answers questions in the following areas:

- Frequency of interaction
- Discussing Academics
- Discussing Behavior
- Social interaction
- Professionally Helpful Relationship

While empirically these layers are found to have very high correlations, modeling them in a way that demonstrates those correlations is challenging.

Goals

An ideal model would solve the following problems:

1. Comparable layouts
2. Layer-wise dimensionality reduction
3. Manually tunable between comparable and expressive

Model

The model is based on the widely used Latent Space Model (LSM) for social networks. This model has the following form:

$$\prod_k \prod_{ij} \sigma_{ijk}^{y_{ijk}} (1 - \sigma_{ijk})^{1 - y_{ijk}}$$

$$\sigma_{ijk} = \text{logit}^{-1}(\alpha_k - d(z_{ik}, z_{jk}))$$

The original model has two simple elements: one, a binomial likelihood for the edges, making for easy and stable inference. And two, an edge probability based on the distance of the two node's latent variables, z_i, z_j. The farther apart these variables are, the less likely an edge is. The original model uses the observed edges to infer possible latent positions for the nodes that satisfy these relationships.

Methods

To fit these models, three methods were tested on the following likelihood:

$$\prod_k \prod_{ij} \sigma_{ijk}^{y_{ijk}} (1 - \sigma_{ijk})^{1 - y_{ijk}} + \lambda \sum_i \sum_k |\epsilon_{ik}|$$

1. Proximal Gradient Descent
2. Coordinate-Wise Optimization
3. Hamiltonian Monte Carlo methods (for comparison)

The first was used as a natural approach to the lasso penalized problem, the second was used to prevent the model from "spinning" during inference, and the third is the typical way in which these models were fit, used for combarability.

Takeaway

Convex methods have valuable contributions for fitting complex Latent Space Models, and their extensions

Model's Advantages

Our model has the additional inclusion of a hierarchy on the latent variables:

$$z_{ik} = b_i + \epsilon_{ik}$$

each node has a separate latent position to explain its ties within a single layer of the multigraph, but these multiple positions are all tied together by a single "base" position for that node b_i, and a layer-specific perturbation, ϵ_{ik}.

This allows two major advantages. One is graph-wise regularization with a group lasso penalty, allowing entire redundant layers to be removed for meaningful dimensionality reduction. Second, regularization allows a manual tuning between layouts that are expressive, and those that are comparable, solving the difficult problem of visual comparability.

Results

Figure 1: Figure caption

Nunc tempus venenatis facilisis. Curabitur suscipit consequat eros non porttitor. Sed a massa dolor, id ornare enim:

Treatments	Response 1	Response 2
Treatment 1	0.0003262	0.562
Treatment 2	0.0015681	0.910
Treatment 3	0.0009271	0.296

Table 1: Table caption

A Simple Test

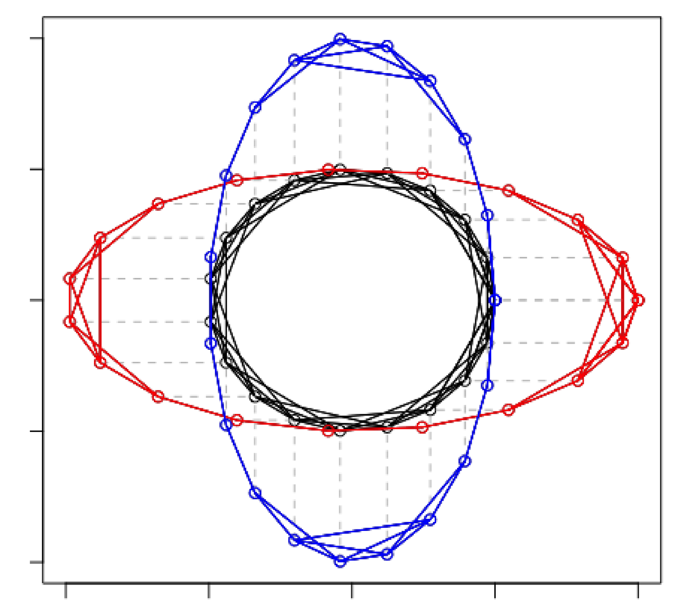

Figure 2: A Stylized Multigraph

Conclusion

We found that proximal gradient methods are an effective tool for refining an HLSM fit, given a good starting point.

References

Hoff, Peter D., Adrian E. Raftery, and Mark S. Handcock. *Latent space approaches to social network analysis.* Journal of the american Statistical association 97.460 (2002): 1090-1098.

Salter-Townshend, Michael, and Tyler H. McCormick. *Latent space models for multiview network data.* Technical Report 622, Department of Statistics, University of Washington, 2013.

Salter-Townshend. M, *Personal communication*, 2017

Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. *Stan: A probabilistic programming language.* Journal of Statistical Software 76(1). DOI 10.18637/jss.v076.i01, 2017

Loewi, A., *Parent-Teacher Relationships and Student Outcomes*, 2017

Acknowledgements

The authors would like to thank the Network Analysis group in the Department of Statistics and Data Science for their invitation to present, and the helpful feedback they contributed.