

UCLA

UCLA Previously Published Works

Title

Comment

Permalink

<https://escholarship.org/uc/item/6nx307ct>

Journal

Journal of the American Statistical Association, 112(520)

ISSN

0162-1459

Author

Handcock, Mark S

Publication Date

2017-10-02

DOI

10.1080/01621459.2017.1389737

Peer reviewed

Discussion of
“Nonparametric Bayes Modeling of
Populations of Networks”

by Durante, Dunson and Vogelstein

February 22, 2018

Author: Mark S. Handcock¹

Durante, Dunson and Vogelstein (DDV) are to be congratulated on producing an accessible paper on flexibly modeling population distributions for networks. The framework has its roots in the hierarchical projection model of Hoff et al. (2002). In that model, the nodes are represented by unknown directions in an underlying “social space” and, conditional on the directions, the ties are independent. Specifically:

$$\log \text{odds}(A_{uv} = 1 | X_u, X_v, c_{uv}, \beta) = \beta \cdot c_{uv} + X_u \cdot X_v$$

where, as in DDV, $\{A_{uv}\}$ is the binary tie variable between nodes u and v , $\{X_u\}$ are the directions of the nodes in the social space $\mathbf{R}^k$ and $\{c_{u,v}\}$ denote observed dyad-specific characteristics. This framework has been greatly developed by Peter Hoff (Hoff, 2005, 2007, 2008) extending it to a GLM-like class of additive and multiplicative effects (AME) latent factor models:

$$\log \text{odds}(A_{uv} = 1 | X_u, X_v, c_{uv}, \beta) = \beta \cdot c_{uv} + D_{uv} + \epsilon_{uv} \quad (1)$$

where $D = \{D_{uv}\}$ has singular value decomposition $D_{uv} = \sum_{r=1}^R \lambda_r X_{ur} X_{vr}$ and ϵ_{uv} is independent noise. This is a powerful and flexible way to allow dimensional-reduction on the systematic part of the dependency, D . Typically, in the application of this model only a single network is observed and the model is not further embedded.

DDV adapt this approach to hierarchically model populations of networks, having observed draws from the population. The replicates facilitate modeling the population as comprised of H sub-populations where sub-population h has model $P_h(A = a)$ and proportion ν_h . The model for each sub-population is similar to (1) with shared and sub-population specific terms:

$$\log \text{odds}_h(A_{uv} = 1 | \dots) \equiv \gamma_{uv}^{(h)} = Z_{uv} + D_{uv}^{(h)}$$
$$D_{uv}^{(h)} = \sum_{r=1}^R \lambda_r^{(h)} X_{ur}^{(h)} X_{vr}^{(h)}$$

The key features of the model in DDV are then:

- A mixture model to represent sub-population heterogeneity
- Simple sub-population modeling via conditional dyadic-independence

¹Professor and Chair of Statistics, Department of Statistics, University of California, Los Angeles, CA 90095-1554 (E-mail: handcock@ucla.edu).

- Model the dyad probabilities flexibly but tractably via a low-rank approximation (AME)

Hence DDV build on Hoff’s AME foundation and well-established mixture ideas to flexibly model more complex and realistic population distributions for networks.

They show that, due to the finiteness of A , the PMF for A can be represented this way for sufficiently large H and R . It is important to note that this assumes the conditional dyadic-independence assumption holds. A feature of the model is that it entails a swarm of parameters/latent variables that are directed by “shrinkage priors” to encourage sparsity/parsimony.

A second feature is that the conditional dyadic-independence assumption reduces the models ability to represent higher-order effects e.g., triadic structure, transitivity, path length. The effect can be seen by considering three dyads in a triad. If we knew exactly the three dyadic probabilities we would still not know the triadic probability. These models and that of AME hope that the triadic level effects are captured by the known covariates and the latent dyadic level variables. This will often be the case, as in Minhas et al. (2016), but need not be. A potential solution when the form of non-dyadic effects is hypothesized is to add exponential-family random graph terms to the sub-population models via:

$$P_h(A = a) = \frac{\exp\{\gamma^{(h)} \cdot a\}}{\kappa(\gamma^{(h)})} \quad (2)$$

where $g(a)$ are the hypothesized higher-order network statistics. However, this would add significant computational complexity.

Such approaches adapt the model to improve the parsimony and interpretability in circumstances where specific network summary measures are important in a given application (e.g., in DDV’s neuroscience application homophily by hemisphere, measures of transitivity, measures of centrality). In general, and also in the application in the paper, it is important to add nodal/dyadic attributes as they are directly interpretable and scientifically interesting e.g. intra verses inter hemispherical connections, spatial location, lobe membership. It would also be important to add terms for individual nodal heterogeneity in the propensity to form ties (again, as in AME).

The model’s primary output are posterior distributions of network summary measures e.g., network density, homophily by hemisphere, measures of transitivity, measures of centrality and degree distribution. Major findings are then assessed by these being large/small relative to what is “expected”. In DDV the reference populations are various Rényi and Erdős (1959) network models. However, these are seldom realistic models. More sophisticated reference comparisons are necessary for model assessment. For example, the question of the presence of transitivity beyond homophily in the network is difficult to answer with the current reference set.

The great strength of the model and AME is the ability to flexibly model the dyadic probabilities. However, is the model primarily a means to approximate the network generating process or should the components be interpreted and analyzed? I believe there is

great value in analyzing the components of the model to better understand the population generating process, especially the mixture proportions, the shared similarity matrix and the component-specific deviation matrices. It is important to develop numerical and graphical summaries involving the model components, especially $X^{(h)}$. There is an older sub-field on “optimal scaling” of multivariate categorical data (Gifi, 1990). This area considered similar notions of low-rank approximations to discrete multivariate data e.g. correspondence analysis, homogeneity analysis. They focused on visualizing data and structure, for example, using Biplots to visualize both node directions and tie variables. There are a lot of ideas in this literature that would be of value here.

In the neuroscience application, the data are counts of white matter fibers connecting the $V = 68$ regions (i.e. nodes). Binary networks are induced by the presence of a single fiber between the regions. However, the fiber counts are very variable (1-30,000 for dyads) so that using binary networks seems to throw away much of the information that a model could use. A weighted version of the model is called for, for example, similar to that in Krivitsky et al. (2009). A count model would also reduce the effects of false positives on fibers - See Butts (2003).

DDV have two images from each of 21 subjects. We expect greater between-subject variation than within-subject. It is natural to add this repeated observation to the model with a random effect. The subjects have many recorded characteristics such as age, sex, fiducial location, etc. These are likely to be sources of large variation between networks and incorporating them into the model as covariates would improve the interpretability. A bound of $H = 30$ clusters is used for $V = 68$ nodes (34 per hemisphere). A discussion of the effective number of clusters and their nature would provide insight. It would also be interesting to know if there were local modes in the posterior for this application. Do they expect this to be an issue in other applications?

In summary, the model and paper present a significant advance in modeling populations of networks. It is a very powerful approach, building on the AME foundation. The conditional dyadic-independence assumption can be limiting in some applications. A model of this complexity certainly requires further analysis and development. More realistic and sophisticated ways to interpret the model are needed. Given the complexity of the model, its true role will only be clear when it has been applied to many populations of populations of networks. For this, open-source and user-friendly software is required to allow its behavior to be observed and reproduced by the community of researchers.

Acknowledgements

This work was supported by supported by the National Science Foundation (NSF, grant numbers SES-1357619, IIS-1546259), the National Institute of Child Health and Human Development of the National Institutes of Health (NICHD, grant numbers R21HD075714 and R24-HD041022).

References

- Butts, C. T. (2003). Network inference, error, and informant (in)accuracy: a Bayesian approach. *Social Networks* 25(2), 103–140.
- Gifi, A. (1990). *Nonlinear Multivariate Analysis*. New York: Wiley.
- Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association* 100, 286–295.
- Hoff, P. D. (2007). Modeling homophily and stochastic equivalence in symmetric relational data. In *Proceedings of the 20th International Conference on Neural Information Processing Systems, NIPS'07, USA*, pp. 657–664. Curran Associates Inc.
- Hoff, P. D. (2008, Oct). Multiplicative latent factor models for description and prediction of social networks. *Computational and Mathematical Organization Theory* 15(4), 261.
- Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002, December). Latent space approaches to social network analysis. *Journal of the American Statistical Association* 97(460), 1090–1098.
- Krivitsky, P. N., M. S. Handcock, A. E. Raftery, and P. D. Hoff (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks* 31, 204–213.
- Minhas, S., P. D. Hoff, and M. D. Ward (2016, November). Inferential Approaches for Network Analyses: AMEN for Latent Factor Models. *ArXiv e-prints*.
- Rényi, A. and P. Erdős (1959). On random graphs, I. *Publicationes Mathematicae Debrecen* 6, 290–291.