# Discussion of "Co-citation and Co-authorship Networks of Statisticians"

Joshua Daniel Loyal* and Yuguo Chen*

*Department of Statistics, University of Illinois at Urbana-Champaign

We want to congratulate the authors on a fascinating article containing an insightful analysis and their hard work curating the high-quality co-citation and co-authorship networks. These data sets alone are a valuable contribution to the statistics profession, which will undoubtedly inspire future data science projects and advances in methodology. In fact, we are eager to use these networks in our own classrooms and research. Furthermore, the authors use these networks to tackling exciting questions in network science that go beyond the familiar problems of edge imputation and predicting node labels. In doing so, the authors perform a terrific analysis accompanied by exciting new methodology. This analysis serves as a great first step in understanding these networks, and the ideas initiated in this paper will certainly stimulate many further research questions. For example, how do individuals influence the research trajectory of others? Or, how do the components of the proposed "research map" change over time? As statisticians, we have a first-hand understanding of the complex system these networks describe, which can help us contextualize these problems and validate our inferences. As such, we look forward to this data set becoming a standard benchmark to test new models and scalable inference procedures.

A central challenge of the work is rigorously quantifying the time-varying research patterns and trends of the statistics community, which naturally leads to the statistical modeling of dynamic networks. The authors skillfully use various dynamic block models to uncover statisticians' community structure. In the remainder of this discussion, we focus on an alternative statistical network model known as latent space models. Specifically, we briefly describe the latent space modeling approach, highlight five further research questions, and demonstrate how latent space models may be used to answer them. Although other models, such as block models, may be appropriate to tackle these questions as well, we hope that this discussion gives future researchers an expanded toolset to investigate this rich data source.

Latent space models (LSMs) are a popular approach to modeling networks first proposed by Hoff et al. (2002) for static networks and later generalized to dynamic networks by Sarkar and Moore (2006) and Sewell and Chen (2015). These models embed the nodes of a

network into a low-dimensional latent space, which can provide meaningful visualizations and insights into the evolution of a network. In particular, consider $T$ binary undirected networks on a common set of $n$ nodes, and let $A_1, \ldots, A_T$ be their adjacency matrices with entries $A_t(i,j)$. Also, let $\mathbf{u}_{it} \in \mathbb{R}^d$ be the latent position of the $i$th node at time $t$. Dynamic LSMs posit that

$$\mathbb{P}(A_t(i,j) = 1) = f(h_\psi(\mathbf{u}_{it}, \mathbf{u}_{jt}), \boldsymbol{\theta}), \tag{1}$$

where $h_\psi$ is a similarity function that depends on parameters $\psi$, $f$ is an inverse-link function, and $\boldsymbol{\theta}$ are additional parameters. To capture temporal correlations, the latent positions evolve over time as Markov processes:

$$\mathbf{u}_{i1} \overset{\text{iid}}{\sim} N(0, \tau^2 I_d), \quad \mathbf{u}_{it} \sim N(\mathbf{u}_{i(t\ \ 1)}, \sigma^2 I_d), \quad t = 1, \ldots, T.$$

Furthermore, one assumes $A_1, \ldots, A_T$ are independent given the latent positions. As defined, LSMs are flexible models that can capture various properties of dynamic networks.

**1. The role of node and dyad attributes.** Incorporating additional features such as author characteristics (e.g., institution, department, academic rank, etc.) could yield interesting insights into the statistics community's co-citation and co-authorship patterns. As the authors observe in the text: "*collaborations may be driven by many factors (e.g., geographical proximity, academic genealogy, cultural ties).*" In Section 3, the authors answer this question by associating clusters inferred with a degree-corrected block model with attributes in a post hoc manner. Another approach is to incorporate the features directly into the network model. The LSM framework can formally quantify the effect of covariates on edge formation by using the following likelihood in Equation (1):

$$\text{logit}\{\mathbb{P}(A_t(i,j) = 1)\} = \beta_t^{\mathrm{T}} \mathbf{X}_{ijt} + \mathbf{u}_{it}^{\mathrm{T}} \mathbf{u}_{jt},$$

where $\mathbf{X}_{ijt}$ is a vector of dyad-specific covariates and $\beta_t$ is a time-varying vector of coefficients. This approach can be understood as a generalized bilinear mixed-effect model (Hoff, 2005, 2021). The latent positions are mean-zero random-effects ($\mathbb{E}[\mathbf{u}_{it}^{\mathrm{T}} \mathbf{u}_{jt}] = 0$) that capture residual network correlations such as transitivity. For example, we can use this model to investigate whether geographical proximity has had a decreasing effect over time on co-authorship as virtual communication platforms became popular.

**2. Inferring an evolving research map.** Just as the co-authorship community structure changes over time, it is reasonable to assume that the research areas of the research map do not remain static from 1991 to 2021. In fact, statistical network analysis has emerged as a popular research topic during this time. An alternative to the mixed-membership model for community detection involves clustering the nodes according to their positions in latent space (Handcock et al., 2007; Sewell and Chen, 2017). To infer an evolving community structure, Loyal and Chen (2022) focused on the following LSM

likelihood

$$\text{logit}\{\mathbb{P}(A_t(i,j)=1)\} = \beta_0 - \|\mathbf{u}_{it} - \mathbf{u}_{jt}\|_2,$$

and proposed a Bayesian nonparametric approach that can infer additions, deletions, splits, and mergers of communities. This model could elicit changes in statistics research areas when applied to the co-citation networks.

**3. Measuring research attraction.** We can use dynamic LSMs to answer our previous question on how individuals influence the research trajectory of others through a concept called edge attraction (Sewell and Chen, 2015). The edge attraction between nodes $i$ and $j$ measures the tendency of node $i$ to move through the latent space in the direction of another node $j$. Sewell and Chen (2015) developed a test for the presence of edge attraction between two nodes. It would be exciting to develop a similar concept for the research trajectories estimated by the dynamic DCMM to study the co-movement of statisticians' research interests.

**4. Accounting for co-citation and co-authorship counts.** When constructing the co-citation and co-authorship networks, the authors convert the weighted networks of counts into unweighted networks by applying a threshold to the edge weights. This procedure may affect the detected communities since it equates edges with low and high counts. It would be interesting to compare how the research map and co-authorship communities change (or not) when accounting for an edge's strength. In the context of LSMs, the model accounts for weighted edges by assuming the dyads in the networks, $A_t(i,j)$, arise from a generalized mixed model

$$g(\mathbb{E}[A_t(i,j)]) = \beta^{\mathrm{T}}\mathbf{X}_{ijt} + h_{\psi}(\mathbf{u}_{it}, \mathbf{u}_{jt}),$$

where $g$ is a link function. Sewell and Chen (2016) introduced likelihoods for various weighted networks, including networks with count-valued edges. As before, a clustering model can be applied to the latent positions to detect communities in the networks.

**5. Pooling information across co-citation and co-authorship networks.** The analysis in Section 3 indicates that both co-citation and co-authorship relations contain information about statistics research areas with many communities corresponding to statistics sub-fields. It would be interesting to combine these two relations by viewing the co-authorship and co-citation networks as components of a dynamic multilayer network, a collection of dynamic networks defined on a common set of nodes. Specifically, let $A_t^k$ indicate the adjacency matrix for relation $k$ (i.e., co-citation or co-authorship) measured at time $t$ with entries $A_t^k(i,j)$. To infer structure shared across the two relations, Loyal and Chen (2021) proposed modeling these adjacency matrices with a shared dynamic latent space as follows:

$$\text{logit}\{\mathbb{P}(A_t^k(i,j)=1)\} = \theta_{it}^k + \theta_{jt}^k + \mathbf{u}_{it}^{\mathrm{T}}\Lambda_k\mathbf{u}_{jt},$$

where $\theta_{it}^k \in \mathbb{R}$ models degree heterogeneity across time and relation, and $\Lambda_k$ is a diagonal matrix that allows the relations to apply different weights to the shared latent features. One can infer communities shared by the co-citation and co-authorship relations by clustering the latent positions.

Again, we want to congratulate the authors for a fine contribution. The authors do a tremendous job developing methods and theory to answer complex questions in network science. In particular, it is exciting to see the power of modern statistical network analysis in uncovering information about our academic community. We look forward to the ideas presented in this paper and the co-citation and co-authorship networks stimulating more exciting research in the future.

# References


Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007), "Model-Based Clustering of Social Networks," *Journal of the Royal Statistical Society, Series A*, 170, 301–354.

Hoff, P. D. (2005), "Bilinear Mixed-Effects Models for Dyadic Data," *Journal of the American Statistical Association*, 100, 286–295.

— (2021), "Additive and Multiplicative Effects Network Models," *Statistical Science*, 36, 34–50.

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.

Loyal, J. D. and Chen, Y. (2021), "An Eigenmodel for Dynamic Multilayer Networks," *arXiv preprint arxiv:2103.12831*.

— (2022), "A Bayesian Nonparametric Latent Space Approach to Modeling Evolving Communities in Dynamic Networks," *Bayesian Analysis*, in press.

Sarkar, P. and Moore, A. W. (2006), "Dynamic Social Network Analysis using Latent Space Models," in *Advances in Neural Information Processing Systems*, pp. 1145–1152.

Sewell, D. K. and Chen, Y. (2015), "Latent Space Models for Dynamic Networks," *Journal of the American Statistical Association*, 110, 1646–1657.

— (2016), "Latent Space Models for Dynamic Networks with Weighted Edges," *Social Networks*, 44, 105–116.

— (2017), "Latent Space Approaches to Community Detection in Dynamic Networks," *Bayesian Analysis*, 12, 351–377.