Journal of Business & Economic Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Discussion of "Co-citation and Co-authorship Networks of Statisticians" by Pengsheng Ji, Jiashun Jin, Zheng Tracy Ke, and Wanshan Li

Peter W. MacDonald, Elizaveta Levina & Ji Zhu

To cite this article: Peter W. MacDonald, Elizaveta Levina & Ji Zhu (2022) Discussion of "Co-citation and Co-authorship Networks of Statisticians" by Pengsheng Ji, Jiashun Jin, Zheng Tracy Ke, and Wanshan Li, Journal of Business & Economic Statistics, 40:2, 492-493, DOI: 10.1080/07350015.2022.2041423

To link to this article: https://doi.org/10.1080/07350015.2022.2041423

Published online: 21 Apr 2022.

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2022, VOL. 40, NO. 2, 492–493
https://doi.org/10.1080/07350015.2022.2041423

Discussion of "Co-citation and Co-authorship Networks of Statisticians" by Pengsheng Ji, Jiashun Jin, Zheng Tracy Ke, and Wanshan Li

Peter W. MacDonald, Elizaveta Levina, and Ji Zhu

Department of Statistics, University of Michigan, Ann Arbor, MI

1. Introduction

We congratulate the authors on an interesting paper and on making an important contribution to the network analysis community through compiling a large new dataset which will spur further work on multilayer, dynamic and other complex network settings. This discussion focuses on the paper's particular methods and applications in dynamic network analysis. Complexity of dynamic network data leads to many necessary analyst choices in both data processing and network modeling. Where possible, we will compare the choices made in this paper with other possibilities from recent literature on dynamic network analysis.

One of the important points of the paper is that much of our network data has always been dynamic. For instance, communication networks consisting of sent and received E-mails come with time stamps, whether we choose to incorporate them or not. Developing statistical methods that take advantage of this time varying structure will lead to greater efficiency, novel insights, and generally allow us to take full advantage of rich modern datasets like the one featured in this paper.

2. Choices in Data Processing

Dynamic network data typically arrive in one of two general settings, which we term *snapshots* and *events*. Snapshot data, consisting of time stamped networks collected at prespecified times, generally arise from the active querying of a complex system at those times. Conversely, event data, consisting of time stamped dyadic events between nodes, is a natural outcome from passive observation of a complex system. It is common, however, as this paper does, to aggregate event data into fixed time windows, which results in a format similar to snapshot data. Typically, this makes it easier to extend single network methods to the aggregated data, but there is a growing literature on directly modeling event data (Crane and Dempsey 2018; Matias, Rebafka, and Villers 2018; Kreiß, Mammen, and Polonik 2019, among others), including first steps toward models for nondyadic (hyperedge) events (Mulder and Hoff 2021).

Windowing, the choice of how to aggregate the observed events, is an important stage in the processing of dynamic network data. This paper chooses to summarize 30 years of data into 21 overlapping time windows of varying length from 5 to 10 years in Section 2, and then summarizes the same data into three overlapping time windows in Section 3. For spectral methods like those applied here, varying window length gives the analyst direct control of the edge density and ultimately the eigengap, especially important for methods which operate on snapshots independently, and thus, require sufficient signal strength for each snapshot. Overlapping and nonoverlapping windows will influence downstream modeling assumptions, with overlapping windows suggesting serial dependence among snapshot entries. As noted by Kim et al. (2018), overlapping windows can also aid interpretation by stabilizing results over time. However, stability of results to window choice is also important, and would help increase confidence in the final results.

The node set will not necessarily remain constant across dynamic network snapshots. In Section 2, the authors restrict the node set to nodes present in the first snapshot; for citations that makes sense, since a paper can continue to be cited indefinitely. However, the node set of the coauthorship network in Section 3 changes across the three snapshots. Since the proposed estimation method operates on each snapshot independently, nodes have no effect on the analysis of snapshots where they do not appear. An alternative approach might be to explicitly model the arrival and departure of nodes from the system, as done, for example, in the dynamic stochastic block model (SBM) proposed by Matias and Miele (2017). This can potentially allow for more efficient information sharing across snapshots.

3. Choices in Network Modeling

The citation network analysis in Section 2 works with 21 network snapshots constructed by aggregating overlapping time windows. The dynamic degree-corrected mixed-membership (DCMM) model assumes independence of the network snapshots after accounting for the latent community structure. Dependence across networks is a challenging new area of research, including recent work on multiple community detection with dependence (Yuan and Qu 2021), and applications of more classical time series models to individual node-pair series (Jiang, Li, and Yao 2021).

CONTACT Peter W. MacDonald *pwmacdon@umich.edu* Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1382.

The dynamic DCMM model presented here, along with many other dynamic extensions of the SBM, and more general network latent space models can be categorized according to which parameters are constant over time and which are allowed to vary. In contrast to the dynamic DCMM model, Pensky and Zhang (2019) and Matias and Miele (2017) propose dynamic extensions of the SBM in which both the community memberships and community structure matrix are allowed to vary. In order to still share information across time, Pensky and Zhang (2019) assume smoothness of the parameters, while Matias and Miele (2017) community membership is governed by a Markov chain with nontime varying parameters. These models allow for greater flexibility and potentially better fit to the data, but they are typically challenging to fit, and can present issues in model identifiability and interpretability, as we discuss below.

An often cited property of network latent space models, including the SBM, is their intrinsic nonidentifiability to a class of transformations, which could be permutations (SBM), rotations (RDPG, the random dot product graph model), or indefinite orthogonal transformations (generalized RDPG (Rubin-Delanchy et al. 2020)). In the dynamic setting, this can lead to time varying nonidentifiability, in which the parameters at each snapshot are unknown up to a snapshot-specific unknown transformation. In the case of the dynamic DCMM model, parameter sharing (a constant $K \times K$ community structure matrix P) makes the model well identified, limiting the nonidentifiability to a single unknown transformation.

Although the dynamic DCMM model is well identified, the authors still need to account for time varying nonidentifiability in their estimation procedure. An intermediate stage of their estimation procedure embeds each network snapshot, and a naive approach could produce embeddings $\{\hat{r}_i^{(t)}\}_{i=1}^n$ for each snapshot with arbitrarily rotated (misaligned) columns. While the theoretically aligned embeddings may exhibit smoothness over time, this smoothness can be lost due to misalignment. The interpretability of a dynamic network model via plots of embedded trajectories like Figure 2, or those presented in Sewell and Chen (2015), also relies on correct alignment.

The reference network projection approach taken in Section 2.2 can be viewed as a particular method of aligning a collection of $(K-1)$-dimensional network embeddings (which may be the end goal of analysis or an intermediate step followed by clustering) by representing them all in the same basis. The authors note that it also serves as a denoising step. We also note that it relies on the assumption of homogeneity of the community structure matrix over time, so that projection onto the reference network eigenvectors does not interfere with the signal strength in the other snapshots.

Other approaches could align embeddings at consecutive times by solving a (possibly indefinite) Procrustes problem (Sanna Passino et al. 2021). This approach is a simple post-processing step which is logical under the assumption of network smoothness over time, but it does not necessarily fix issues caused by misalignment in the modeling and estimation stages. Alignment can be ensured during the estimation stage by jointly embedding a concatenated object, such as an omnibus adjacency matrix (Levin et al. 2017). However, the omnibus adjacency matrix is an $nT \times nT$ object that can be computationally unwieldy for operations like singular value decomposition. Estimation efficiency of the omnibus embedding depends on the homogeneity of the network over time (Draves and Sussman 2021).

4. Conclusion

In this discussion, we have used the methodology and analysis presented in the accompanying paper to consider new opportunities presented by dynamic network data, but also new issues which require careful consideration in data processing, modeling and interpretation. At the data processing stage, we have highlighted how control over windowing and aggregation can influence signal strength, smoothness across snapshots, and heterogeneity of the (possibly time varying) node set. At the modeling stage we discuss dependence, the key issue of alignment, and argue that nonidentifiability in network models requires careful consideration in the dynamic setting. The authors of this paper have revealed many insights about their new dataset by taking advantage of its dynamic structure, but as with any statistical analysis, these insights depend on choices, and the number of choices grows quickly with the complexity of the data. Recognizing these choices and their potential alternatives is a necessary step toward a complete and principled framework for dynamic network analysis.

Funding

Levina's research is partially supported by NSF DMS grants 1916222 and 2052918. Zhu's research is partially supported by NSF DMS grant 1821243.

References

Crane, H., and Dempsey, W. (2018), "Edge Exchangeable Models for Interaction Networks," *Journal of the American Statistical Association*, 113, 1311–1326. [492]

Draves, B., and Sussman, D. L. (2021), "Bias-Variance Tradeoffs in Joint Spectral Embeddings," *arXiv:2005.02511 [math, stat]*. [493]

Jiang, B., Li, J., and Yao, Q. (2021), "Autoregressive Networks," *arXiv:2010.04492 [stat]*. [492]

Kim, B., Lee, K. H., Xue, L., and Niu, X. (2018), "A Review of Dynamic Network Models with Latent Variables," *Statistics Surveys*, 12, 105–135. [492]

Kreiß, A., Mammen, E., and Polonik, W. (2019), "Nonparametric Inference for Continuous-Time Event Counting and Link-Based Dynamic Network Models," *Electronic Journal of Statistics*, 13, 2764–2829. [492]

Levin, K., Athreya, A., Tang, M., Lyzinski, V., and Priebe, C. E. (2017), "A Central Limit Theorem for an Omnibus Embedding of Multiple Random Dot Product Graphs," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 964–967. [493]

Matias, C., and Miele, V. (2017), "Statistical Clustering of Temporal Networks Through a Dynamic Stochastic Block Model," *Journal of the Royal Statistical Society*, 79, 1119–1141. [492,493]

Matias, C., Rebafka, T., and Villers, F. (2018), "A Semiparametric Extension of the Stochastic Block Model for Longitudinal Networks," *Biometrika*, 105, 665–680. [492]

Mulder, J., and Hoff, P. D. (2021), "A Latent Variable Model for Relational Events with Multiple Receivers," arXiv:2101.05135 [stat]. [492]

Pensky, M., and Zhang, T. (2019), "Spectral Clustering in the Dynamic Stochastic Block Model," *Electronic Journal of Statistics*, 13, 678–709. [493]

Rubin-Delanchy, P., Cape, J., Tang, M., and Priebe, C. E. (2020), "A Statistical Interpretation of Spectral Embedding: The Generalised Random Dot Product Graph," arXiv:1709.05506 [cs, stat]. [493]

Sanna Passino, F., Bertiger, A. S., Neil, J. C., and Heard, N. A. (2021), "Link Prediction in Dynamic Networks Using Random Dot Product Graphs," *Data Mining and Knowledge Discovery*, 35, 2168–2199. [493]

Sewell, D. K., and Chen, Y. (2015), "Latent Space Models for Dynamic Networks," *Journal of the American Statistical Association*, 110, 1646–1657. [493]

Yuan, Y., and Qu, A. (2021), "Community Detection with Dependent Connectivity," *The Annals of Statistics*, 49, 2378–2428. [492]