Journal of Data Science 21 (3), 443–445
July 2023

DOI: 10.6339/23-JDS213EDI
Editorial

Editorial: Advances in Network Data Science

Yuguo Chen[1], Daniel Sewell[2], Panpan Zhang[3,*], and Xuening Zhu[4]

[1]*Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA*
[2]*Department of Biostatistics, University of Iowa, Iowa City, IA 52246, USA*
[3]*Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN 37203, USA*
[4]*School of Data Science, Fudan University, Shanghai, China*

This special issue features nine articles on “Advances in Network Data Science”. Data science is an interdisciplinary research field utilizing scientific methods to facilitate knowledge and insights from structured and unstructured data across a broad range of domains. Network data are proliferating in many fields, and network data analysis has become a burgeoning research in the data science community. Due to the nature of heterogeneity and complexity of network data, classical statistical approaches for network model fitting face a great deal of challenges, especially for large-scale network data. Therefore, it becomes crucial to develop advanced methodological and computational tools to cope with challenges associated with massive and complex network data analyses. This special issue highlights some recent studies in the area of network data analysis, showcasing a variety of contributions in statistical methodology, two real-world applications, a software package for network generation, and a survey on handling missing values in networks.

Five articles are published in the Statistical Data Science Section. Wang and Resnick (2023) employed point processes to investigate the macroscopic growth dynamics of geographically concentrated regional networks. They discovered that during the startup phase, a self-exciting point process effectively modeled the growth process, and subsequently, the growth of links could be suitably described by a non-homogeneous Poisson process.

Komolafe et al. (2023) described an approach for community detection in text networks. With a computationally efficient divide-and-conquer approach, this approach identifies groups of highly similar documents while allowing unique, dissimilar documents to form their own “miscellaneous” cluster. The authors provided a compelling and motivating example in the domain of patient safety research, yielding insights into trends of medical errors which cost tens of thousands of deaths and billions of dollars annually in the U.S. alone.

Chen and Chen (2023) modeled dynamic transport network as a time series of relational matrices with matrix factor models, where the observed surface network is assumed to be driven by a latent dynamic transport network with lower dimensions. In application to the international trade flow, the authors made interesting discoveries about trading hubs, centrality, trends, patterns, and change points of trading policies.

Ouyang et al. (2023) tackled social network clustering with a mixed membership model that allows each entity in the network to possess more than one membership. The likelihood of the membership parameters given the observed adjacency matrix is expressed in terms of a similarity function of the membership probability vectors of each pair of nodes. In a Bayesian framework with Markov chain Monte Carlo, the authors demonstrated inferences with real and simulated networks.

Shang et al. (2023) investigated the A/B testing problem in a network setting, where the stable-unit-treatment-value assumption does not hold due to communication and information

[*] Corresponding author. Email: panpan.zhang@vumc.org.

spreading among neighbors. The authors introduced a nonparametric approach utilizing graph cluster randomization, which computes the p-value for the observed average treatment effect by performing permutation tests at the cluster level. The test was shown to maintain its size and have substantial power in numerical studies.

The Data Science in Action Section contains two articles. Frazier et al. (2023) analyzed data from 484 students that enrolled in a large U.S. public university. The dataset included demographic details and support network information of the students. By utilizing a decision tree-based predictive model, the authors identified that various types of support played a significant role in predicting academic achievement, contingent on the students' race and gender.

O'Malley et al. (2023) aimed to find optimal physician shared-patient networks in the context of medical technology diffusion. The authors used a five-factor experiment that produces 80 distinct projections of the bipartite patient-physician mixing matrix to a unipartite physician network derived from the referral path data. The projections and their underlying factors were evaluated based on network feature heterogeneity across 2,219 hospitals and in their ability in improve predicting a hospital's adoption of a novel cardiac intervention. The factorial design setting is expected to be useful as a general methodology in network analysis.

In the Section of Computing in Data Science, Yuan et al. (2023) presented a generic, user-friendly, and efficient implementation of preferential attachment network generation with their R package *wdnet*. Additional exclusive features include allowance of multiple edges at a time, heterogeneous reciprocal edges, and user-specified preference functions. This article is expected to become a primary reference for the *wdnet* package with many citations.

The Data Science Review Section contains one article on an important topic, imputation of missing values in social network data. After a review of eight imputation methods for networks, Xu et al. (2023) conducted a simulation study to compare them under various practical scenarios. The effectiveness of an imputation method depends on the missing data type, the missing mechanism, the complexity of the social networks, and the network features of interest.

The special issue received strong support from the network research community and the Journal of Data Science editorial team. We are grateful to all the authors for their contributions and to the referees for their constructive feedback in helping our authors improve their manuscripts. The reproducibility-checking team, which consists of graduate students in Statistics at Renmin University of China, did a fantastic job in making sure the authors' code supplements work as expected in reproducing the reported results. They are to be commended for establishing the reproducibility as a shining point of the journal.

As network data science rapidly evolves, we hope that this special issue will be of interest to all network analysis scientists and attract more younger generation researchers into this exciting field.

References

Chen EY, Chen R (2023). Modeling dynamic transport network with matrix factor models: An application to international trade flow. *Journal of Data Science*, 21(3): 490–507. https://doi.org/10.6339/22-JDS1065

Frazier A, Silva J, Meilak R, Sahoo I, Broda M, Chan D (2023). Decision tree-based predictive models for academic achievement using college students' support networks. *Journal of Data Science*, 21(3): 557–577. https://doi.org/10.6339/21-JDS1033

Komolafe T, Fong A, Sengupta S (2023). Scalable community extraction of text networks for automated grouping in medical databases. *Journal of Data Science*, 21(3): 470–489. https://doi.org/10.6339/22-JDS1038

O'Malley AJ, Ran X, An C, Rockmore DN (2023). Optimal physician shared-patient networks and the diffusion of medical technologies. *Journal of Data Science*, 21(3): 578–598. https://doi.org/10.6339/22-JDS1064

Ouyang G, Dey DK, Zhang P (2023). A mixed-membership model for social network clustering. *Journal of Data Science*, 21(3): 508–522. https://doi.org/10.6339/23-JDS1109

Shang H, Shi X, Jiang B (2023). Network A/B testing: Nonparametric statistical significance test based on cluster-level permutation. *Journal of Data Science*, 21(3): 523–537. https://doi.org/10.6339/23-JDS1112

Wang T, Resnick SI (2023). Common growth patterns for regional social networks: A point process approach. *Journal of Data Science*, 21(3): 446–469. https://doi.org/10.6339/21-JDS1021

Xu Z, Hai J, Zhang Z YY (2023). Comparison of methods for imputing social network data. *Journal of Data Science*, 21(3): 599–618. https://doi.org/10.6339/22-JDS1045

Yuan Y, Wang T, Yan J, Zhang P (2023). Generating general preferential attachment networks with R package wdnet. *Journal of Data Science*, 21(3): 538–556. https://doi.org/10.6339/23-JDS1110