The Journal of Open Source Software

PANINIpy: Package of Algorithms for Nonparametric Inference with Networks In Python

Alec Kirkley[1,2,3] and **Baiyue He**[1]

1 Institute of Data Science, University of Hong Kong, Hong Kong **2** Department of Urban Planning and Design, University of Hong Kong, Hong Kong **3** Urban Systems Institute, University of Hong Kong, Hong Kong

DOI: 10.21105/joss.07312

Software

- Review
- Repository
- Archive

Editor: Vissarion Fisikopoulos

Reviewers:

- @ankurankan
- @gchure

Submitted: 19 September 2024
Published: 14 November 2024

Summary

Complex networks provide a highly flexible representation of the relational structure within a variety of real-world systems, from city streets to the Internet (Barabási, 2016). The topology and dynamics of real network data are often too complex to summarize or visualize using traditional data analysis methods, which has triggered a substantial research movement within multiple fields—including physics, computer science, sociology, mathematics, and economics among others—to develop new tools for statistical inference and machine learning specifically suited for networks.

Research on complex network inference has the goal of learning meaningful structural and dynamical regularities in network data in a manner that is often independent of the particular application of interest but relies on fundamental principles that govern a wide range of networked systems, such as transitivity, degree heterogeneity, and assortativity (Newman, 2018). A substantial amount of research within complex network inference over the last two decades has focused on the construction and evaluation of algorithms for *community detection*—the task of inferring groups of nodes that exhibit particularly strong connectivity or that have shared roles or features (Fortunato, 2010). Another major area of interest is *network reconstruction* (Peel et al., 2022), which aims to infer statistically significant functional connections from time series or other activity patterns as well as identify spurious correlations and missing edges in partially observed noisy network data. A third notable focus area within complex network inference is the clustering of network populations or multilayer networks arising in longitudinal and cross-sectional studies (Young et al., 2022).

Although community detection, network reconstruction, and network population clustering are some of the most widely researched areas in complex network inference, there are a broad range of tasks for which there is active development of new methods. For example, there is a large new body of work aimed at inferring statistically significant structure in higher order networks (Battiston et al., 2021) and networks with different types of metadata on the nodes and/or edges (Fajardo-Fontiveros et al., 2022).

Statement of Need

Due to their discrete, relational, and heterogeneous nature, complex networks present new obstacles for statistical inference. Many inference objectives on networks are intrinsically combinatorial and produce complex summaries in the form of sets or partitions. These factors make scalability and interpretability of critical importance for practical algorithms, which are not often easily accommodated within learning frameworks that focus on continuous ordered data. There are also a number of ways uncertainty can be introduced in the collection of a network dataset, whether through measurement error, sampling bias, or fluctuations across experimental settings in longitudinal or cross-sectional studies. These factors underscore the importance of developing new principled, nonparametric methods for extracting structural and dynamical regularities in networks that do not rely on ad hoc parameter choices or heuristics, allowing them to be robust in the presence of noise. These methods should be tailored and optimized for specific inference settings (e.g. network population clustering or hub detection) for the best performance in practice.

PANINIpy is a flexible and easy-to-use collection of nonparametric statistical inference methods for unsupervised learning with network data. These methods are unified in their motivation from fundamental principles–currently, the Minimum Description Length (MDL) principle underlies all the methods in *PANINIpy*—and their lack of dependence on arbitrary parameter choices that can impose unwanted biases in inference results. *PANINIpy* is highly accessible for practitioners as its methods do not require the user to manually tune any input parameters and everything is written from scratch in pure Python to be optimized for each task of interest.

PANINIpy fills an important gap in the software space by focusing on nonparametric inference methods for tasks beyond community detection and network reconstruction, for which there are many well developed and maintained Python packages (including Graph-Tool, NetworkX, iGraph, and netrd among others). There are no existing Python packages allowing for the breadth of network inference problems tackled by *PANINIpy*, which provides methods for network hub identification, temporal and multilayer network aggregation, spatially contiguous regionalization, and network backboning among other methods. By providing a unified package with these methods, users can identify parsimonious summaries of their network data from multiple perspectives, all comparable on the absolute scale of data compression in bits (for the MDL-based methods). *PANINIpy* does not have the extensive dependency requirements of existing packages and tailors its data structures for each specific network inference problem for efficient algorithmic performance and easy maintenance. *PANINIpy* therefore provides an important complement to existing network analysis packages in Python such as NetworkX that focus on network metrics and network visualization (with some methods for community detection), as well as Graph-Tool and netrd which focus on community detection and network reconstruction respectively.

Current Modules

Modules can be flexibly added to the package as needed. All modules take as input a standard representation of a network (either as an edgelist or an adjacency list in Python). The existing modules at the time of this publication are:

- **hypergraph_binning** (Kirkley, 2024c): Methods for identifying MDL-optimal temporally contiguous partitions of event data between distinct node sets (e.g. users and products).
- **population_clustering** (Kirkley et al., 2023): Methods for performing clustering of observed network populations, multilayer network layers, or temporal networks. Also includes method for generating synthetic network populations (Young et al., 2022).
- **distributional_regionalization** (Kirkley, 2022): Methods for performing MDL-based regionalization on distributional (e.g. census) data over space.
- **hub_identification** (Kirkley, 2024b): Methods for inferring hub nodes in a network using different information theoretic criteria.
- **community_regionalization** (Morel-Balbi & Kirkley, 2024): Perform contiguous regionalization of spatial network data, applicable to a wide class of community detection objectives.
- **network_backbones** (Kirkley, 2024a): Perform global and local network backboning for a weighted network.

Please refer to the documentation for details on the methodology, implementation, and usage for each module.

Acknowledgments

This work was supported by an HKU Urban Systems Institute Fellowship Grant and the Hong Kong Research Grants Council under ECS–27302523 and GRF-17301024.

References

Barabási, A.-L. (2016). *Network science*. Cambridge University Press.

Battiston, F., Amico, E., Barrat, A., Bianconi, G., Ferraz de Arruda, G., Franceschiello, B., Iacopini, I., Kéfi, S., Latora, V., Moreno, Y., & others. (2021). The physics of higher-order interactions in complex systems. *Nature Physics*, *17*(10), 1093–1098. https://doi.org/10.1038/s41567-021-01371-4

Fajardo-Fontiveros, O., Guimerà, R., & Sales-Pardo, M. (2022). Node metadata can produce predictability crossovers in network inference problems. *Physical Review X*, *12*(1), 011010. https://doi.org/10.1103/physrevx.12.011010

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3-5), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

Kirkley, A. (2022). Spatial regionalization based on optimal information compression. *Communications Physics*, *5*, 249. https://doi.org/10.1038/s42005-022-01029-4

Kirkley, A. (2024a). Fast nonparametric inference of network backbones for graph sparsification. *arXiv:2409.06417*. https://doi.org/10.48550/arXiv.2409.06417

Kirkley, A. (2024b). Identifying hubs in directed networks. *Physical Review E*, *109*, 034310. https://doi.org/10.1103/physreve.109.034310

Kirkley, A. (2024c). Inference of dynamic hypergraph representations in temporal interaction data. *Physical Review E*, *109*, 054306. https://doi.org/10.1103/physreve.109.054306

Kirkley, A., Rojas, A., Rosvall, M., & Young, J.-G. (2023). Compressing network populations with modal networks reveals structural diversity. *Communications Physics*, *6*, 148. https://doi.org/10.1038/s42005-023-01270-5

Morel-Balbi, S., & Kirkley, A. (2024). Bayesian regionalization of urban mobility networks. *Physical Review Research*, *6*, 033307. https://doi.org/10.1103/physrevresearch.6.033307

Newman, M. (2018). *Networks*. Oxford University Press. https://doi.org/10.1093/oso/9780198805090.001.0001

Peel, L., Peixoto, T. P., & De Domenico, M. (2022). Statistical inference links data and theory in network science. *Nature Communications*, *13*(1), 6794. https://doi.org/10.1038/s41467-022-34267-9

Young, J.-G., Kirkley, A., & Newman, M. E. (2022). Clustering of heterogeneous populations of networks. *Physical Review E*, *105*(1), 014312. https://doi.org/10.1103/physreve.105.014312