# PANINIpy: Package of Algorithms for Nonparametric Inference on Networks in Python

**Authors:**

- **Alec Kirkley**
  Institute of Data Science, University of Hong Kong, Hong Kong
  Department of Urban Planning and Design, University of Hong Kong, Hong Kong
  Urban Systems Institute, University of Hong Kong, Hong Kong
  **ORCID**: 0000-0001-9966-0807
- **Baiyue He**
  Institute of Data Science, University of Hong Kong, Hong Kong
  **ORCID**: XXXX-XXXX-XXXX-XXXX

**Tags:**
Python, complex networks, nonparametric inference, minimum description length, unsupervised learning

**Date:**
24 August 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 2013). 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 networks such as transitivity, degree heterogeneity, and assortativity (Newman 2018). For example, a major research interest in this area over the last two decades has been the construction and evaluation of a vast array of algorithms for *community detection*, which aim to infer highly connected subsets of nodes to summarize the mesoscale structure of a network (Fortunato 2010). Another major area of interest is *network reconstruction* (Peel,

Peixoto, and De Domenico 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 focus area within complex network inference is the clustering of network populations or multilayer networks arising in longitudinal and cross-sectional studies (Young, Kirkley, and Newman 2022).

Although community detection, network reconstruction, and clustering network populations 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, Guimerà, and Sales-Pardo 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 emphasize the importance of developing new principled and flexible 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.

`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—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 without reliance on existing packages. `PANINIpy` therefore provides an important complement to existing network analysis packages such as `NetworkX` that focus primarily on network metrics, network visualization, and community detection.

## Related Software Packages

There are number of existing Python packages containing individual methods that perform nonparametric inference with networks, but none that are unified under this scope with the ease-of-use of `PANINIpy`. The `Graph-Tool` (Peixoto 2014) package includes a number of principled Bayesian methods for complex network inference, many of which are nonparametric. As its core functionalities are implemented in C++, `Graph-tool` is also quite efficient given the computational demand of the inference problems it considers. However, the data structures in Graph-tool are often challenging to navigate for new users and its optimization routines are largely dependent on MCMC methods which are highly flexible but tricky to tune. Other popular packages such as `NetworkX` (Hagberg, Swart, and Schult 2008) and `iGraph` (Csardi and Nepusz 2006) also have methods for complex network inference but are much broader in scope, being used primarily for network summary statistics and visualization. `PANINIpy` fills an important gap in the software space for network inference methods with very simple dependencies in pure Python.

## 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).

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

## Acknowledgments

The authors thank Sebastian Morel-Balbi for useful discussions. 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, Albert-László. 2013. "Network Science." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371 (1987): 20120375.

Battiston, Federico, Enrico Amico, Alain Barrat, Ginestra Bianconi, Guilherme Ferraz de Arruda, Benedetta Franceschiello, Iacopo Iacopini, et al. 2021. "The Physics of Higher-Order Interactions in Complex Systems." *Nature Physics* 17 (10): 1093–98.

Csardi, Gabor, and Tamas Nepusz. 2006. "The Igraph Software Package for Complex Network Research." *InterJournal* Complex Systems: 1695. https://igraph.org.

Fajardo-Fontiveros, Oscar, Roger Guimerà, and Marta Sales-Pardo. 2022. "Node Metadata Can Produce Predictability Crossovers in Network Inference Problems." *Physical Review X* 12 (1): 011010.

Fortunato, Santo. 2010. "Community Detection in Graphs." *Physics Reports* 486 (3-5): 75–174.

Hagberg, Aric, Pieter J Swart, and Daniel A Schult. 2008. "Exploring Network Structure, Dynamics, and Function Using NetworkX." Los Alamos National Laboratory (LANL), Los Alamos, NM (United States).

Kirkley, A. 2022. "Spatial Regionalization Based on Optimal Information Compression." *Communications Physics* 5: 249. https://arxiv.org/abs/2111.01813.

———. 2024a. "Identifying Hubs in Directed Networks." *Physical Review E* 109: 034310. https://arxiv.org/abs/2312.03347.

———. 2024b. "Inference of Dynamic Hypergraph Representations in Temporal Interaction Data." *Physical Review E* 109: 054306. https://arxiv.org/abs/2308.16546.

Kirkley, A., A. Rojas, M. Rosvall, and J.-G. Young. 2023. "Compressing Network Populations with Modal Networks Reveals Structural Diversity." *Communications Physics* 6: 148. https://arxiv.org/abs/2209.13827.

Morel-Balbi, S., and A. Kirkley. 2024. "Urban Boundary Delineation from Commuting Data with Bayesian Stochastic Blockmodeling: Scale, Contiguity, and Hierarchy." https://arxiv.org/abs/2405.04911.

Newman, Mark. 2018. *Networks*. Oxford university press.

Peel, Leto, Tiago P Peixoto, and Manlio De Domenico. 2022. "Statistical Inference Links Data and Theory in Network Science." *Nature Communications* 13 (1): 6794.

Peixoto, Tiago P. 2014. “The Graph-Tool Python Library.” *Figshare*. https://doi.org/10.6084/m9.figshare.1164194.

Young, Jean-Gabriel, Alec Kirkley, and Mark EJ Newman. 2022. “Clustering of Heterogeneous Populations of Networks.” *Physical Review E* 105 (1): 014312.