# ScholarCodeCollective: A Python Package for Nonparametric Inference in Network Data

24 August 2024

## ScholarCodeCollective: A Python Package for Nonparametric Inference in Network Data

**Authors:** - Baiyue He (Institute of Data Science, University of Hong Kong, Hong Kong) - Alec Kirkley (Institute of Data Science, University of Hong Kong, Hong Kong)

**Date:** 24 August 2024

**Tags:** Python, network analysis, graph theory, hypergraph binning, community detection, spatial regionalization, nonparametric inference, unsupervised machine learning

### Summary

In the modern era of data-driven research and network science, analyzing complex systems for meaningful patterns from intricate data structures such as social networks, urban regions, and hypergraph structures requires robust tools for tasks like spatial regionalization, temporal partitioning, and community detection. Traditional methods often encounter limitations in efficiency and scalability, particularly when working with large, multidimensional, and temporally dynamic datasets across diverse networks. To overcome these challenges, we introduce ScholarCodeCollective, a Python package designed to facilitate nonparametric inference for unsupervised learning tasks involving network data. Under the principles of parsimony and the Minimum Description Length, this package provides a versatile framework that is both flexible and capable of incorporating new methods in the future.

### Statement of Need

The growing complexity and heterogeneity of data across various scientific domains have increased the need for advanced analytical tools capable of uncovering underlying patterns in a flexible, data-driven manner without imposing restrictive assumptions. Network data poses unique challenges due to its intricate structure, dynamic characteristics, and varying scales. Traditional parametric methods often struggle to manage these complexities, leading to oversimplifications that overlook critical insights, as highlighted in recent studies on large-scale network modeling (Goldenberg et al. 2010; Vu, Hunter, and Schweinberger 2013). While efficient Python modules like Graph-tool exist for graph manipulation and statistical analysis within the network community, their core structures and algorithms are implemented in C++ (Peixoto 2014), creating difficulties for some researchers who wish to apply these tools to their own studies. The Minimum Description Length (MDL) principle has been a framework for model selection and complexity control, particularly in applications that require parsimonious models in unsupervised learning tasks. (Grünwald 2007; Rissanen 2007).

ScholarCodeCollective, under the principles of MDL and developed entirely in Python, is designed to fill this gap by offering a comprehensive suite of tools for nonparametric inference in network data across various unsupervised learning tasks. Nonparametric methods have become increasingly important in network analysis due to their ability to model complex relational data without relying on predefined distributional assumptions, thus making them highly adaptable to diverse types of networks, especially in dynamic and large-scale networks (Teh et al. 2004; Hoff 2009). However, traditional nonparametric approaches can be computationally demanding and challenging to generalize across different network structures and data modalities (Newman 2018). There is urgent need for such a unified framework that can apply nonparametric principles across various network analysis tasks, ensuring consistency, scalability, and interpretability of the results.

The key strength of the ScholarCodeCollective package is its flexibility. By employing a nonparametric approach, the package can be readily extended with new methods and techniques as they emerge, ensuring its continued relevance and adaptability to the evolving demands of network analysis. Future updates might include new modules for novel network structures, integration with other machine learning frameworks, or enhancements to existing algorithms driven by user feedback and advancements in the field. This flexibility stems from the package's unified underlying principles, making ScholarCodeCollective as a robust tool for addressing both current and future challenges in network analysis.

## Module Features

One of the core modules of the package is the implementation of the hypergraph binning algorithm, which is derived from the work of Kirkley (2024b). By minimizing the description length (DL), this package aids researchers in compressing and interpreting dynamic interaction data, making it easier to identify and analyze temporally contiguous partitions within large datasets. Traditional graph-based methods often struggle with the complexities of temporal interactions and hypergraph structures However, this algorithm efficiently processes large-scale event data to uncover underlying structures, offering both exact dynamic programming and faster greedy agglomerative methods for inferring dynamic hypergraph representations from temporal data.

Another core module of the package is its ability to identify modal networks and cluster graph populations, based on the work of Kirkley et al. (2023). The robust MDL-based clustering algorithm enables researchers to compress and identify structural diversity within large sets of node-aligned graphs, offering deeper insights into underlying network structures. The algorithm accommodates various network types, including bipartite and directed networks, and provides flexible clustering options, to uncover structural diversity within populations, identify representative modal networks.

The third core module of the package is designed to optimally partition spatial units into regions based on the distribution of various categories, such as demographic groups. This approach, derived from the work of Kirkley (2022), enables researchers to identify spatial regions that minimize the information required to describe the distributional data, for better understanding spatial heterogeneity in fields. Traditional regionalization methods often fall short in capturing the full complexity of distributional data in spatial analysis. In contrast, the MDL-based regionalization algorithm optimally partitions spatial units by considering the distribution of categories across the region to minimize information loss, leading to more accurate and meaningful analyses of spatial data.

The fourth core functionality of the package features a powerful algorithm for identifying hub nodes in directed networks using MDL principles. As detailed in the work of Kirkley (2024a), this method detects of hub nodes that optimally compress the network's structural information. The algorithm supports multiple encoding schemes, including Erdős–Rényi (ER) and Configuration Model (CM) encodings, applicable to both simple and multigraph representations. Traditional hub detection methods, such as degree centrality, often fail to capture the intricate dynamics between nodes in directed or weighted networks, where the MDL-based algorithm identifies hubs while accounting for the directed or weighted nature of the network, resulting in more accurate and insightful identification of central nodes in complex networks.

The current final core module offers a powerful tool for spatial regionalization, specifically designed to delineate urban boundaries based on commuting data using a Bayesian Stochastic Blockmodeling approach. This method, as detailed in the work of Morel-Balbi and Kirkley (2024), this method facilitates the identification of contiguous regions within urban areas by minimizing the Description Length (DL) of a weighted stochastic block model, as stochastic block models and their nonparametric extensions have become a standard approach for community detection and network modeling (Karrer and Newman 2011; Fortunato and Hric 2016). Traditional regionalization techniques often struggle with ensuring spatial contiguity while managing the complexities and hierarchical nature of urban boundaries. However, this module addresses these challenges by implementing an MDL-based greedy agglomerative algorithm within spatial networks, offering a robust and scalable solution for the task of urban boundary delineation.

## Application

### Binning Temporal Hypergraphs

Using data from an e-commerce platform as example, where interactions between users and purchased goods or web visits are recorded over time, the hypergraph binning algorithm from ScholarCodeCollective can provide valuable insights. By applying the exact dynamic programming method, researchers can identify fine-grained structures in the data, revealing high-resolution temporal clusters, while the greedy method offers a faster yet slightly less precise solution. Both methods contribute to compressing the event data, enabling researchers to uncover underlying temporal clusters. These insights can be used to arrange promotions or stock up on items prior to periods of frequent purchases by specific user groups, revealing trends such as seasonal buying patterns or the effects of market advertising.

### Clustering Network Populations

When studying a population of social networks, each representing interactions between individuals over time, a researcher can utilize the population clustering module to identify common interaction patterns from modal networks and quantify the diversity within the population. This approach can reveal various social structures or the presence of distinct communities within the population. Companies or political parties can use uncover communities of interest by clustering networks of users based on their interaction patterns, to target specific user groups to enhance the effectiveness of their engagement strategies.

### Regionalization with Distributional Data

If a researcher is studying the spatial distribution of demographic groups within a metropolitan area, they can use the distributional regionalization module to partition the area into regions that minimize the information required to describe the demographic distribution. This approach leads to more accurate and meaningful spatial analyses, potentially revealing distinct neighborhoods with similar demographic profiles or areas of high diversity. By clustering neighborhoods based on demographic distributions, urban planners can identify regions with comparable characteristics, which aids in public resource allocation and policy development.

### Identifying Network Hubs

The identification of hub nodes is crucial for understanding the structural and functional dynamics of networks, particularly in contexts like transportation and information dissemination (Barabási and Albert 1999). By using the MDL-based hub identification to determine which nodes serve as hubs in the network, offering a more comprehensive approach than traditional centrality measures. The algorithm not only identifies these hubs but also quantifies how effectively they compress the network's structural information, providing deeper insights into their significance. Similarly, in a transportation network, the hub identification module can be used to pinpoint key transportation hubs, such as major airports or train stations. These hubs are vital for optimizing the flow of people and traffics as the prime targets for infrastructure development.

### Regionalization with Community Detection

Consider a scenario where an urban planner is tasked with redeveloping a metropolitan area based on commuting patterns. The community regionalization module provides a sophisticated approach to regionalization that accounts for both spatial contiguity and hierarchical structure. By applying the greedy agglomerative algorithm to the spatial and flow data of the region, the planner can identify boundaries that minimize the DL, which are invaluable for planning public health, education, and housing, ultimately enhancing public services distribution across urban areas.

## References

Barabási, Albert-László, and Réka Albert. 1999. "Emergence of Scaling in Random Networks." *Science* 286 (5439): 509–12.

Fortunato, Santo, and Darko Hric. 2016. "Community Detection in Networks: A User Guide." *Physics Reports* 659: 1–44.

Goldenberg, Anna, Alice X Zheng, Stephen E Fienberg, Edoardo M Airoldi, et al. 2010. "A Survey of Statistical Network Models." *Foundations and Trends® in Machine Learning* 2 (2): 129–233.

Grünwald, Peter D. 2007. *The Minimum Description Length Principle.* MIT press.

Hoff, Peter D. 2009. *A First Course in Bayesian Statistical Methods.* Vol. 580. Springer.

Karrer, Brian, and Mark EJ Newman. 2011. "Stochastic Blockmodels and Community Structure in Networks." *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 83 (1): 016107.

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.

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

Rissanen, Jorma. 2007. *Information and Complexity in Statistical Modeling*. Springer Science & Business Media.

Teh, Yee, Michael Jordan, Matthew Beal, and David Blei. 2004. "Sharing Clusters Among Related Groups: Hierarchical Dirichlet Processes." *Advances in Neural Information Processing Systems* 17.

Vu, Duy Q, David R Hunter, and Michael Schweinberger. 2013. "Model-Based Clustering of Large Networks." *The Annals of Applied Statistics* 7 (2): 1010.