NetPointLib: Library for Large-Scale Spatial Network Point Data Fusion and Analysis

Yunfan Kang
University of Illinois Urbana-Champaign
Urbana, Illinois, USA
yfkang@illinois.edu

Fangzheng Lyu
Virginia Tech
Blacksburg, Virginia, USA
fangzheng@vt.edu

Shaowen Wang
University of Illinois Urbana-Champaign
Urbana, Illinois, USA
shaowen@illinois.edu

ABSTRACT

Network-constrained events, including for example traffic accidents and crime incidents, are widespread in urban environments. Understanding spatial patterns of these events within network spaces is essential for deciphering the underlying dynamics and supporting informed decision-making. The fusion and analysis of network-constrained point data pose significant computational challenges, particularly with large datasets and sophisticated algorithms. In this context, we introduce NetPointLib, a computationally efficient library designed for processing and analyzing large-scale event data in network spaces. NetPointLib utilizes the capabilities of high-performance computing (HPC) environments including ROGER supercomputer, ACCESS resources, and the CyberGISX platform, providing a scalable and accessible framework for conducting network point data fusion and pattern analysis and supporting computational reproducibility. The library encompasses several algorithmic implementations, including the network local K function and network scan statistics, to enable researchers and practitioners to identify spatial patterns within network-constrained data. This is achieved by harnessing the computational power of HPC resources, facilitating advanced spatial analysis in an efficient and scalable manner.

CCS CONCEPTS

• **Software and its engineering** → **Software libraries and repositories**; • **Information systems** → **Geographic information systems**; **Web services**.

KEYWORDS

Spatial Network, Point Pattern Analysis, Geographic Information Science and Systems (GIS), CyberGIS, High-Performance Computing

ACM Reference Format:
Yunfan Kang, Fangzheng Lyu, and Shaowen Wang. 2024. NetPointLib: Library for Large-Scale Spatial Network Point Data Fusion and Analysis. In *Practice and Experience in Advanced Research Computing (PEARC '24), July 21–25, 2024, Providence, RI, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3626203.3670615

This work is licensed under a Creative Commons Attribution International 4.0 License.

PEARC '24, July 21–25, 2024, Providence, RI, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0419-2/24/07
https://doi.org/10.1145/3626203.3670615

1 INTRODUCTION

As the volume of available spatial data continues to grow, the proliferation of spatial data across various domains—ranging from urban planning and public health to transportation and environmental monitoring—has presented unique challenges and opportunities. The large volume, high velocity, and extensive variety characteristics of spatial big data necessitate scalable analytical tools to extract information, identify patterns, and inform impactful decisions.

A significant portion of spatial data is inherently tied to or significantly influenced by network spaces, such as transportation systems, utility infrastructures, and waterways. This network-constrained nature of spatial data, from the travel patterns of individuals influencing crime locations [13] to the role of road characteristics in traffic accidents [3], requires advanced analytical methods that transcend traditional planar spatial techniques. Planar methods often fall short, producing inaccuracies when applied to network spaces due to their fundamentally different natures, underscoring the need for network-specific analytical approaches [7].

While point pattern analysis and spatial scan statistics have long been established in spatial analysis, adapting these methods to network contexts represents a contemporary research frontier. Statisticians and geospatial scientists are aware of the importance of analyzing events in network spaces and continue to improve the concepts and statistical computation methods to better capture the unique properties of spatial network patterns[12]. However, the high computational demands of these methods have limited their scalability, restricting their analysis to datasets comprising only limited numbers of data points [1, 2, 4]. Creating datasets suitable for these analyses also presents unique challenges. Recent studies continue to utilize SANET 3 [9, 10] and SANET 4 [6] for data fusion purposes. However, these tools are only compatible with ArcGIS versions 9 and 10, which have been deprecated for over a decade. Additionally, accessing and deploying these outdated extensions can be quite challenging.

Harnessing High-Performance Computing (HPC) offers a path forward, enabling the analysis of extensive data sets with complex network constraints. Yet, translating theoretical advancements into practical, scalable solutions remains a hurdle, limiting the broader adoption and application of network-based spatial analysis techniques. Recognizing this gap, we introduce a comprehensive Python library, NetPointLib, designed to leverage HPC for efficient, scalable creation and analysis of network-constrained spatial data. The implementation and sample data for NetPointLib is available at https://github.com/cybergis/NetPointLib. This library facilitates large-scale data fusion and advanced spatial-temporal analysis, including point pattern analysis, network scan statistics, and event

forecasting, tailored for network spaces. The paper presents the first version of the library, outlining its foundational concepts, capabilities, and potential applications. While already a useful and efficient tool, the library is subject to continuous improvement, with future versions planned to incorporate broader functionalities and enhancements based on user feedback and evolving research needs. The NetPointLib is deployed on the CyberGISX platform [14] (https://cybergisxhub.cigi.illinois.edu/), which has emerged as a critical resource for computation- and data-intensive geospatial research and education. CyberGISX provides seamless access to an online Jupyter Notebook environment backed by advanced cyberinfrastructure, including the first geospatial supercomputer named ROGER, while shielding the complexity of managing cyberinfrastructure access from users [11].

CyberGISX significantly enhances computational reproducibility by facilitating the sharing of computational notebooks, data, methods, and results, thus bridging the gap between theoretical frameworks and practical implementations. Through this integration, we strive to democratize access to advanced network-based spatial analysis capabilities, enabling a wide range of applications across diverse domains, including transportation network analysis, crime pattern detection, epidemiological studies, and resource allocation optimization. By bridging the gap between theory and practice, NetPointLib aims to unlock the full potential of network-constrained spatial big data, fostering innovation in fields where network spaces play a central role.

2 NETPOINTLIB: LIBRARY FOR SPATIAL NETWORK POINT DATA FUSION AND ANALYSIS

This section provides an in-depth look at NetPointLib, a comprehensive package for the fusion and analysis of spatial network point data. As shown in Figure 1, NetPointLib is structured into two main subpackages: Data Creation and Fusion, and Spatial Network Point Analysis, each designed to streamline workflows for users.

2.1 Data Creation and Fusion

The Data Creation and Fusion subpackage is engineered to merge real or synthetic point datasets with spatial networks. It outputs data in compliance with SANET specifications while being compatible with modern analytical tools and techniques in Python, such as Shapely and NumPy.

For real-world applications, users specify both the target spatial network and event point dataset. Spatial networks are formatted as NetworkX graphs, facilitating the direct incorporation of road networks from the Open Street Map via the OSMnx package. This subpackage supports simplified topology graphs and those associated with geometries, broadening its applicability. Event data points are managed as Pandas dataframes, with longitude and latitude coordinates for each event. Utilizing CyberGIS-Compute [14] as middleware, this subpackage harnesses HPC resources to execute data fusion tasks. A divide-and-conquer strategy partitions a spatial network into manageable grids of sub-networks, enhancing processing speed through parallelization. This approach ensures efficient data projection by indexing sub-network edges with R-trees and later recombining them into a unified network, preserving data integrity and spatial accuracy.

Synthetic data generation is another key feature, allowing for the insertion of data points onto network edges based on specified distributions. The default Gaussian distribution aids in Monte Carlo simulations for network k-function and scan statistics analysis, providing a versatile tool for spatial data science research.

2.2 Spatial Network Point Analysis

The Spatial Network Point Analysis subpackage provides implementations of several widely adopted algorithms that can be parallelized on HPC and used directly in notebooks for users to analyze spatial point patterns in network spaces. The datasets created from the Data Creation and Fusion subpackage can be directly analyzed using the algorithms provided.

The Point Pattern Analysis module includes the implementation of the network global auto K function and the network local auto K function [7], which are spatial statistical analysis methods used to assess the clustering of point features within a network. The network global auto K function assesses clustering tendencies and spatial relationships. The network local auto K function determines whether the event points on the sub-network are clustered or not. Chained with the sub-network enumeration strategy, the network local auto K function is able to identify all statistically robust clusters in the given network and can be used to identify the location of a criminal or event hotspot.

The Network Scan Statistic module implements the network-based Kulldorff spatial scan statistic algorithm [9, 10]. The algorithm scans through the network with specified network-based search windows and identifies clusters by performing likelihood ratio tests of events within a study area against events outside of the study area. The iso-distance subnetwork definition is used to define the shape of the default search window and the total length of the line segment within the search window is used to define its size. Poisson and Bernoulli distributions are included while the user can pass the definition of the distributions as an argument to the network scan statistic function. The network work scan statistic algorithm is also able to perform expectation-based space-time network anomaly detection with the simulation method in the Data Creation and Fusion subpackage and time-series forecasting methods [5].

The library's ongoing development is underscored, with plans to incorporate deep learning for enhanced spatial-temporal event forecasting within network spaces. This commitment to expanding NetPointLib's functionalities highlights the library's potential to remain at the forefront of spatial network analysis capabilities.

2.3 Computation Efficiency and CyberGIS Capability

To project event data points onto a network, we use an R-tree to identify the nearest network edge for each point, followed by projection and insertion. Given n points and m edges, sequential projection using OSMnx and Shapely libraries has a time complexity of $O(n \log m)$. To enhance efficiency for large-scale networks, we divide the network into smaller, overlapping sub-networks. This reduces the complexity to $O(\log(\frac{m}{k}))$ per point within each sub-network, where k is the number of partitions. Leveraging parallel

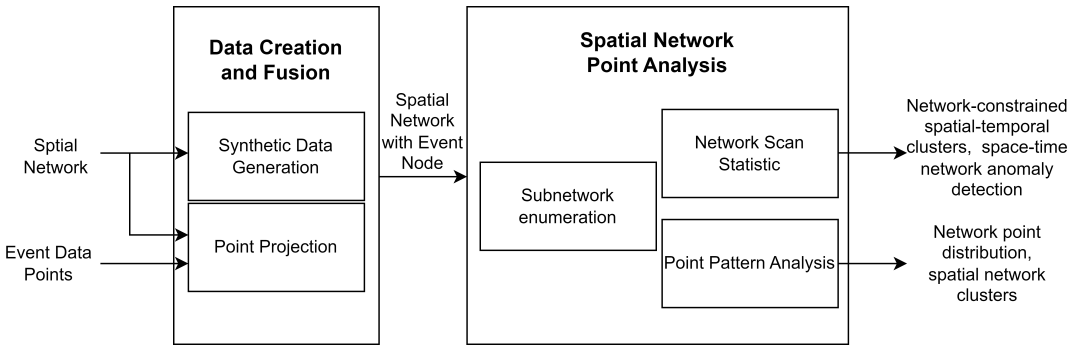

Figure 1: NetPointLib overview

processing in an HPC environment reduces the overall time complexity to $O(\frac{n}{p}\log(\frac{m}{k}))$, where p is the number of parallel processes. This strategy, enabled by CyberGISX and HPC resources, significantly optimizes the data fusion procedure.

3 CHICAGO CRIME ANALYSIS WITH NETPOINTLIB

This section showcases the application of NetPointLib to process and analyze spatial data using the computational power of HPC in a cloud environment. We demonstrate this through a case study involving crime data in Chicago [8], specifically from February 5 to 27, 2024, comprising 14,417 data points, alongside the city's road network, which includes 258,124 nodes and 809,614 edges.

NetPointLib, accessible as an open-source library on GitHub and through a dedicated online notebook on the CyberGISX platform, facilitates seamless data fusion on the platform. Users can access CyberGISX for free after authenticating with CILogon and navigating to the specific notebook designed for NetPointLib operations.

3.1 Fusion of the Crime and Road Network Datasets

To integrate the Chicago crime dataset with the road network, users employ the *multiMapMatching.match_points_to_network* function, specifying "Latitude" and "Longitude" for the x and y coordinates, respectively. This fusion task is executed on HPC resources, allowing users to run complex computations online without taxing their local machines.

Figure 2 a) visualizes the crime data points, now accurately projected onto the Chicago road network, with crime locations marked in red against the grey lines of the road network. Additionally, the library supports the creation of synthetic datasets for analysis. To perform such synthetic data creation, the *multiMapMatching.generate_points_on_network* function takes the spatial network, the target point process, and the number of points to be generated as the input. Users can optionally specify an attribute for calculating the length of network edges in topologically simplified graphs; otherwise, the default behavior is to compute edge lengths using the coordinates of the endpoints. Figure 2 b) shows the synthetic data generated on the Chicago road network with the event points highlighted in red. The size of the dots representing the event points is set to be a smaller value for visualization. The number of points generated matches the real dataset, and their distribution assumes complete spatial randomness. The process of data fusion and synthetic data generation, leveraging HPC-based parallelization, is completed in less than 1 hour. By contrast, the projection and interpolation methods provided by the shapely library can only process less than 200 points within the same amount of time under the same configuration.

3.2 Analysis Using Spatial Network Point Analysis

With the data now fused, analysis can proceed using NetPointLib's Spatial Network Point Analysis capabilities. Comparing Figure 2 a) and b), the distribution of the actual crime cases appears to be dispersed in the northwest portion of the city and clustered on the east side. To further validate these observations within the network space and identify the actual locations of crime hotspots, the *analysis.network_local_k* function is invoked using the fused dataset as input. The network segments highlighted in red in Figure 2 c) indicate areas where crime clustering exceeds the 95% confidence level for statistical significance, as determined by the network local auto k function. This analysis identifies hotspots by evaluating clusters within isodistance subnetworks, with a maximum travel distance set to 1,600 meters. The *analysis.network_scan* function can also be utilized to identify the most significant, as well as secondary, hotspots. This function compares the actual number of observed crimes within 1-kilometer network segments to a baseline case according to a Poisson process and calculates the likelihood ratios. Figure 2 d) identifies the most significant crime hotspot. The expected number of crimes on the sub-network was calculated to be 70.24, whereas the actual observed was 209, highlighting a significant deviation and pinpointing a critical area for further investigation or resource allocation.

4 CONCLUSION AND FUTURE WORK

This paper describes NetPointLib, a comprehensive library designed to streamline the fusion, generation, analysis, and visualization of large-scale spatial network point data. The integration of HPC with NetPointLib is pivotal for managing and analyzing such data, as demonstrated in the case study of crime data analysis in Chicago. The task of processing and analyzing such complex datasets, with

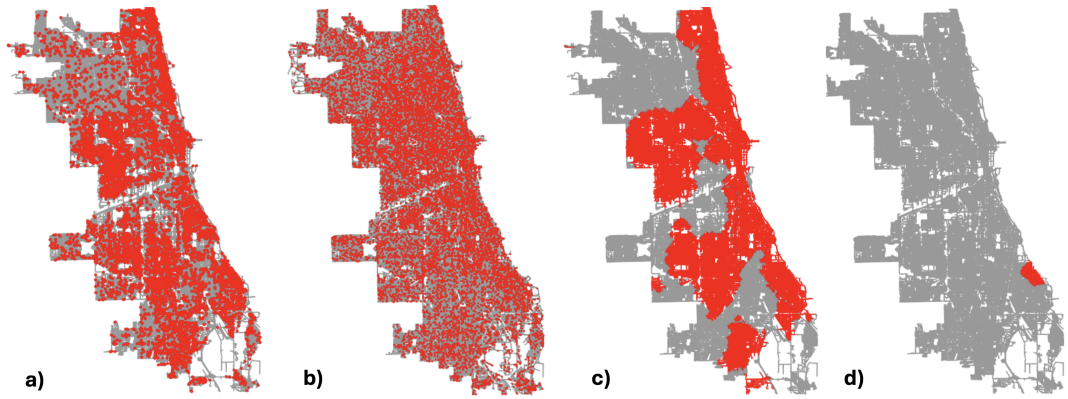

Figure 2: a) Chicago road network with crime data, b) synthetic data, c) network local auto *k* function hotspots, d) most significant network scan statistic hotspot

thousands of events across the vast network infrastructure, is beyond the capability of standard PCs due to constraints like limited multiprocessing capabilities and insufficient RAM. For example, a workstation equipped with an Intel Xeon W-2223 processor and 64GB of RAM took approximately two days to complete the tasks that, through the integration of HPC and CyberGIS-Compute, were accomplished within just four hours. This integration not only dramatically accelerates processing times but also runs remotely, thereby conserving local computational resources. Moreover, the use of CyberGIS-Compute ensures that the tool is accessible from anywhere, enhancing user convenience and facilitating broader adoption. The CyberGISX platform further supports computational reproducibility, ensuring that results are consistent and verifiable across different instances. This robust integration transforms complex spatial data analysis into a more feasible and efficient process, opening up new possibilities for researchers and practitioners working with spatial network data.

As we look to the future, our focus will remain on enhancing the utility and accessibility of NetPointLib. Key areas of development include refining the API and the usability of the notebooks, expansion of analysis capabilities, incorporating advanced algorithms to exploit HPC, and enabling spatial-temporal event forcasting based on deep learning approaches. As we continue to develop and refine this library, our goal is to democratize access to advanced spatial analysis tools, enabling researchers and decision-makers to uncover deep insights into network-constrained phenomena and make informed decisions. We invite the research community to join us in this endeavor, contributing ideas, expertise, and feedback to shape the future of spatial network analysis.

ACKNOWLEDGMENTS

This research and associated materials are based in part upon work supported by the National Science Foundation under grant numbers: 2112356, 2118329, and 2321070. Our computational work used ROGER, which is a geospatial supercomputer supported by the CyberGIS Center for Advanced Digital & Spatial Studies and the School of Earth, Society, & Environment at the University of Illinois Urbana-Champaign.

REFERENCES

[1] Ottmar Cronie, Mehdi Moradi, and Jorge Mateu. 2020. Inhomogeneous higher-order summary statistics for point processes on linear networks. *Statistics and Computing* 30, 5 (Sept. 2020), 1221–1239. https://doi.org/10.1007/s11222-020-09942-w

[2] Nicoletta D'Angelo, Giada Adelfio, Antonino Abbruzzo, and Jorge Mateu. 2022. Inhomogeneous spatio-temporal point processes on linear networks for visitors' stops data. *The Annals of Applied Statistics* 16, 2 (June 2022), 791–815. https://doi.org/10.1214/21-AOAS1519 Publisher: Institute of Mathematical Statistics.

[3] Nicoletta D'Angelo, David Payares, Giada Adelfio, and Jorge Mateu. 2022. Self-exciting point process modelling of crimes on linear networks. *Statistical Modelling* (2022), 1471082X221094146.

[4] Matthias Eckardt and Mehdi Moradi. 2023. Marked spatial point processes: current state and extensions to point processes on linear networks. https://doi.org/10.48550/arXiv.2309.01511 arXiv:2309.01511 [stat].

[5] Chance Haycock, Edward Thorpe-Woods, James Walsh, Patrick O'Hara, Oscar Giles, Neil Dhir, and Theodoros Damoulas. 2020. An expectation-based network scan statistic for a covid-19 early warning system. *arXiv preprint arXiv:2012.07574* (2020).

[6] David S Lamb, Joni A Downs, and Chanyoung Lee. 2016. The network K-function in context: examining the effects of network structure on the network K-function. *Transactions in GIS* 20, 3 (2016), 448–460.

[7] Atsuyuki Okabe and Kokichi Sugihara. 2012. *Spatial analysis along networks: statistical and computational methods.* John Wiley & Sons.

[8] Chicago Data Portal. 2024. Chicago Crimes Map. https://data.cityofchicago.org/Public-Safety/Crimes-Map/dfnk-7re6/. [Online; accessed 27-Feb-2024].

[9] Shino Shiode and Narushige Shiode. 2020. A network-based scan statistic for detecting the exact location and extent of hotspots along urban streets. *Computers, Environment and Urban Systems* 83 (Sept. 2020), 101500. https://doi.org/10.1016/j.compenvurbsys.2020.101500

[10] Shino Shiode and Narushige Shiode. 2022. Network-Based Space-Time Scan Statistics for Detecting Micro-Scale Hotspots. *Sustainability* 14, 24 (Jan. 2022), 16902. https://doi.org/10.3390/su142416902 Number: 24 Publisher: Multidisciplinary Digital Publishing Institute.

[11] Shaowen Wang. 2010. A CyberGIS framework for the synthesis of cyberinfrastructure, GIS, and spatial analysis. *Annals of the Association of American Geographers* 100, 3 (2010), 535–557.

[12] Shaowen Wang, Mary Kathryn Cowles, and Marc P. Armstrong. 2008. Grid computing of spatial statistics: using the TeraGrid for G(d) analysis. *Concurrency and Computation: Practice and Experience* 20, 14 (2008), 1697–1720. https://doi.org/10.1002/cpe.1294 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.1294

[13] David Weisburd. 2015. The law of crime concentration and the criminology of place. *Criminology* 53, 2 (2015), 133–157.

[14] Dandong Yin, Yan Liu, Anand Padmanabhan, Jeff Terstriep, Johnathan Rush, and Shaowen Wang. 2017. A CyberGIS-Jupyter framework for geospatial analytics at scale. In *Proceedings of the practice and experience in advanced research computing 2017 on sustainability, success and impact.* 1–8.