Journal of Electronic Research and Application, 2023, Volume 7, Issue 6
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Manifold Structure Analysis of Tactical Network Traffic Matrix Based on Maximum Variance Unfolding Algorithm

Hao Shi*, Guofeng Wang, Rouxi Wang, Jinshan Yang, Kaishuan Shang

China Academy of Electronics and Information Technology, Beijing 100086, China

Corresponding author: Hao Shi, haoshi@mail.ustc.edu.cn

Abstract: As modern weapons and equipment undergo increasing levels of informatization, intelligence, and networking, the topology and traffic characteristics of battlefield data networks built with tactical data links are becoming progressively complex. In this paper, we employ a traffic matrix to model the tactical data link network. We propose a method that utilizes the Maximum Variance Unfolding (MVU) algorithm to conduct nonlinear dimensionality reduction analysis on high-dimensional open network traffic matrix datasets. This approach introduces novel ideas and methods for future applications, including traffic prediction and anomaly analysis in real battlefield network environments.

Keywords: Manifold learning; Maximum Variance Unfolding (MVU) algorithm; Nonlinear dimensionality reduction

Online publication: November 29, 2023

1. Introduction

Since the Gulf War, the nature of warfare has transitioned from traditional large-scale mechanized conflicts to localized warfare under high-tech conditions. With the rapid advancement of technologies such as computers, network information, and artificial intelligence, modern warfare has become increasingly informatized, networked, and intelligent. The shift in the form of modern warfare from platform-centered to network-centered is a significant change. Unlike platform-centered warfare focused on weapon platforms, network-centered warfare is a networked and information-based integrated form centered on the network. It stands as the fundamental and crucial combat style for local wars under information technology conditions, integrating intelligence, command, communication, computers, electronic warfare, information warfare, combat support, and firepower strikes. This approach constructs a network-centric battlefield environment, connecting various combat units, weapon platforms, and information systems in the battlefield through the network. It achieves continuous improvement in intelligence sharing, enhances battlefield situational awareness, accelerates combat decision-making and commanding, and accomplishes almost real-time combat coordination for tactical purposes. The data link network facilitates space-time cooperative operations among various combat units in systematic and integrative warfare and serves as the infrastructure for constructing a network-centric battlefield

environment. It also acts as the neural center of the network-centric warfare system.

With the rapid development of military science and technology, the level of informatization and intelligence of modern weapons and equipment is rapidly advancing. The number and types of equipment that can be connected to the tactical data link network will increase exponentially, leading to a more complex topology and data traffic characteristics of the network. Data link network traffic acts as the carrier of information, flowing like the "blood" in the large-scale and complex battlefield network environment. In the face of this increasingly complex network environment, understanding how to process and analyze battlefield network communication traffic data, extract traffic characteristics, perceive the network situation, detect anomalies, and predict network traffic is a crucial prerequisite for building a more robust and real-time battlefield network.

Currently, most analyses and research on network traffic data are primarily conducted on a single link in isolation [1-5]. However, the emergence of the traffic matrix provides an opportunity to analyze network traffic data characteristics from the perspective of the entire network [6]. Research revealed that data traffic on different links in a network is not independent; instead, it often exhibits similar traffic characteristics [6]. Lakhina *et al.* employed a linear analysis method based on the Principal Component Analysis (PCA) algorithm [7]. As the complexity of the network structure increases, data traffic may exhibit more complex nonlinear characteristics. Therefore, this article utilizes a nonlinear dimensionality reduction algorithm to conduct dimensionality reduction analysis of complex battlefield networks, providing a new idea and method for network situation awareness and data analysis in future systematic operations involving complex battlefield networks.

2. Basic theory

This section primarily introduces the mathematical theory related to manifold learning.

(1) Topology: Let τ be a subset family of the non-empty set X. If τ satisfies the following constraints:

- $X, \emptyset \in \tau$;
- if $A,B \in \tau$, then $A \cap B \in \tau$;
- if $\tau_1 \in \tau$, then $\cup_{A \in \tau 1} A \in \tau$;

then is called a topology of X.

(2) Topology space: If τ is a topology of the set X, then the pair (X, τ) is called a topological space.

(3) Homeomorphism: Let X and Y be two topological spaces. If $f:X \rightarrow Y$ is one-to-one mapping, f and $f^{-1}:Y \rightarrow X$ both are continuous, then f is called homeomorphic mapping or homeomorphism.

(4) Hausdoff space: Let (X, τ) be topology space, $\forall x, y \in X, x \neq y, \exists U_x$ and U_y, s.t. $U_x \cap U_y = \emptyset$, then (X, τ) is a Hausdoff space, where U_x and U_y are the neighborhoods of x and y.

(5) Manifold: Let X be a Hausdoff space, $\forall x \in X$, there exists an open set neighborhood U which is homeomorphic to the Euclidean space $\mathbb{R}^D$, then X is a D-dimensional topological manifold, referred to as a D-dimensional manifold.

(6) Manifold learning: For data set $X=\{x_1,x_2,\cdots,x_N\} \subset \mathbb{R}^D$, assume that any point in X can be generated by $Y=\{y_1,y_2,\cdots,y_N\} \subset \mathbb{R}^d$ through a nonlinear mapping. The goals of manifold learning are:

- To get $Y=\{y_1,y_2,\cdots,y_N\} \subset \mathbb{R}^d$, which is the low-dimensional coordinates of X;
- To get $f^{-1}:\mathbb{R}^D \rightarrow \mathbb{R}^d$, which is a nonlinear mapping from the high-dimensional input space to the low-dimensional output space.

3. Maximum variance unfolding algorithm

3.1. Brief introduction

Manifold learning assumes that high-dimensional input data are approximately situated on a low-dimensional manifold embedded in the high-dimensional space. The global feature preservation method calculates low-dimensional coordinates by performing eigendecomposition on the similarity inner product matrix, constructed from the global similarity matrix, aiming to retain the global geometric features of the output data in the low-dimensional space. Typically, this algorithm involves three steps: (1) constructing a neighborhood graph from the input data; (2) building an inner product matrix based on the global similarity measure matrix; (3) conducting eigenvalue decomposition of the inner product matrix to derive the low-dimensional embedding coordinates of the input data.

Maximum variance unfolding (MVU) is a manifold learning algorithm proposed under local isometric constraints [8-11]. Its fundamental concept is to "expand" non-adjacent data as far apart from each other as possible while maintaining the distance between neighboring points on the neighborhood graph unchanged. For instance, envision a string of curly necklaces as a one-dimensional manifold embedded in a two-dimensional space. Each data point in the high-dimensional space represents a node on the necklace. The MVU idea is akin to unfolding the "necklace" in the low-dimensional space, illustrated in **Figure 1**. Theoretically, the process of "unfolding" high-dimensional data using MVU can be formulated as a quadratic programming problem, where $max \sum_{ij} \|y_i - y_j\|^2$ which satisfies the following constraints

(1) $\|y_i - y_j\|^2 = \|x_i - x_j\|^2$ where x_i and y_j are neighbors of each other

(2) $\sum_i y_i = 0$

where constraint (1) is a local isometric constraint, ensuring that the Euclidean distance between neighboring points remains unchanged after dimensionality reduction, while constraint (2) is used to eliminate the centralization constraint of the translational degree of freedom [12,13]. By defining the Gram inner product matrix $K_{ij} = y_i y_j$, the above non-convex optimization quadratic programming problem can be transformed into a convex optimization semidefinite programming (SDP) problem, where $maxtr(K)$ satisfies

(1) $K_{ii} - 2K_{ij} + K_{jj} = \|x_i - x_j\|^2$, where x_i and y_j are neighbors of each other

(2) $\sum_{ij} K_{ij} = 0$

(3) $K \geq 0$

where (3) is a positive semi-definite constraint, which ensures that this SDP has the optimal solution. By solving this SDP problem, the Gram matrix K can be obtained, and the d-dimensional embedding coordinates can be represented by the eigenvectors corresponding to the d largest eigenvalues of K.

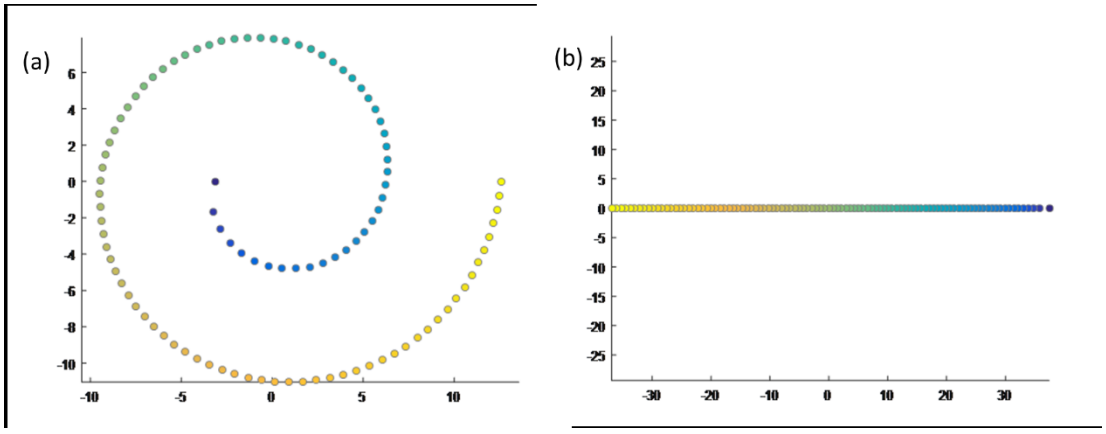

Figure 1. (a) The two-dimensional "necklace" data; **(b)** MVU embedding result from "necklace" data

3.2. MVU algorithm process

(1) Neighborhood graph construction by using the k-NN method or ε-ball method.

(2) SDP: The objective function is constructed under local isometric constraints and centralization constraints, and converted into a SDP problem. Then the SDP problem is solved to obtain the positive semi-definite Gram matrix .

(3) Spectral decomposition: Perform eigen-decomposition of the Gram matrix to obtain the low-dimensional coordinates.

The embedding results of MVU on public data sets are shown in **Figure 2**.

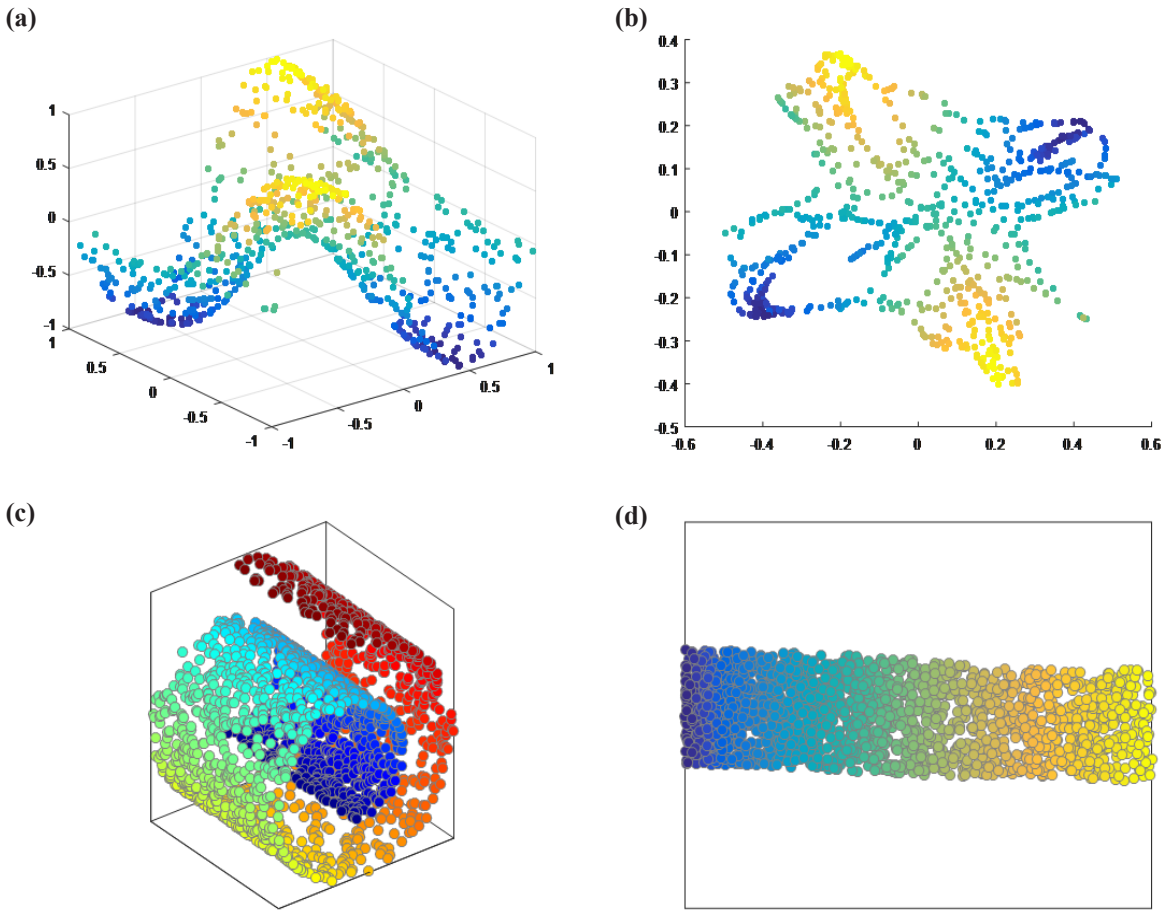

Figure 2. (a) TwinPeak data set; **(b)** TwinPeak data set embedding result by MVU; **(c)** SwissRoll data set; **(d)** SwissRoll data set embedding result by MVU

3.3. Algorithm analysis

MVU is a global manifold learning algorithm based on isometry. If there exists a subset of the Euclidean space equidistant from the manifold where the high-dimensional input data is situated, MVU can accurately restore the low-dimensionality of the high-dimensional input data. Additionally, because MVU does not require the calculation of geodesic distance between input data, it can yield genuine dimensionality reduction results even for non-convex data sets. However, the MVU algorithm has notable drawbacks. Firstly, the time complexity and space complexity of MVU for solving the SDP problem are both $O((kN)^3)$, and the time complexity of eigenvalue decomposition of the Gram matrix during the solution of low-dimensional embedded coordinates is $O(N^3)$. Consequently, the substantial computational complexity significantly hampers the application of the

MVU algorithm on large-scale data sets. Secondly, due to the stringent local isometric constraints, the MVU algorithm may perform poorly due to "short-circuit edges" resulting from noise data during the construction of the neighborhood graph.

4. Dimensionality reduction analysis of tactical network traffic data

4.1. Network traffic matrix modeling

Let Ω represent the non-empty set of all nodes in the network, with $|\Omega|=N$. The flow matrix can be naturally represented by a three-dimensional non-negative hypermatrix $X(t)$, where each element is denoted as $X_{i,j}(t)$. Each element in the traffic matrix signifies the traffic measurement value from the source node i to the destination node j within the time period $(t, t+\Delta t) \subset T$, covering the entire measurement period. Real-time measurement of the traffic matrix size is challenging, so the algebraic mean of the traffic at a discrete time interval Δt is typically used as the measurement value for that period. For different combat missions, the time interval can be chosen based on the specific circumstances. In a decentralized battlefield network consisting of N combat units, the traffic matrix dimension obtained in any observation period is N^2. Consequently, the dimension of traffic matrix data acquired in one measurement period will be 225 in a network comprising 15 nodes. Directly analyzing such a high-dimensional traffic matrix poses challenges in terms of computational and storage complexity. Therefore, the MVU algorithm is employed for dimensionality reduction on the high-dimensional network traffic matrix data to facilitate storage and analysis.

In this paper, we utilize public traffic matrix datasets to simulate real battlefield networks, sourced from http://www.cs.utexas.edu/~yzhang/research/AbileneTM. This dataset collection spans 6 months of Internet traffic matrix data from the Abilene backbone network, comprising a total of 24 data files, with each data file containing 2016 traffic matrices. The network consists of 12 PoP points, making each traffic matrix 144-dimensional, with a sampling interval of 5 minutes and data units measured in bytes.

4.2. Intrinsic dimensionality and residual variance analysis

Determining whether the traffic matrix possesses low-dimensional features is a prerequisite for subsequent dimensionality reduction analysis. Lakhina *et al*. observed that each data flow in the traffic matrix can be expressed as the weighted sum of a small number of eigenvalue flows [7]. As depicted in **Figure 3**, a significant portion of the traffic variance is determined by the first few (5 to 10) eigenvalue flows, indicating that the dimensions of these high-dimensional traffic matrices are considerably lower than the number of PoP pairs in the network.

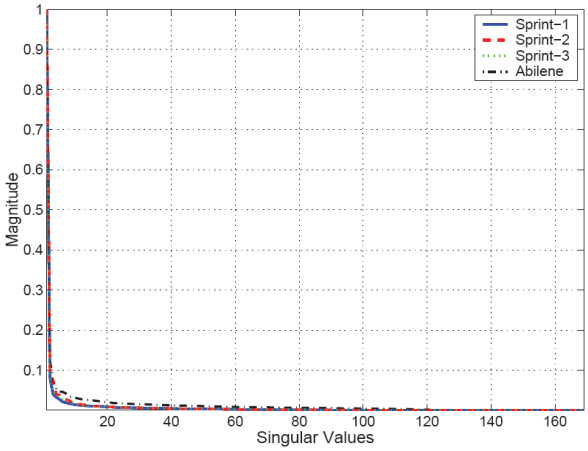

Figure 3. Dimensionality analysis on traffic flow data by PCA

Tenenbaum *et al*. employed the residual variance method to analyze the intrinsic dimensionality of high-dimensional input data [12]. We applied MVU to the public traffic data set, and the results of the residual variance results are presented in **Figure 4**. The intrinsic dimensionality can be identified by identifying the "elbow" point where the curve ceases to significantly decrease with added dimensions. As demonstrated in **Figure 4**, the "elbow" points of the residual variance curves for all four data sets occur at d = 5, signifying that their intrinsic dimensions are much smaller than their original dimensions.

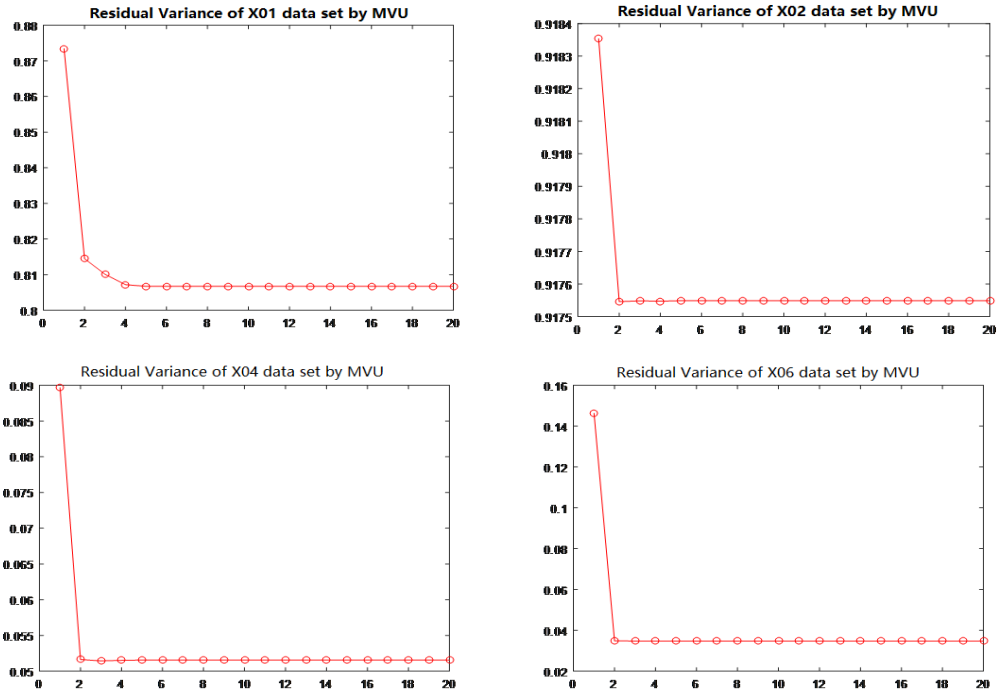

Figure 4. Residual variance results of the open traffic datasets by MVU

The low-dimensional nature of the network traffic matrix is attributed to its spatial correlation. The network comprises edge networks and core networks, and the traffic of different nodes in the core network may originate from the same edge network. This similarity in variation patterns among the different nodes results in a more condensed low-dimensional representation of the high-dimensional traffic matrix.

4.3. Low-dimensional structure analysis

To gain a more intuitive understanding of the low-dimensional structures within the traffic matrix, this section delves deeper into the analysis of the MVU three-dimensional embedding results of the traffic matrix data.

For visual clarity, we present the three-dimensional embedding result of datasets X01, X06, X11, and X18 in **Figure 5**. These four datasets exhibit diverse structures in the low-dimensional embedding space, illustrating relationships among global network-wide traffic during sampling time intervals. The embedding result of dataset X11 demonstrates nearly linear characteristics, suggesting a linear variation pattern and features within the corresponding network traffic. In the results of datasets X01 and X18, some isolated points are noticeable, possibly linked to abnormal situations in the network. Through this analysis, it becomes evident that various network traffic matrix datasets showcase distinct structural characteristics in their low-dimensional embedding space.

By utilizing manifold learning methods to reduce the dimensionality of high-dimensional traffic matrix data and analyze low-dimensional structures, valuable insights can be gleaned. This approach enables the

extraction of information such as traffic trends or anomalies, providing a novel method for analyzing internet traffic from a network-wide perspective.

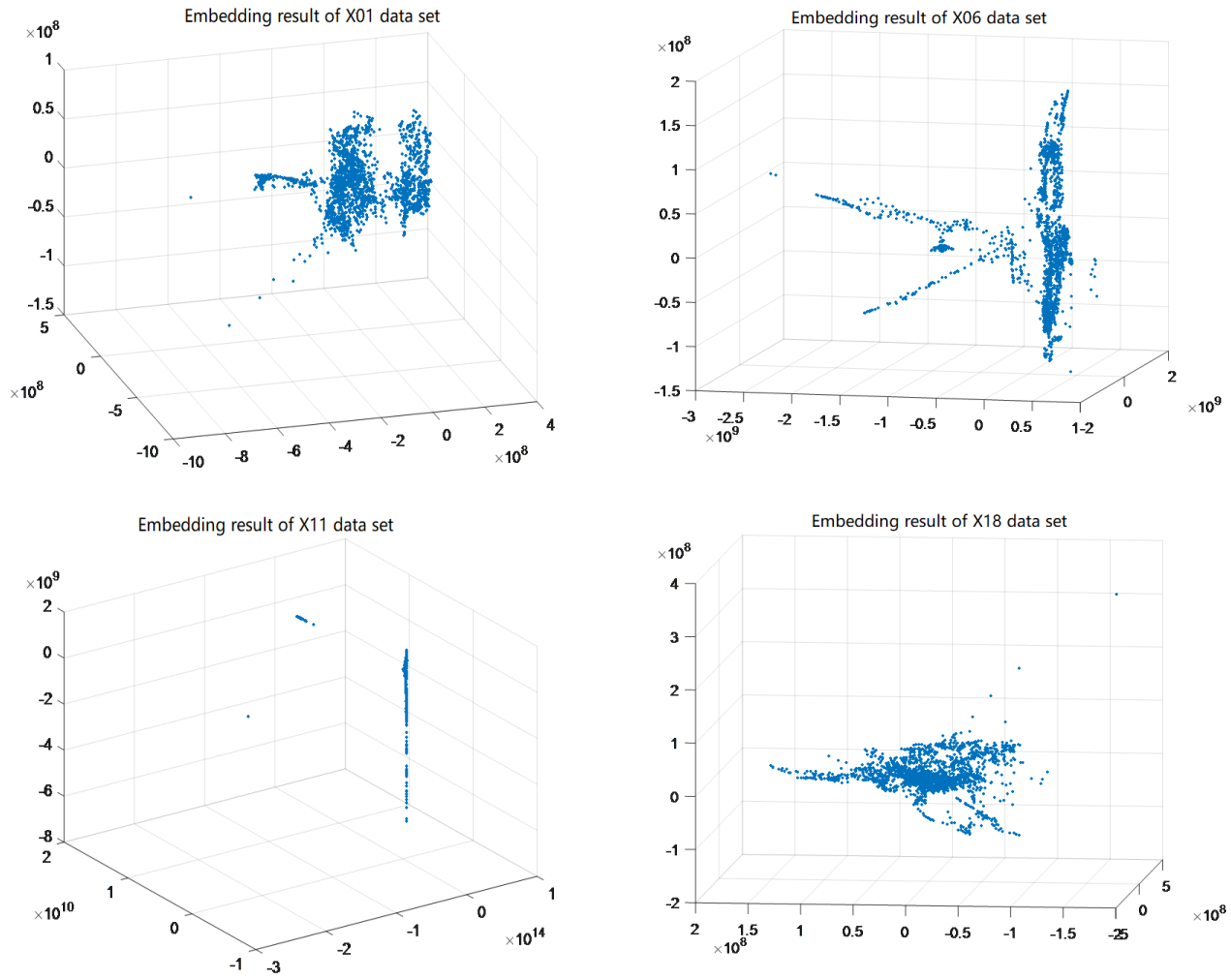

Figure 5. MVU embedding result of different traffic datasets

5. Conclusion

In this paper, we applied the MVU algorithm to open datasets to analyze the low-dimensional structure of network traffic. The experimental results demonstrate the validity of our approach. However, the MVU algorithm still faces challenges due to its high complexity and weak robustness, thereby restricting its application to large-scale, noisy datasets. In future work, our emphasis will be on enhancing the computational efficiency and robustness of the MVU algorithm.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Papagiannaki N, Taft N, Zhang Z, et al., 2003, 22nd Annual Joint Conference of the IEEE Computer and

Communications Societies, March 30–April 3, 2003: Long-Term Forecasting of Internet Backbone Traffic: Observations and Initial Models. IEEE INFOCOM 2003, San Francisco, vol 2, 1178–1188. https://doi.org/10.1109/INFCOM.2003.1208954

[2] Zhang Y, Roughan M, Lund C, et al., 2003, SIGCOMM '03: Applications, Technologies, Architectures, and Protocols for Computer Communications, August 25–29, 2003: An Information-Theoretic Approach to Traffic Matrix Estimation. ACM SIGCOMM, Karlsruhe, 301–312. https://doi.org/10.1145/863955.863990

[3] Soule A, Nucci A, Cruz R, et al., 2004, How to Identify and Estimate the Largest Traffic Matrix Elements in a Dynamic Environment. ACM SIGMETRICS Performance Evaluation Review, 32(1): 73–84. https://doi.org/10.1145/1012888.1005698

[4] Medina A, Taft N, Salamatian K, et al., 2002, Traffic Matrix Estimation: Existing Techniques and New Directions. ACM SIGCOMM Computer Communication Review, 32(4): 161–174. https://doi.org/10.1145/964725.633041

[5] Crovella M, Kolaczyk E, 2003, 22nd Annual Joint Conference of the IEEE Computer and Communications Societies, March 30–April 3, 2003: Graph Wavelets for Spatial Traffic Analysis. IEEE INFOCOM 2003, San Francisco, vol 3, 1848–1857. https://doi.org/10.1109/INFCOM.2003.1209207

[6] Tune P, Roughan M, 2013, Internet Traffic Matrices: A Primer, in Recent Advances in Networking, ACM SIGCOMM eBook, vol 1, 108–163. https://sigcomm.org/education/ebook/SIGCOMMeBook2013v1_chapter3.pdf

[7] Lakhina A, Papagiannaki K, Crovella M, et al., 2004, SIGMETRICS '04/Performance '04: Proceedings of the Joint International Conference on Measurement and Modeling of Computer Systems, June 10–14, 2004: Structural Analysis of Network Traffic Flows. ACM SIGMETRICS, New York, 61–72. https://doi.org/10.1145/1005686.1005697

[8] Weinberger KQ, Packer B, Saul LK, 2005, Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics, January 6–8, 2005: Nonlinear Dimensionality Reduction by Semidefinite Programming and Kernel Matrix Factorization. Society for Artificial Intelligence and Statistics, Barbados, 381–388.

[9] Weinberger KQ, Saul LK, 2004, Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 27–July 2, 2004: Unsupervised Learning of Image Manifolds by Semidefinite Programming. IEEE-CS\DATC, Washington, vol 2, II-988–II-995. https://doi.org/10.1109/CVPR.2004.1315272

[10] Weinberger KQ, Saul LK, 2006, AAAI '06: Proceedings of the 21st National Conference on Artificial Intelligence, July 16–20, 2006: An Introduction to Nonlinear Dimensionality Reduction by Maximum Variance Unfolding. AAAI, Boston, vol 2, 1683–1686.

[11] Weinberger KQ, Sha F, Saul LK, 2004, Proceedings of the 21st International Conference on Machine Learning, July 4–8, 2004: Learning a Kernel Matrix for Nonlinear Dimensionality Reduction. ACM, Banff, 839–846.

[12] Tenenbaum JB, de Silva V, Langford LC, 2000, A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science, 290(5500): 2319–2323. https://doi.org/10.1126/science.290.5500.2319