Topological clustering of multilayer networks

Monisha Yuvaraj[a,1], Asim K. Dey[a,b,1], Vyacheslav Lyubchich[c,1], Yulia R. Gel[a,1], and H. Vincent Poor[b,1,2]

[a]Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX 75080; [b]Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544; and [c]Chesapeake Biological Laboratory, University of Maryland Center for Environmental Science, Solomons, MD 20688

Contributed by H. Vincent Poor, February 9, 2021 (sent for review September 24, 2020; reviewed by David L. Banks, Hiba Baroud, and Amy Braverman)

Multilayer networks continue to gain significant attention in many areas of study, particularly due to their high utility in modeling interdependent systems such as critical infrastructures, human brain connectome, and socioenvironmental ecosystems. However, clustering of multilayer networks, especially using the information on higher-order interactions of the system entities, still remains in its infancy. In turn, higher-order connectivity is often the key in such multilayer network applications as developing optimal partitioning of critical infrastructures in order to isolate unhealthy system components under cyber-physical threats and simultaneous identification of multiple brain regions affected by trauma or mental illness. In this paper, we introduce the concepts of topological data analysis to studies of complex multilayer networks and propose a topological approach for network clustering. The key rationale is to group nodes based not on pairwise connectivity patterns or relationships between observations recorded at two individual nodes but based on how similar in shape their local neighborhoods are at various resolution scales. Since shapes of local node neighborhoods are quantified using a topological summary in terms of persistence diagrams, we refer to the approach as clustering using persistence diagrams (CPD). CPD systematically accounts for the important heterogeneous higher-order properties of node interactions within and in-between network layers and integrates information from the node neighbors. We illustrate the utility of CPD by applying it to an emerging problem of societal importance: vulnerability zoning of residential properties to weather- and climate-induced risks in the context of house insurance claim dynamics.

multilayer network | clustering | topological data analysis | persistence diagram | insurance risk

Many modern human-made systems, e.g. critical infrastructures integrating operations of vital societal physical and cyber services such as power systems, telecommunication, and transportation, as well as a broad range of natural phenomena from human brain functionality to socioenvironmental ecosystems and virus–host interactomes, exhibit a sophisticated, highly interdependent structure (1–7). Modeling such interdependency can be addressed with multilayer graphs, resulting in a recent surge of interest in the interdisciplinary analysis of complex multilayer networks. A multilayer network accounts for relationships among multiple layers of connectivity (i.e., networks), where each layer represents a system or subsystem. Dictated by emerging applications in security and resilience of critical infrastructures to natural hazards, terrorist activities, and cyber threats (8–14), one of the primary goals of such studies nowadays is to better understand which segments of the multilayer network are most vulnerable to a particular hazard and to develop proactive strategies for optimal partitioning, thereby isolating unhealthy components and mitigating the risk of further failure propagation (15–17).

Similar to the case of unilayer networks, the objective of optimal partitioning, or clustering, of multilayer networks is to unveil meaningful patterns of node groupings and to divide nodes into communities, by accounting for the different interaction types nodes can be involved both within and in-between layers of the considered multilayer graph. Still, contrary to unilayer networks, clustering of multilayer graphs remains a substantially less-developed area (18–20), and most currently existing methods are based on an adaptation of conventional clustering approaches for unilayer networks such as stochastic block models (21–24) and layer aggregation in the spectral domain (25–28) to the multilayer case. However, the clustering of multilayer graphs poses a number of new, specific research challenges. First, partitioning of multilayer graphs requires accounting for both the important relationships between nodes in the same layer and interactions among nodes in different layers. Second, such layers, as, for instance, in the case of critical infrastructures formed by transportation and power grid networks, may exhibit disparate local and global structural properties, making application of clustering methods originating in the unilayer network analysis and based on an aggregation of the layer information infeasible. Finally, higher-order network structures, in the context of both unilayer and multilayer graphs, are often shown to exhibit stronger signals of community existence than lower-order pairwise connectivity patterns which are assessed at the level of individual nodes and edges (29, 30). This phenomenon becomes particularly important in conjunction with resilience analysis of highly interdependent systems such as critical infrastructures when the focus is on evaluating how multiple interconnected entities of the systems, for example electric power substations, transportation hubs, and telecommunication towers, jointly respond to natural disasters and cyber attacks. Nevertheless, clustering of multilayer networks while accounting for higher-order connectivity structures remains in its infancy.

Significance

Multilayer network clustering is used in such diverse areas as optimal islanding of critical infrastructures, analysis of trade agreements, and monitoring ecological interaction patterns. We propose a perspective on multilayer network clustering based on the concept of shape. By invoking the machinery of topological data analysis, we first study a shape of each node neighborhood and then group nodes based on how similar shapes of their local neighborhoods are. The significance of this methodology can be viewed through an emerging problem of sustainability of house insurance to climate risks. The topological perspective opens possibilities for more systematic, robust, and mathematically rigorous integration of higher-order network properties and their interplay to the analysis of complex networks.

Author contributions: M.Y., A.K.D., V.L., Y.R.G., and H.V.P. designed research; M.Y., A.K.D., V.L., Y.R.G., and H.V.P. performed research; M.Y., A.K.D., V.L., Y.R.G., and H.V.P. contributed new reagents/analytic tools; M.Y., A.K.D., V.L., Y.R.G., and H.V.P. analyzed data; and M.Y., A.K.D., V.L., and Y.R.G. wrote the paper.

Reviewers: D.L.B., Duke University; H.B., Vanderbilt University; and A.B., Jet Propulsion Laboratory.

The authors declare no competing interest.

[1]M.Y., A.K.D., V.L., Y.R.G., and H.V.P. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: poor@princeton.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2019994118/-/DCSupplemental.

Published May 18, 2021.

To address these challenges, we introduce the concepts of topological data analysis (TDA) to studies of complex multilayer networks and propose a topological approach to network clustering. TDA is an emerging methodology at the interface of algebraic topology and data science (31–34) offering a mathematically rigorous machinery for analysis of data shape. In particular, TDA allows one to glean a deeper insight into hidden mechanisms behind the data-generating process by analyzing both topological and geometric properties of the observed data through multiple-resolution lenses. While TDA has been proven to deliver high utility in a very diverse set of applications, from cancer gene expression to financial fraud to ichthyology (35–38), TDA concepts have not yet propagated into clustering analysis of complex networks. The key idea behind our topological network clustering is to group nodes based on how similar in shape their local neighborhoods are. In particular, the proposed topological approach is based on the comparison of local topology and geometry around each node using persistence diagrams and, hence, is termed "clustering using persistence diagrams" (CPD). The topological CPD approach to network clustering allows both for systematic accounting of heterogeneous higher-order properties of within and in-between network layers and for integrating the important information from the node neighbors and their interactions. In contrast to earlier (not network-focused) TDA-based clustering approaches such as Mapper (39) and ToMATo (40), which both act in conjunction with some additional clustering algorithms, CPD is a stand-alone clustering approach and does not require a filter function, which is characteristic of Mapper. Furthermore, compared to clustering using Betti numbers, a TDA-based clustering algorithm for spatiotemporal data (41), CPD simultaneously accounts for multiple topological summaries and their interdependencies and, as a result, shows more stable performance, especially in application to sparse heterogeneous graph-structured data. Finally, the area of the CPD applicability is well beyond complex networks and also includes multivariate point clouds and sets of functions.

We illustrate the application of our CPD algorithm and the utility of topological concepts for clustering of complex networks in application to a multilayer climate-insurance network. The insurance industry currently experiences major challenges due to the impact of climate dynamics expressed in the rising frequency and intensity of adverse weather events, including the so-called low-individual but high-cumulative-impact events such as higher-than-normal precipitation and stronger-than-usual wind speeds. For example, ref. 42 shows that 38% of insurance companies view climate risk as a core business issue, with implications for governance, strategy, risk management, and operations, while 29% of the companies consider climate risk as a sustainability issue which is evolving to a core business issue. In turn, often-neglected low-individual but high-cumulative-impact adverse weather events, coupled with aging critical city infrastructures, increasingly lead to accidents of various scales and property depreciation. One of the first tasks toward better assessment of climate risks and development of more efficient mitigation strategies is the identification of areas that show higher vulnerability not only due to the magnitude of climate trends but also due to economic and sociodemographic patterns. However, climate variables, insured property characteristics, and associated insurance claim dynamics tend to exhibit complex dependence structures that are often nonlinear and nonstationary in space and time. As a result, similarity measures based on Euclidean distances and conventional geographic proximity might not be appropriate metrics for optimal partitioning of such data. As shown by refs. 43–48, such a sophisticated dependence structure in climate variables can be addressed with complex networks. However, no analysis has been done to capture the multivariate spatiotemporal dependency for classifying the insurance risk exposure and informing the risk mitigation strategies. We address this knowledge gap by introducing a multilayer complex network based on climate and home insurance variables and by developing vulnerability zoning based on the topological CPD approach. The proposed peril map based on shape similarities in environmental and sociodemographic characteristics allows for more accurate modeling of climate risk than conventional tools based on simple geographic proximity.

Multilayer Networks

Consider a single-layer network modeled by a graph $G = (V, E, \omega)$, where V is the set of nodes and $E \subset V \times V$ is the set of edges. The total number of nodes in G is $n = |V|$. Here $\omega: V \times V \mapsto \mathbb{R}$ is an edge weight function such that each edge $e_{uv} \in E$ has a weight ω_{uv}. The adjacency matrix A is symmetric, i.e., $A_{ij} = A_{ji}$.

A multilayer network can be modeled by a multilayer graph, $\mathcal{G}$, consisting of m nonoverlapping layers, where each layer is modeled with a weighted graph G_i with an associated adjacency matrix A_i, $i = 1, \dots, m$. The elements of the set $\mathcal{A} = \{A_1, A_2, \dots, A_m\}$ are referred to as the within-layer matrices, representing connections along a single layer, known as intralayer links.

To model dependencies between two graphs, G_k and G_l with their adjacency matrices, A_k and A_l, respectively ($k, l = 1, 2, \dots, m$; $k \neq l$), we consider one-to-one symmetric interconnectivity of nodes in the corresponding graphs. As a result, we obtain a set of cross-layer adjacency matrices $\mathcal{D}_p = \{A_{l,k}, k \neq l\}$ that specifies the edges between nodes in different layers, where p is the number of dependencies. That is, a multilayer network, $\mathcal{G}$, has a set $E_I(\mathcal{G})$ of interlayer links that connect nodes across layers, i.e., for each edge $(u, v) \in E_I(\mathcal{G})$ we have $u \in V(G_k)$ and $v \in V(G_l)$ for $k \neq l$ (49, 50). The supra-adjacency matrix of the multilayer network $\mathcal{G}$ is defined as a block-matrix structure:

$$\mathfrak{A} = \left(\begin{array}{c} A_1 \ \dots \ A_{1k} \cdots A_{1m} \\ \hline \vdots \quad \ddots \quad \vdots \quad \ddots \quad \vdots \\ \hline A_{l1} \cdots A_{k=l} \cdots A_{lm} \\ \hline \vdots \quad \ddots \quad \vdots \quad \ddots \quad \vdots \\ \hline A_{m1} \cdots A_{mk} \cdots A_m \end{array}\right).$$

The diagonal elements corresponding to the set $\mathcal{A}$ are within-layer matrices. Off-diagonal matrices A_{lk} for $k, l = 1, 2, \dots, m$; $k \neq l$ represent interlayer links that connect nodes in layer G_k to nodes in layer G_l (2, 51–53).

Similarity-Based Networks. Edges and edge weights in a multilayer network can be defined using various application-tailored relationships and measures. In cases when there exists no application-driven notion of edges, e.g. as flight routes in air transportation networks or transmission lines in power grid networks, edges are typically constructed based on some measure ω of similarity between nodes, resulting in a so-called similarity-based network. One of the most widely used similarity measures is a correlation coefficient, with applications of correlation-based networks ranging from finance (54–56) to brain sciences (57–59) to climatology (60–62). In this study, we follow a similar route for our specific case study of climate-insurance multilayer networks and construct edges in $\mathcal{G}$, based on the maximum correlation achieved upon nonlinear nonparametric transformations of observed variables.

In particular, let X and Y be two time series representing the same variable observed at different nodes (when defining $\mathcal{A}$) or different variables recorded at the same node (when defining $\mathcal{D}$). (For instance, in the context of our case study X and Y may represent precipitation levels at two locations or precipitation and insurance claim records reported at the same location.) Our goal is to find a nonlinear transformation of X and Y such that the correlation between the transformed variables is maximized. This step is performed using alternating conditional expectations (ACE), which is an algorithm to find the best-fitting additive model resulting in the maximum linear effect between the transformed response and predictors:

$$\omega = r^*(X, Y) = \max_{\phi,\theta} r[\phi(X), \theta(Y)], \qquad [1]$$

where r^* is the maximal correlation between the optimal transformations $\phi(X)$ and $\theta(Y)$ of X and Y, respectively. To find such transformations, the errors $e^2(\theta, \phi) = \mathrm{E}[\theta(Y) - \phi(X)]^2$ are alternatingly minimized first with respect to $\theta(Y)$ (keeping $\mathrm{E}(\theta^2) = 1$), then with respect to $\phi(X)$ for a given $\theta(Y)$. The solutions can be written as

$$\theta(Y) = \frac{\mathrm{E}[\phi(X)|Y]}{\|\mathrm{E}[\phi(X)|Y]\|}; \quad \phi(X) = \mathrm{E}[\theta(Y)|X], \qquad [2]$$

where $\|\cdot\| = \sqrt{\mathrm{E}(\cdot)^2}$.

The minimization process begins with an initial guess for one of the functions ($\theta(Y) = Y/\|Y\|$). Each iteration performs one pair of the single-function minimizations using Eq. **2**, until a complete iteration pass fails to decrease e^2. The algorithm converges to the optimal transformations θ and ϕ (63).

Node Embedding. Extraction of meaningful information from complex networks is computationally and memory-intensive. Node embedding provides a framework to combat both these issues by transforming the network into a low-dimensional space while preserving structural information. The variety of methods for node embedding can be divided into two categories: matrix factorization methods and random walk methods (see refs. 64 and 65 and references therein). The former enjoy a strong theoretical backing as the approach relies on matrix factorization techniques with tractable optimization functions that converge. Here we use multilayered network embedding (MANE), which is an extended form of matrix factorization for multilayer networks.

To describe MANE, let $F_i \in \mathbb{R}^{n_i \times d_i}$ for all nodes in the i-th layer ($i = 1, \ldots, m$), where n_i is number of nodes in the i-th layer and d_i is the embedding dimension. The objective of the algorithm is to find a low-dimensional vector representation that retains the node proximity in the topological structure of the network. The objective function is

$$\max_{F_i} tr(F_i^\top L_i F_i) + \alpha \sum_{j=1}^{m} tr(F_i^\top D_{ij} F_j F_j^\top D_{ij}^\top F_i), \qquad [3]$$

where D_{ij} denotes network dependency between layers i and j; L_i is the normalized Laplacian matrix, and the embedding representation F_i is a matrix such that $F_i^\top F_i = I$ $(\forall i = 1, \ldots, m)$ (66). The embedding is essentially obtained by concatenating the top d_i eigenvectors of $L_i + \alpha \sum_{j=1}^{m} D_{ij} F_j F_j^\top D_{ij}^\top$. The first term in Eq. **3** corresponds to embedding of a single-layer network to a low-dimensional representation which aims to preserve the node proximity in the original single-layer structure. The second term in Eq. **3** corresponds to embedding cross-layer connectivity (i.e., dependency across single-layer networks). Here the idea is to use interplay among node latent features in different layers as an approximation to real dependencies.

An advantage of using a matrix factorization technique is the reduced number of tuning parameters. Parameter tuning in most of these methods most often involves random walk procedures that require a selection of the walk length, number of random walks, etc. The parameters are usually tuned on a labeled training dataset.

Background on Topological Data Analysis

To infer meaningful and actionable inferences from the embeddings, clustering algorithms can be employed. Forming clusters based on dynamics of shape helps in discovering persistent clusters of nodes that follow similar patterns. TDA is useful here, due to its inherent reliance on similarity graphs.

Consider an (edge)-weighted graph G. If we select a certain threshold (or scale) $\epsilon_j > 0$ and keep only edges with weights $\omega_{uv} \leq \epsilon_j$, we obtain a graph G_j with an associated adjacency matrix $A_{uv} = \mathbb{1}_{\omega_{uv} \leq \epsilon_j}$. Now, changing the threshold values $\epsilon_1 < \epsilon_2 < \ldots < \epsilon_n$ results in a hierarchical nested sequence of graphs $G_1 \subseteq G_2 \subseteq \ldots \subseteq G_n$ that is called a "network filtration." One of the widely used simplicial complexes is the Vietoris–Rips (VR) complex. The VR complex at threshold ν_j is defined as $VR_j = \{\sigma \subset V | \omega_{uv} \leq \nu_j \text{ for all } u, v \in \sigma\}$.

Armed with the filtration, we assess changes in topological summaries of the network to detect long-lived (or persistent) features over a wide range of thresholds ϵ_j. The objective is to detect features which are long-lived (or persistent) over varying thresholds ϵ (32, 67, 68). Such persistent features are likely to characterize the intrinsic system organization.

The lifespans of topological features under VR filtration can be represented with a barcode plot where each bar depicts the lifespan of each topological feature. Births and deaths of topological features under VR filtration are also visualized with a persistence diagram (PD), where each topological feature is denoted by a point with (x, y) coordinates corresponding to the birth and death times, respectively. Hence, features with longer lifespans, i.e., stronger persistence, are those points that are far from the main diagonal and are considered as topological signals. For a more detailed description see *SI Appendix*, section 1. PD captures the geometry and topology of the data and hence can be used in different learning tasks. Next, we introduce a clustering algorithm based on the PD of the data.

CPD

We propose a method for clustering of multilayer networks which is motivated by the following two overarching queries. First, in contrast to supervised community detection and classification, unsupervised learning of multilayer networks is still noticeably less developed (69, 70). Second, most current approaches for clustering of multilayer networks are based on graph embedding into a Euclidean space via graph spectral decomposition and, as such, do not explicitly account for local underlying graph geometry and topology. Our goal is to cluster multilayer networks in the unsupervised setting from a perspective of data shape similarities recorded at multiple resolutions. To systematically quantify the shape dynamics of multilayer networks at evolving similarity scales we introduce the multilens tools of TDA into the method CPD.

The rationale behind our clustering approach is the following. Two points are close enough to be grouped into one cluster if their local neighborhoods are similar in shape at all resolution scales. To compare the shapes we conduct the following steps:

1) Consider $X_n = (x_1, \ldots, x_n)$ in some metric space $(\mathbb{X}, \mathbb{D})$.
2) Set resolution thresholds $\nu_1 < \nu_2 < \ldots < \nu_K$ and construct a VR filtration $VR_{\nu_1}(N(i)) \subseteq VR_{\nu_2}(N(i)) \subseteq \ldots \subseteq VR_{\nu_K}(N(i))$.
3) Compute a local topological summary of x_i in the form of a persistence diagram $PD(i)$, $i = 1, \ldots, n$.
4) For all local neighborhoods $N(i)$ of x_i and $N(j)$ of x_j, $i, j = 1, 2, \ldots, n$, compute a pairwise topological or data shape dissimilarity as the 2-Wasserstein distance (37, 71) between their respective persistence diagrams $PD(i)$ and $PD(j)$:

$$W_2(x_i, x_j) = W_2(PD(i), PD(j)) \qquad [4]$$

$$= \left(\inf_{\gamma} \sum_{x \in PD(i) \cup \Delta} \|x - \gamma(x)\|_\infty^2 \right)^{1/2},$$

where $\Delta = \{(x, x) | x \in \mathbb{R}\}$ and γ is taken over all bijective maps from $PD(i) \cup \Delta$ to $PD(j) \cup \Delta$, counting multiplicities. The 2-Wasserstein distance allows us to systematically quantify how similar shapes of the two node neighborhoods are. That is, we count and compare all loops, voids, and other topological features in each node neighborhood.

5) Form a distance graph $\mathbb{G}$ over $W_2(N(i), N(j))$, $i, j = 1, 2, \ldots, n$, with adjacency matrix A, where

$$A_{ij} = \begin{cases} 1, & \text{if } W_2^{ij} \geq \kappa \\ 0, & \text{otherwise.} \end{cases}$$

The cutpoint κ is defined via elbow plot or cross-validation.

6) The connected components of $\mathbb{G}$ are our resulting clusters.

Hence, CPD utilizes both the distance function and local geometric information around the points. *SI Appendix*, *Algorithm 1* outlines the proposed CPD approach to clustering of multilayer networks. Fig. 1 shows a schematic illustration of the CPD algorithm.

We compare the results of CPD to some of the most widely used unsupervised learning algorithms, namely (complete linkage) hierarchical clustering (72–74) and K-medoids (75–77). The K-medoids, also known as partitioning around medoids, is a variant of K-means that is more robust to noise and outliers because it uses an actual, most centrally located point as the cluster center instead of a mean.

However, the most widespread implementation of K-medoids is using Euclidean distances, which ignore the geometry of the data. Therefore, in our study, we propose a K-medoids concept using Wasserstein distances—partitioning around Wasserstein medoids (K-PaWM)—that focuses on local geometry.

Performance Measures. The performance of the clustering algorithms is compared using internal cluster validation measures, specifically, mean within sum of squares (WSS), mean between sum of squares (BSS), and WB-ratio (78, 79).

Assume the data $x_1, \ldots, x_n$ in some metric space S with metric *dist* are partitioned into exhaustive and nonoverlapping sets $C_1, \ldots, C_k$. The centroids of these clusters are $r_1, \ldots, r_k$. The WSS and BSS are used as performance measures of the algorithm to group similar objects together and to differentiate between two groups of objects. Lower WSS and higher BSS are indicators of the efficacy of a clustering algorithm. The WB-ratio (80) is defined as WB-ratio $= WSS/BSS$, where

$$WSS = \frac{1}{|C_i|} \sum_{x \in C_i} dist(x, \mu)^2, \quad \mu = \frac{1}{|C_i|} \sum_{x \in C_i} x;$$

$$BSS = \frac{1}{k} \sum_{r=1}^{k} dist(r_i, r_k)^2, \quad c = \frac{1}{n} \sum_{x=1}^{n} x.$$

A smaller WB-ratio indicates that the clusters are tight and well-separated, while larger ratios indicate the opposite. These metrics are all calculated using Wasserstein distance, as Euclidean measures would not provide an adequate assessment of the algorithm's performance.

Experiments

We illustrate the utility of the topological clustering for multilayer networks in application to simulated and real data, particularly, climate-insurance networks. Clustering performance is measured by three internal cluster validation metrics, i.e., WSS, BSS, and WB-ratio. We compare the clustering performance of the CPD algorithm with respect to the following three methods: hierarchical, K-medoids, and K-PaWM algorithms.

Here we focus on the results of applying multilayer network analysis and topological clustering to the real home insurance data. (Detailed experiments on the two simulated multilayer networks are presented in *SI Appendix*. Our analysis of the simulated data suggests that CPD tends to deliver better clustering performance compared to benchmark methods for both simulated networks.) There are two datasets used in our case study. The first dataset comprises information about home insurance claims due to precipitation-induced damage for 504 forward sortation areas (FSA) in Ontario, Canada, over a 10-y period (2002 to 2011). The dataset contains the following information for each FSA: number of insurance claims, total amount of losses incurred by an insurance company (CAD), date of damage, average age of the houses (years), and average credit score. To remove the effects of risk exposure evolving over time due to sociodemographic growth, the number and amount of claims are normalized by the number of homes insured in the postal area on each day (81). To remove the effect of inflation, home insurance losses are converted to the prices of 2002 (CAD2002) using a metropolitan area composite index of apartment building construction. The insurance claim data are provided by one of the largest insurance companies in North America. The second dataset comprises daily precipitation (millimeters) obtained for the same 10-y period (2002 to 2011) from the ERA-Interim reanalysis product (82) with 0.1° spatial grid resolution.

The distributions of the number of claims and incurred losses are alike spatially (Fig. 2). The spatial pattern of credit scores is the least noticeable among all considered variables, whereas the precipitation has the strongest pattern, showing an increase in

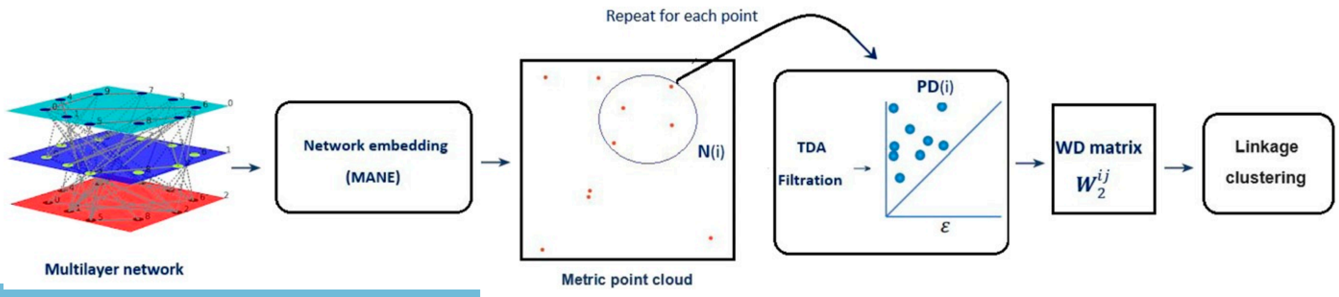

Fig. 1. Pipeline of the CPD algorithm for a three-layer network.

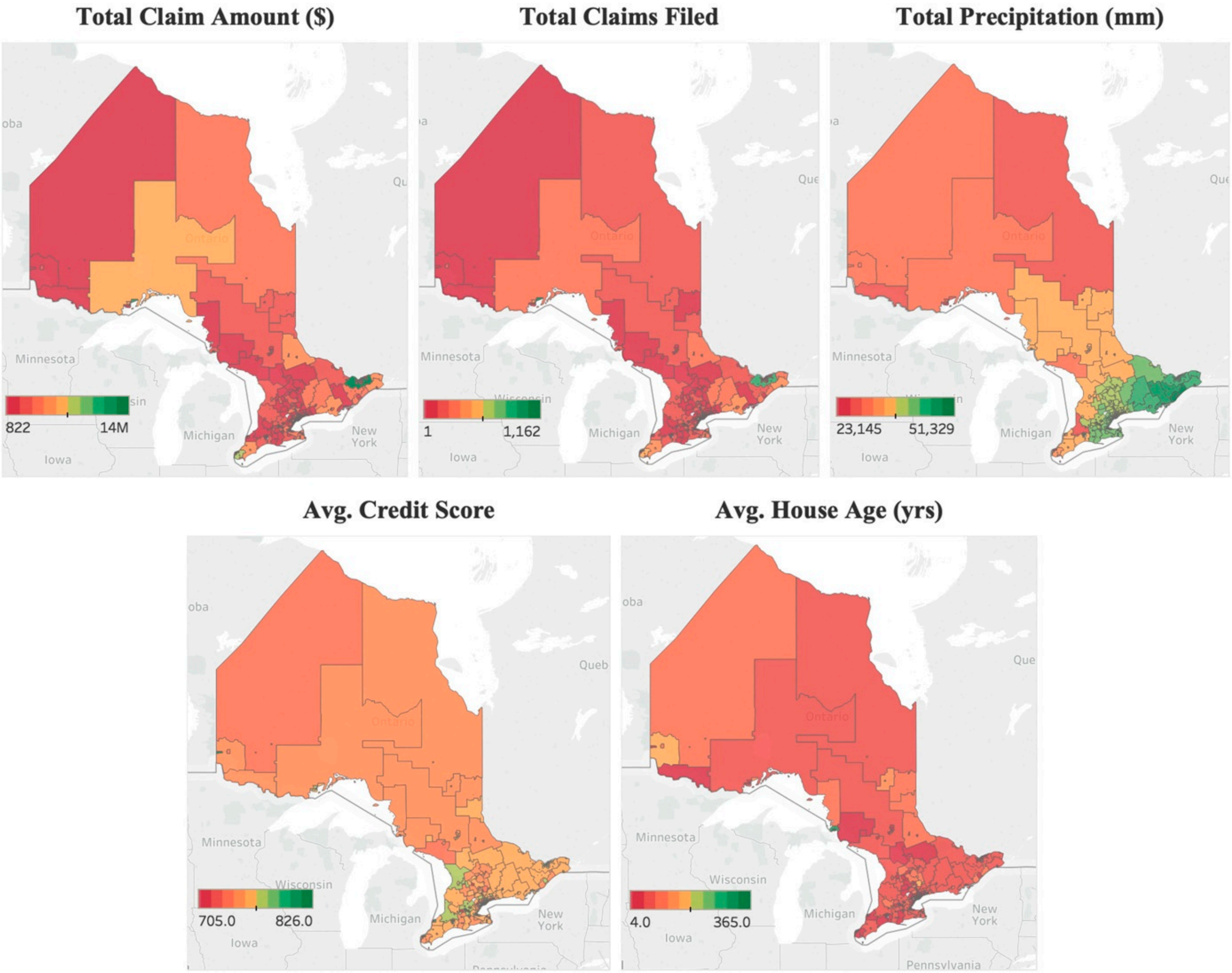

Fig. 2. Spatial distribution of the variables.

the southeast direction (toward the regions of Lake Ontario and Lake Erie; Fig. 2).

We aggregate the daily data by week and remove or substitute missing and anomalous values of each variable by an average value. As a result, for each FSA we have 520 weekly observations (for the 10-y period) of the following five variables: number of home insurance claims (N), associated losses (L), average age of the houses (H), average credit score (C), and total precipitation (P).

We construct a five-layer network, where each layer corresponds to one of the five variables and has 504 nodes (FSA). Each layer is a fully connected network, where both within-layer and cross-layer edge weights are determined by the ACE approach. The final multilayer network has seven cross-layer dependencies (Table 1).

The MANE algorithm requires two parameters: the balancing parameter α and the embedding dimension d. The parameter values are set based on the findings of ref. 66. Since in our case the contribution of the factors to insurance losses is unknown, all of the factors are weighted equally, and hence $\alpha = 1$. A number of embedding dimensions have been tried and the performance does not improve after $d = 30$. Hence, d is set to 30.

The embedded nodes of the resulting network are clustered using the proposed method of CPD and the baseline method of hierarchical clustering. Based on the elbow plot, the hierarchical clustering is stopped at 10 clusters. The CPD algorithm has two parameters, namely, the filtration length and the number of nearest neighbors. We select a filtration length of 30 and 30 nearest neighbors based on tuning experiment. The choice of 30 for both is made for computational efficiency. The K-medoids algorithm has only one parameter, namely, the number of clusters that is chosen using the elbow plot. As a result, 13 clusters are formed using K-medoids and 10 using K-PaWM.

Table 2 compares the four clustering algorithms based on three validation metrics. The WB-ratio indicates that the methods that use Wasserstein distances perform segmentation up to 10 times better than those that use Euclidean distances.

The CPD algorithm delivers competitive performance in identifying clusters, with the BSS at least three times higher

Table 1. Climate-insurance multilayer network

Layers	G_N, G_L, G_H, G_C, G_P
Within-layer adjacency matrix	A_N, A_L, A_H, A_C, A_P
No. of nodes in each layer	$n_i = 504$, $i = 1, 2, \ldots, 5$
Cross-layer dependency matrix	A_{NL}, A_{PL}, A_{PN}, A_{NH}, A_{NC}, A_{LH}, A_{LC}
Intralayer and interlayer edges	Each layer and cross-layer is a complete graph

Table 2. Summary of the internal validation measures

Metric	CPD	Hierarchical	*K*-medoids	
			Wasserstein	Euclidean
WSS	89.35	142.18	51.44	139.27
BSS	71.24	23.13	24.34	14.11
WB-ratio	1.25	23.13	2.11	9.87

than by other algorithms. On the other hand, K-PaWM has the lowest WSS and is comparable to CPD. Both methods that use Euclidean metrics perform poorly on these measures. We now turn to assessing clustering performance in terms of interpretability.

Fig. 3 presents the number of clusters and their spatial locations in Ontario, Canada, delivered by the four clustering algorithms. CPD yields the maximum number of 15 clusters (Table 3). All four methods tend to form a large cluster in the northwest of Ontario. To profile the clusters from the four methods, we study the differences in cluster means of the various attributes.

The CPD has formed two large clusters (Clusters 1 and 2) with over 70% of the FSAs (Table 3). These two clusters have similar precipitation, credit scores, average numbers of claims, and average losses per claim, but Cluster 2 is in the more urban areas of Ontario and has a higher-than-average house age. Cluster 3 has the highest average precipitation and consequently has the highest average number of claims and the highest average loss per FSA. Remarkably, the within-cluster variability of the attributes, e.g., precipitation and average number of claims and losses, is smaller with CPD than with the other three clustering methods (*SI Appendix*, Tables S6–S9). Also, the intercluster variability of the attributes is higher with CPD than with the

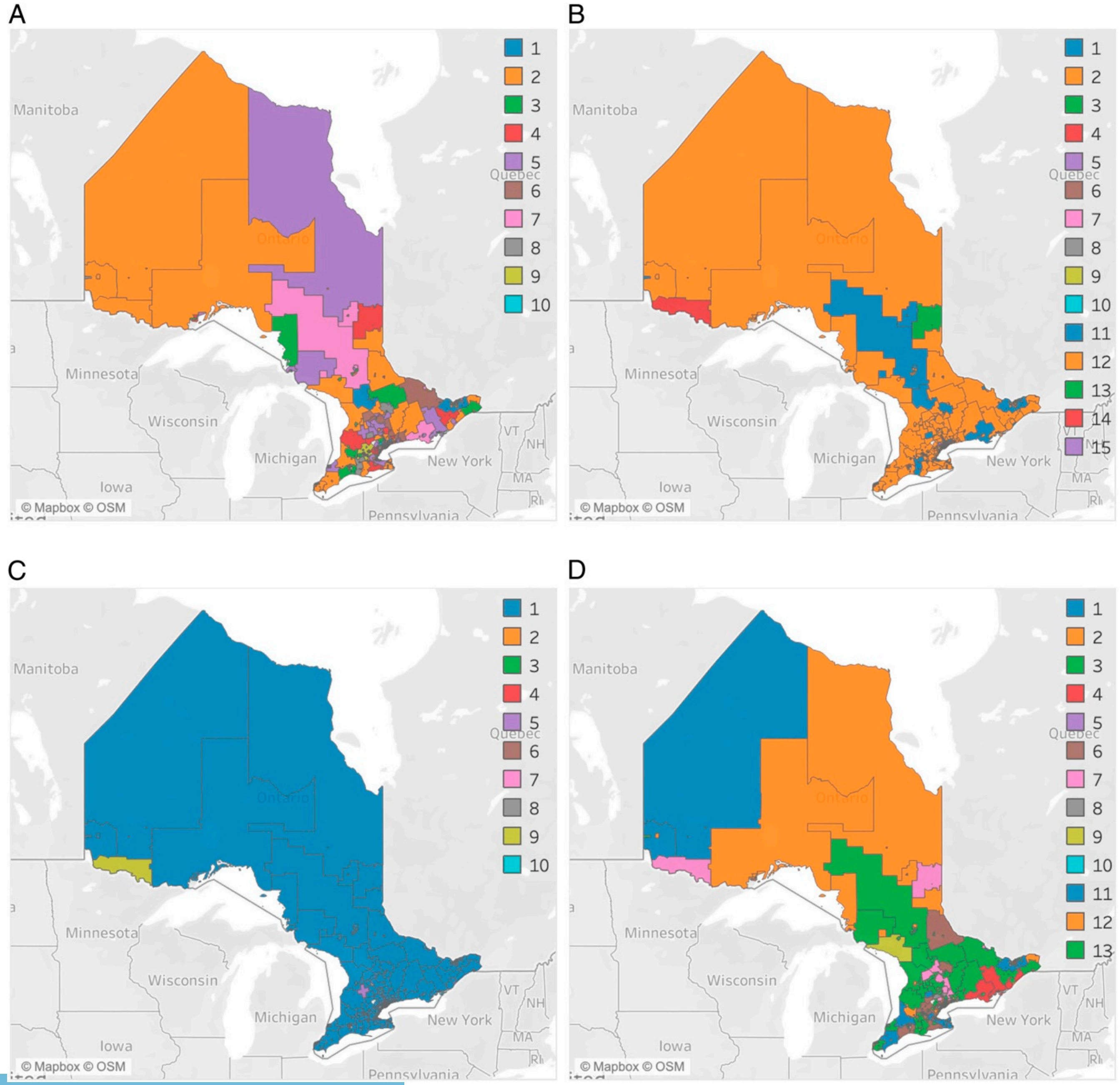

Fig. 3. Cluster labels assigned to the postal areas in southern Ontario, Canada: (*A*) *K*-PaWM, (*B*) CPD, (*C*) hierarchical, and (*D*) *K*-medoids.

Table 3. Profile (average) of the CPD clusters

Clusters	Count of FSA	Average credit score	Average precipitation	Average claim amount, $	Average no. of claims	Average house age, y
1	89	758	74	2,345	0.17	121
2	367	757	76	2,125	0.17	63
3	1	757	81	14,505	0.58	77
4	2	757	69	2,214	0.16	25
5	35	754	76	2,499	0.14	68
6	1	762	77	11,592	0.45	24
7	1	761	78	3,469	0.18	198
8	1	752	76	465	0.06	107
9	1	767	76	1,260	0.11	18
10	1	755	78	268	0.03	44
11	1	760	76	290	0.03	93
12	1	776	73	581	0.05	33
13	1	750	69	2,945	0.18	102
14	1	747	55	159	0.02	32
15	1	758	54	760	0.08	43
Total	504	757	75	2,216	0.17	74

Red shades represent very small values of the variable, yellow shades indicate values near the center, and green shades represent very large values. For two clusters, the same colored shades for a particular variable imply that the two clusters are similar in terms of that variable.

other three clustering methods. Hence, CPD appears to better capture the dynamics of the relationships of the atmospheric, sociodemographic, and insurance factors, as CPD performs clustering based on the intrinsic data shape similarities.

In contrast, hierarchical clustering does not make such a distinction and has ∼86% of the FSAs in one cluster (Cluster 1; *SI Appendix*, Table S5). The other ∼14% of the FSAs are divided among three clusters. Cluster 2 has the highest average precipitation and consequently has the second-highest average number of claims per FSA despite the lowest average house age. As Figs. 2 and 3 suggest, hierarchical clustering reflects segmentation based predominantly on precipitation only and largely disregards all other insurance-relevant information. This phenomenon is due to the fact that hierarchical clustering tends to split the dataset based on the magnitude of data variability, and precipitation exhibits the highest variability range.

In K-medoids with Wasserstein distance (i.e., K-PaWM), the FSAs are approximately evenly distributed in 10 clusters (*SI Appendix*, Table S3). Cluster 1 has the highest average number of claims and the highest average amount of losses. The classical K-medoids with Euclidean distance, on the other hand, has 13 clusters with heterogeneous cluster size (*SI Appendix*, Table S4). We observe that topological K-PaWM also tends to split the clusters based on the interplay of the atmospheric, sociodemographic, and insurance factors, while K-medoids does not capture the signal similarly well and splits out haphazard clusters as primarily being overindexed on precipitation data.

The graphs in Fig. 3 *C* and *D*, which are purely based on Euclidean distances, show geographically contiguous clusters due to precipitation similarities, while TDA methods (e.g., Fig. 3*B*) do not necessarily deliver geographically contiguous clusters.

Conclusion and Discussion

In this paper we have introduced the emerging machinery of TDA tools into the analysis of complex multilayer networks. We have developed a topological clustering algorithm (CPD) based on a multilens comparison of intrinsic data shapes in multilayer networks using a TDA concept of persistence diagram. We have validated the utility of the CPD approach with respect to conventional algorithms based on Euclidean distance, specifically, hierarchical clustering and K-medoids. Furthermore, we have proposed a modified version of K-medoids, K-PaWM, based on a topological similarity measure (i.e., Wasserstein distance). We have found that both topological approaches (i.e., CPD and K-PaWM) are competitive alternatives to conventional clustering tools when data exhibit a complex spatiotemporal dependence structure, including but not limited to geocoded multilayer networks.

We have illustrated the utility of the proposed topological clustering in application to a joint analysis of climate-insurance networks. While climate networks have been considered before, climate-induced insurance networks and multilayer climate-insurance networks have never been studied. The proposed methodology for peril maps and vulnerability zoning for the weather- and climate-induced risks developed in this paper offers multiple potential benefits to mitigate the impact of climate change for the insurance industry, policy makers, and society in general. First, insurance companies can adjust the range of insurance products offered in less- or more-vulnerable zones (e.g., extra insurance against moisture in basements or lower premium rates for houses without a basement), offer various "risk-smart" incentives for homeowners (e.g., reduced insurance premiums, if homeowners use a new generation of wind-resistant or energy-efficient roofing materials), and improve calculations for net amounts at risk (83–85). Second, many other businesses start investing in risk-prone areas on multiple fronts, e.g., modified buildings or homes, green energy, and hybrid vehicles. Insurance companies can further assist the development of these businesses by providing appropriate insurance solutions and joint initiatives for customer incentives. Third, a better understanding of risky zones allows policy makers to implement various mitigation strategies in a timely manner. For example, for more vulnerable areas, city officials may require use of more moisture-resistant wall materials for the construction of new buildings, pay closer attention to the maintenance of aging infrastructure, and integrate derived vulnerability zoning to the atlases of future urban expansion. Fourth, a more accurate zoning of vulnerable areas due to weather- and climate-induced risks allow homeowners to make more informed decisions on buying or selling houses, to reduce property damage as well as to avoid unintended injuries, and even to save lives via enhancing societal alert levels on potential climate risk in a given region. Hence, the proposed topological clustering approach for deriving more accurate zoning due to natural hazards may be viewed as one of the first steps toward informing society on the risks associated with climate change and, hence, further facilitating the development of a safer and more sustainable environment.

In the future, we plan to expand the proposed topological approach to clustering and classification of dynamic multilayer networks, as well as to derive theoretical stability guarantees for topological graph clustering. Another interesting direction is to combine the concept of similarity-based agglomerative clustering (SBAC) (86–88) with CPD by grouping points with less or more common shape feature in the population. We envision that such topological SBAC might be particularly useful in biomedical imaging, such as tumor detection. More generally, we believe that topological and geometric methods open many new promising perspectives for modeling, analysis, and inference for complex multilayer networks.

Data Availability. Data is available upon request from the corresponding author.

ACKNOWLEDGMENTS. H.V.P. has been partially supported by NSF Grants DMS 1736417, ECCS 1824710, and ECCS 2039716, and Y.R.G. has been partially supported by NSF Grants DMS 1736368, IIS 1633331, ECCS 1824716, and ECCS 2039701. V.L. has been partially supported by NSF Grants IIS 1633355 and CNS 1739823. We thank Ignacio Segovia-Dominguez for TDA diagrams. We also thank three reviewers for constructive comments and suggestions. The earlier version of this paper ("Multilayer Network and Topological Data Analysis for Mitigating Impact of Climate Change on Insurance") received the 2020 Best Student Paper Award from the American Statistical Association Section on Statistics in Defense and National Security.

1. S. Caschili, F. R. Medda, A. Wilson, An interdependent multi-layer model: Resilience of international networks. *Netw. Spatial Econ.* **15**, 313–335 (2015).
2. S. Pilosof, M. A. Porter, M. Pascual, S. Kéfi, The multilayer nature of ecological networks. *Nat. Ecol. Evol.* **1**, 101 (2017).
3. M. Pedersen, A. Zalesky, A. Omidvarnia, G. D. Jackson, Multilayer network switching rate predicts brain performance. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 13376–13381 (2018).
4. M. Wu *et al.*, A tensor-based framework for studying eigenvector multicentrality in multilayer networks. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 15407–15413 (2019).
5. C. Gomez, A. D. González, H. Baroud, C. D. Bedoya-Motta, Integrating operational and organizational aspects in interdependent infrastructure network recovery. *Risk Anal.* **39**, 1913–1929 (2019).
6. F. Messina *et al.*, COVID-19: Viral–host interactome analyzed by network based-approach model to study pathogenesis of SARS-CoV-2 infection. *J. Transl. Med.* **18**, 1–10 (2020).
7. Z. Kosowska-Stamirowska, Network effects govern the evolution of maritime trade. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 12719–12728 (2020).
8. L. Tang, K. Jing, J. He, H. Stanley, Complex interdependent supply chain networks: Cascading failure and robustness. *Phys. Stat. Mech. Appl.* **443**, 58–69 (2015).
9. H. Baroud, K. Barker, J. E. Ramirez-Marquez, C. M. Rocco, Inherent costs and interdependent impacts of infrastructure network resilience. *Risk Anal.* **35**, 642–662 (2015).
10. I. Bermudez *et al.*, Twitter response to Munich July 2016 attack: Network analysis of influence. *Front. Big Data* **2**, 17 (2019).
11. Q. Li, S. Dong, A. Mostafavi, Modeling of inter-organizational coordination dynamics in resilience planning of infrastructure systems: A multilayer network simulation framework. *PloS One* **14**, e0224522 (2019).
12. E. Nosyreva, L. Massel, "Application of multilayer networks to detect critical energy facilities" in *Proceedings of the VIth International Workshop 'Critical Infrastructures: Contingency Management, Intelligent, Agent-Based, Cloud Computing and Cyber Security' (IWCI 2019)*, L. Massel, N. Makagonova, A. Kopaygorodsky, A. Massel, Eds. (Atlantis Press, 2019), pp. 249–256.
13. I. Bachmann, J. Bustos, B. Bustos, A survey on frameworks used for robustness analysis on interdependent networks. *Complexity* **2020**, 1–17 (2020).
14. N. Yadav, S. Chatterjee, A. Ganguly, Resilience of urban transport network-of-networks under intense flood hazards exacerbated by targeted attacks. *Sci. Rep.* **10**, 10350 (2020).
15. T. Wang, M. Brede, A. Ianni, E. Mentzakis, Characterizing dynamic communication in online eating disorder communities: A multiplex network approach. *Appl. Netw. Sci.* **4**, 12 (2019).
16. P. V. Bindu, R. Mishra, P. Thilagam, Discovering spammer communities in Twitter. *J. Intell. Inf. Syst.* **51**, 503–527 (2018).
17. M. Wu *et al.*, A tensor-based framework for studying eigenvector multicentrality in multilayer networks. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 15407–15413 (2019).
18. R. Interdonato, M. Magnani, D. Perna, A. Tagarelli, D. Vega, Multilayer network simplification: Approaches, models and methods. *Comput. Sci. Rev.* **36**, 100246 (2020).
19. A. Tagarelli, A. Amelio, F. Gullo, Ensemble-based community detection in multilayer networks. *Data Min. Knowl. Discov.* **31**, 1506–1543 (2017).
20. N. Arinik, R. Figueiredo, V. Labatut, Multiple partitioning of multiplex signed networks: Application to European Parliament votes. *Soc. Network.* **60**, 83–102 (2020).
21. T. Vallès-Català, F. A. Massucci, R. Guimerà, M. Sales-Pardo, Multilayer stochastic block models reveal the multilayer structure of complex networks. *Phys. Rev. X* **6**, 011036 (2016).
22. S. Paul, Y. Chen, Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electron. J. Statist.* **10**, 3807–3870 (2016).
23. P. Barbillon, S. Donnet, E. Lazega, A. Bar-Hen, Stochastic block models for multiplex networks: An application to a multilevel network of researchers. *J. Roy. Stat. Soc.* **180**, 295–314 (2017).
24. J. D. Wilson, J. Palowitch, S. Bhamidi, A. B. Nobel, Community extraction in multilayer networks with heterogeneous community structure. *J. Mach. Learn. Res.* **18**, 5458–5506 (2017).
25. D. R. DeFord, S. D. Pauls, Spectral clustering methods for multiplex networks. *Phys. Stat. Mech. Appl.* **533**, 121949 (2019).
26. F. Znidi, H. Davarikia, K. Iqbal, M. Barati, Multi-layer spectral clustering approach to intentional islanding in bulk power systems. *J. Mod. Power Syst. Clean Energy* **7**, 1044–1055 (2019).
27. P. Mercado, F. Tudisco, M. Hein, "Spectral clustering of signed graphs via matrix power means" in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri, R. Salakhutdinov, Eds. (ML Research Press, 2019), pp. 1–11.
28. S. Paul, Y. Chen, Spectral and matrix factorization methods for consistent community detection in multi-layer networks. *Ann. Stat.* **48**, 230–250 (2020).
29. D. J. Watts, S. H. Strogatz, Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
30. A. R. Benson, D. F. Gleich, J. Leskovec, Higher-order organization of complex networks. *Science* **353**, 163–166 (2016).
31. R. Ghrist, Barcodes: The persistent topology of data. *Bull. Am. Math. Soc.* **45**, 61–75 (2008).
32. G. Carlsson, Topology and data. *BAMS* **46**, 255–308 (2009).
33. F. Chazal, B. Michel, An introduction to topological data analysis: Fundamental and practical aspects for data scientists. arXiv [Preprint] (2017). https://arxiv.org/abs/1710.04019 (Accessed 15 January 2020).
34. G. Carlsson, "Persistent homology and applied homotopy theory" in *Handbook of Homotopy Theory*, H. Miller, Ed. (CRC Press, 2019), pp. 297–330.
35. P. Y. Lum *et al.*, Extracting insights from the shape of complex data using topology. *Sci. Rep.* **3**, 1236 (2013).
36. N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, H. A. Harrington, A roadmap for the computation of persistent homology. *EPJ Data Sci.* **6**, 17 (2017).
37. L. Wasserman, Topological data analysis. *Annu. Rev. Stat. App.* **5**, 501–532 (2018).
38. M. R. McGuirl, A. Volkening, B. Sandstede, Topological data analysis of zebrafish patterns. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 5113–5124 (2020).
39. G. Singh, F. Memoli, G. Carlsson,"Topological methods for the analysis of high dimensional data sets and 3D object recognition" in *Eurographics Symposium on Point-Based Graphics 2007* (IEEE, 2007), pp. 91–100.
40. F. Chazal, L. Guibas, S. Oudot, P. Skraba, Persistence-based clustering in Riemannian manifolds. *JACM* **60**, 1–38 (2013).
41. U. Islambekov, Y. R. Gel, Unsupervised space-time clustering using persistent homology. *Environmetrics* **30**, e2539 (2019).
42. M. Golnaraghi, *Climate Change and the Insurance Industry: Taking Action as Risk Managers and Investors* (The Geneva Association, 2018).
43. A. A. Tsonis, K. L. Swanson, P. J. Roebber, What do networks have to do with climate? *Bull. Am. Meteorol. Soc.* **87**, 585–596 (2006).
44. A. Gozolchiani, K. Yamasaki, O. Gazit, S. Havlin, Pattern of climate network blinking links follows El Niño events. *Europhys. Lett.* **83**, 28005 (2008).
45. K. Steinhaeuser, N. V. Chawla, A. R. Ganguly, An exploration of climate data using complex networks. *SIGKDD Explor. Newsl.* **12**, 25–32 (2010).
46. K. Steinhaeuser, A. R Ganguly, N. V. Chawla, Multivariate and multiscale dependence in the global climate system revealed through complex networks. *Clim. Dynam.* **39**, 889–895 (2012).
47. J. Ludescher *et al.*, Very early warning of next El Niño. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 2064–2066 (2014).
48. W. K. Huang, D. S. Cooley, I. Ebert-Uphoff, C. Chen, S. Chatterjee, New exploratory tools for extremal dependence: χ networks and annual extremal networks. *J. Agric. Biol. Environ. Stat.* **24**, 484–501 (2019).
49. M. Kivelä *et al.*, Multilayer networks. *J. Complex Netw.* **2**, 203–271 (2014).
50. N. Wider, A. Garas, I. Scholtes, F. Schweitzer. An ensemble perspective on multi-layer networks. arXiv [Preprint] (2015). https://arxiv.org/abs/1507.00169 (Accessed 12 March 2020).
51. F. W. Takes, W. A. Kosters, W. Boyd, M. Eelke, Heemskerk, Multiplex network motifs as building blocks of corporate networks. *Appl. Netw. Sci.* **3**, 1–22 (2018).
52. J. A. Baggio *et al.*, Multiplex social ecological network analysis reveals how social changes affect community robustness more than resource depletion. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 13708–13713 (2016).
53. A. Aleta, S. Meloni, Y. Moreno, A multilayer perspective for the analysis of urban transportation systems. *Sci. Rep.* **7**, 44359 (2017).
54. T. Millington, M. Niranjan, "Quantifying influence in financial markets via partial correlation network inference" in *11th International Symposium on Image and Signal Processing and Analysis*
textit (ISPA, 2019), pp. 306–311.
55. P. Giudici, G. Polinesi, Crypto price discovery through correlation networks. *Ann. Oper. Res.* **299**, 443–457 (2019).
56. D. R. Williams, P. Rast, Back to the basics: Rethinking partial correlation network methodology. *Br. J. Math. Stat. Psychol.* **73**, 187–212 (2020).
57. M. Hawrylycz *et al.*, Multi-scale correlation structure of gene expression in the brain. *Neural Netw.* **24**, 933–942 (2011).
58. G. Petri *et al.*, Homological scaffolds of brain functional networks. *J. R. Soc. Interf.* **11**, 20140873 (2014).
59. R. Yu *et al.*, Weighted graph regularized sparse brain network construction for MCI identification. *Pattern Recogn.* **90**, 220–231 (2019).
60. M. Lu, U. Lall, J. Kawale, S. Liess, V. Kumar, Exploring the predictability of 30-day extreme precipitation occurrence using a global SST–SLP correlation network. *J. Clim.* **29**, 1013–1029 (2016).
61. A. Karpatne, V. Kumar, "Big data in climate: Opportunities and challenges for machine learning" in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17* (Association for Computing Machinery, New York, 2017), pp. 21–22.

62. N. Ying, D. Zhou, Z. G. Han, Q. H. Chen, Q. Ye, Z. G. Xue, Rossby waves detection in the CO_2 and temperature multilayer climate network. *Geophys. Res. Lett.* **47**, e2019GL086507 (2020).
63. L. Breiman, J. H. Friedman, Estimating optimal transformations for multiple regression and correlation. *J. Am. Stat. Assoc.* **80**, 580–598 (1985).
64. P. Goyal, E. Ferrara, Graph embedding techniques, applications, and performance: A survey. *Knowl. Base Syst.* **151**, 78–94 (2018).
65. W. L. Hamilton, R. Ying, J. Leskovec, Representation learning on graphs: Methods and applications. arXiv [Preprint] (2017). https://arxiv.org/abs/1709.05584 (Accessed 15 January 2020).
66. J. Li, C. Chen, H. Tong, H. Liu, "Multi-layered network embedding" in *Proceedings of SDM18* (SIAM, 2018), pp. 684–692.
67. R. G. Barcodes, The persistent topology of data. *BAMS* **45**, 61–75 (2008).
68. A. Zomorodian, Fast construction of the Vietoris–Rips complex. *Comput. Graph.* **34**, 263–271 (2010).
69. X. Dong, P. Frossard, P. Vandergheynst, N. Nefedov, Clustering with multi-layer graphs: A spectral perspective. *IEEE Trans. Signal Process.* **60**, 5820–5831 (2012).
70. J. Chen, H. Molter, M. Sorge, O. Suchỳ, "Cluster editing in multi-layer and temporal graphs" in *29th International Symposium on Algorithms and Computation (ISAAC 2018)*, W.-L. Hsu, D.-J. Lee, C.-S. Liao, Eds. (Dagstuhl Publishing, 2018), pp. 24:1–24:13.
71. M. Kerber, D. Morozov, A. Nigmetov, "Geometry helps to compare persistence diagrams" in *Proceedings of the 18th Workshop on Algorithm Engineering and Experiments (ALENEX 2016)* (SIAM, 2016), pp. 103–112.
72. M. Charikar, V. Chatziafratis, R. Niazadeh, G. Yaroslavtsev, "Hierarchical clustering for Euclidean data" in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, K. Chaudhuri, M. Sugiyama, Eds. (ML Research Press, 2019), pp. 2721–2730.
73. V. Cohen-addad, V. Kanade, F. Mallmann-Trenn, C. Mathieu, Hierarchical clustering: Objective functions and algorithms. *J. ACM* **66**, 1–42 (2019).
74. D. Defays, An efficient algorithm for a complete link method. *Comput. J.* **20**, 364–366 (1977).
75. L. Kaufman, P. J. Rousseeuw, *Clustering by Means of Medoids* (Elsevier, 1987), pp. 405–416.
76. J. Deng, J. Guo, Y. Wang, A novel K-medoids clustering recommendation algorithm based on probability distribution for collaborative filtering. *Knowl. Base Syst.* **175**, 96–106 (2019).
77. E. Schubert, P. J. Rousseeuw, "Faster *k*-medoids clustering: Improving the PAM, CLARA, and CLARANS algorithms" in *Similarity Search and Applications*, G. Amato, C. Gennaro, V. Oria, M. Radovanović, Eds. (Springer International Publishing, Cham, Switzerland, 2019), pp. 171–187.
78. P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
79. Q. Zhao, M. Xu, P. Fränti, "Sum-of-squares based cluster validity index and significance analysis" in *Adaptive and Natural Computing Algorithms*, M. Kolehmainen, P. Toivanen, B. Beliczynski, Eds. (Springer, Berlin, 2009), pp. 313–322.
80. A. W. F. Edwards, L. L. Cavalli-Sforza, A method for cluster analysis. *Biometrics* **21**, 362–375 (1965).
81. V. Lyubchich, Y. R. Gel, Can we weather proof our insurance? *Environmetrics* **28**, e2433 (2017).
82. D. P. Dee *et al.*, The ERA-interim reanalysis: Configuration and performance of the data assimilation system. *Q. J. of the Roy. Met. Soc.* **137**, 553–597 (2011).
83. A. B. Smith, J. L. Matthews, Quantifying uncertainty and variable sensitivity within the US billion-dollar weather and climate disaster cost estimates. *Nat. Hazards* **77**, 1829–1851 (2015).
84. V. Lyubchich, N. K. Newlands, A. Ghahari, T. Mahdi, Y. R. Gel, Insurance risk assessment in the face of climate change: Integrating data science and statistics. *Wiley Interdis. Rev.: Comput. Stat.* **11**, e1462 (2019).
85. R. W. Katz, *Statistical Issues in Detection of Trends in Losses from Extreme Weather and Climate Events* (CRC Press, 2021), pp. 165–186.
86. C. Li, G. Biswas, Unsupervised learning with mixed numeric and nominal data. *IEEE Trans. Knowl. Data Eng.* **14**, 673–690 (2002).
87. B. Stratman, S. Mahadevan, C. Li, G. Biswas, Identification of critical inspection samples among railroad wheels by similarity-based agglomerative clustering. *Integrated Comput. Aided Eng.* **18**, 203–219 (2011).
88. G. Pettet, S. Nannapaneni, B. Stadnick, A. Dubey, G. Biswas, "Incident analysis and prediction using clustering and bayesian network" in *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)* (IEEE, 2017), pp. 1–8.