PNAS
www.pnas.org

Supplementary Information for

Topological Clustering of Multilayer Networks

Monisha Yuvaraj, Asim K. Dey, Vyacheslav Lyubchich, Yulia R. Gel and H. Vincent Poor

H. Vincent Poor.
E-mail: poor@princeton.edu

This PDF file includes:

Supplementary text
Figs. S1 to S2
Tables S1 to S9
References for SI reference citations

Supporting Information Text

1. Topological Data Analysis

Let $\mathcal{X} = \{\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_n\}$ be a set of data points in a metric space, e.g., a d-dimensional Euclidean space $\mathcal{R}^d$. For an appropriate (dis)similarity measure d, and a particular threshold ϵ_k we can form a distance graph G_k with the associated adjacency matrix $A = \mathbb{1}_{d_{ij} \leq \epsilon_k}$, where d_{ij} is the distance between points X_i and X_j. Changing the scale 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 *graph filtration.* Next, to glean the intrinsic topology and geometry underlying the data from the graph filtration, we associate an *(abstract) simplicial complex* with each G_k, $k = 1, \ldots, N$.

Definition 1 (Abstract simplicial complex) *Let $\mathbb{Y}$ be a discrete set. An abstract simplicial complex is a collection $\mathcal{C}$ of finite subsets of $\mathbb{Y}$ such that if $\sigma \in \mathcal{C}$ then $\tau \in \mathcal{C}$ for all $\tau \subseteq \sigma$. If $|\sigma| = p + 1$, then σ is called a p-*simplex.

These constructs can be thought of as higher order analogs of graphs having both topological and combinatorial structures. The latter serves well for computational purposes to extract various topological summaries from data. A major advantage of the multi-lens perspective is that it avoids the issue of searching for an optimal scale value and associated feature engineering. The choice of a simplicial complex depends on the complexity of the data and which topological features one is interested in highlighting. The *Vietoris–Rips* (VR) simplicial complex is one of the most popular choices in TDA due to the ease of its construction and computational advantages (1–3).

Definition 2 (Vietoris–Rips complex) *Let $\mathbb{X}$ be a point cloud lying in a metric space. A Vietoris–Rips complex on $\mathbb{X}$ at (dis)similarity threshold $\epsilon \geq 0$, denoted by VR_ϵ, is an abstract simplicial complex whose p-simplices, $p = 0, \ldots, d$, consist of points that are pairwise within ϵ distance of each other. Here, d is called the dimension of the complex.*

Armed with the associated VR filtration, $VR_1 \subseteq VR_2 \subseteq \ldots \subseteq VR_N$, we can track the evolution of various qualitative topological features such as connected components, loops and voids that appear and disappear as we move along the filtration. Topological features that persist over the filtration, i.e., features with longer lifespans, are likelier to contain some important intrinsic information about the data generating process. In turn, topological features with short lifespans are typically referred to as *topological noise.* Two of the most widely used descriptors of topological features are *Betti numbers* and *persistent diagrams.*

Definition 3 (Betti number) *The Betti-p number of a simplicial complex $\mathcal{C}$ of dimension d, denoted by $\beta_p(\mathcal{C})$, is defined as the rank of the p-th homology group of $\mathcal{C}$, $p = 0, 1, 2, \ldots, d$.*

For applied data analysis Betti-p numbers have a simpler practical interpretation, i.e., Betti-0 is the number of connected components, Betti-1 is the number of cycles (or loops), etc.

Definition 4 (Persistence diagram) *A multi-set $\mathcal{D}$ of points in $\mathbb{R}^2$ is called a persistence diagram (PD) with x and y coordinates being the birth and death of each topological feature, respectively. Since $d \geq b$, all points in $\mathcal{D}$ are in the half-space on or above $y = x$. Features that are located farther from $y = x$ are said to persist and constitute our main interest in data analysis.*

Fig. S1 shows the TDA filtration process.

2. Clustering based on Persistence Diagrams (CPD)

We introduce a new clustering algorithm for multilayer networks based on topological descriptor persistence diagrams named clustering using persistence diagrams (CPD). Alg. 1 shows the steps of the CPD algorithm.

Fig. S2 presents a toy example which describes the intuitive idea of the CPD clustering algorithm. The points with similar neighboring shapes are grouped into one cluster. For example, red points have heart like neighboring shapes, and hence are form a cluster. Similarly, since blue points have leaf like neighboring shapes, they are grouped together.

We compare the performance of CPD with some widely used standard clustering algorithms, namely, hierarchical clustering and K-medoids. K-medoids is a variant of K-means, where the most centrally located observed point is used as the cluster center instead of a mean. That is, in K-means the objective function is $\mathcal{L} = \arg\min_C \sum_{i=1}^{K} \sum_{x \in C_i} \|x - \bar{x}_i\|_2$, where $\bar{x}_i$ is the mean of the cluster C_i, whereas in K-medoids the objective function is

$$\mathcal{L}' = \arg\min_{C} \sum_{i=1}^{K} \sum_{x \in C_i} d(x, m_i), \qquad [1]$$

where the medoid of C_i is defined as

$$m_i = \arg\min_{x_i \in C_i} \sum_{x_j \in C_i} d(x_i, x_j). \qquad [2]$$

However, like the K-means algorithm, K-medoids also commonly uses Euclidean distances, i.e., in Eq. 1 $d(x, m_i) = \|x - m_i\|_2$, and in Eq. 2 $d(x_i, x_j) = \|x_i - x_j\|_2$. Hence, the K-medoids algorithm based on Euclidean distances tends to ignore the

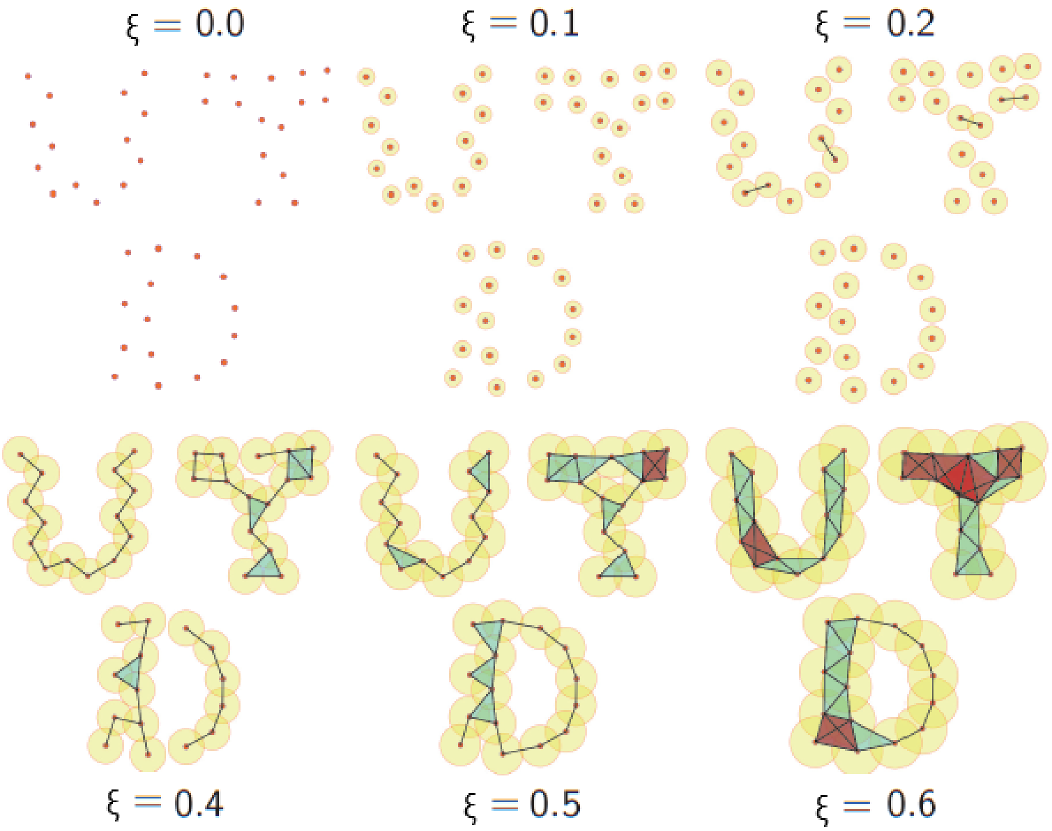

Fig. S1. Illustration of the TDA filtration for a point cloud in a metric space.

Algorithm 1 Clustering using persistence diagrams (CPD)

```
Input: data points $\{X_t\}_{t=1}^{T}$; # of neighbors $n$; dimension $d$; scale sequence $\epsilon[\,]$.
Output: TDA Clusters.
Initialize an increasing sequence $\epsilon[\,]$
for $i = 1$ to $n$ do
    $D$ = pairwise distances for $X[i]$ to all other points
    $Neigbhd[i]$ = lowest $n$ points in $D$
    $PD[i]$ = VR persistence Diagram($Neigbhd[i]$, $\epsilon[\,]$)
    $WD[i]$ = pairwise Wasserstein distance for $PD[i]$
    $C$, $A$ = ParameterSearch($PD[i]$)
    TDA Clusters = connected components of $A$
end for
ParameterSearch (Wasserstein Distances $W[i]$, Initialize range of cut-off points $c$)
for $i = 1$ to length($c$) do
    Adjacency Matrix $A$ =Filter $W[i] for \leq c[i]$
    Cluster = Connected components of $A$
    Within SS = pairwise distances of $Cluster$
    Optimal Cut-Off =$c$ corresp. min(Within SS)
    Final Clusters =$A$ corresp. Optimal Cut-Off
end for
```

geometry of the data. In our study, we propose a new K-medoids algorithm using Wasserstein distances – Partitioning around Wasserstein Medoids (K-PaWM) which focuses on local geometry. That is, we select the (dis)similarity metrics using the Wasserstein distance as

$$\begin{aligned} d(x, m_i) &= W_2\left(x, m_i\right) \qquad [3] \\ &= W_2(PD(i), PD(m_i)), \end{aligned}$$

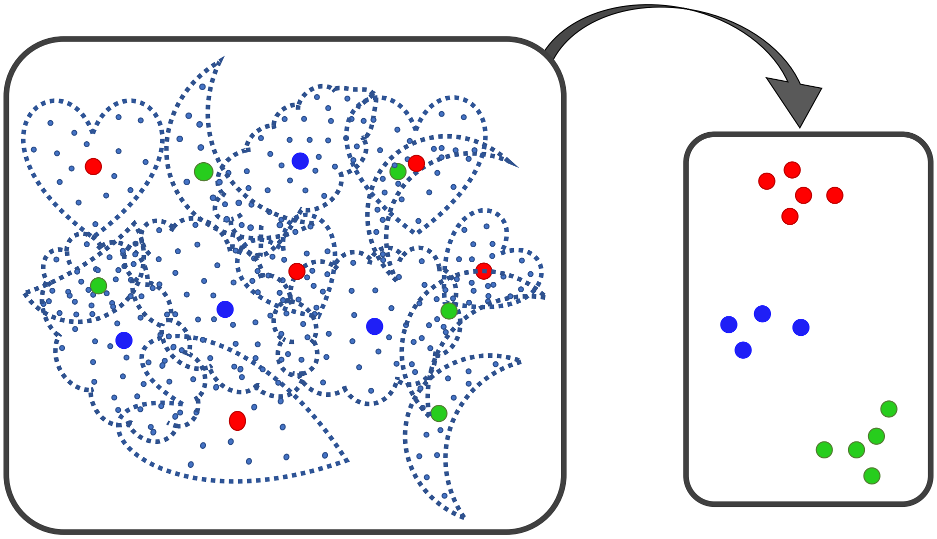

Fig. S2. An example of CPD clustering algorithm.

and

$$\begin{aligned} d(x_i, x_j) &= W_2\left(x_i, x_j\right) \\ &= W_2(PD(i), PD(j)). \end{aligned} \tag{4}$$

3. Experiment on simulated networks

In this section, we apply the CPD clustering algorithm to two simulated multilayer networks. In the first experiment, we generate a two-layer network ($\mathcal{G}^1$), where layers $G_1(500, 700)$ and $G_2(500, 700)$ are constructed based on Erdős–Rényi (ER) graph. We assign random weights ω^1 and ω^2 to the edges of G_1 and G_2, respectively. We link layer G_1 and G_2 with randomly assign 200 cross-layer edges. The adjacency matrix of the multilayer network $\mathcal{G}^1$ is $\begin{pmatrix} A_1 & A_{12} \\ A_{21} & A_2 \end{pmatrix}$, where A_1 and A_2 represent within-layer connections for layer G_1 and G_2, respectively, while A_{12} (A_{21}) represent inter-layer connectivity between G_1 and G_2. Similar to the within-layer case, we assign a random weight to each inter-layer edge.

In the second experiment, we generate a three-layer network ($\mathcal{G}^2$), where layers $G_1(400, 600)$ and $G_3(400, 700)$ are constructed based on ER graph, and layer $G_2(400, 600)$ is constructed based on preferential attachment (PA) model. We randomly set 150 cross-layer edges between G_1 and G_2, 200 cross-layer edges between G_2 and G_3, and 250 cross-layer edges between G_1 and G_3. Like in Experiment 1, we assign a random weight to each within-layer and inter-layer edge.

The adjacency matrix of the multilayer network $\mathcal{G}^2$ is $\begin{pmatrix} A_1 & A_{12} & A_{13} \\ A_{21} & A_2 & A_{23} \\ A_{31} & A_{32} & A_3 \end{pmatrix}$, where diagonal elements represent within-layer connections, and off-diagonal elements represent cross-layer links between two layers. A detailed description of the two multilayer networks ($\mathcal{G}^1$ and $\mathcal{G}^2$) is shown in Table S1.

Table S1. Simulation experiments.

Multilayer network	Experiment 1 ($\mathcal{G}^1$)	Experiment 2 ($\mathcal{G}^2$)
Layers	$G_1(500, 700, \omega^1)$, $G_2(500, 750, \omega^2)$ Edge weight: $\omega^1_{uv} \sim U(0, 3)$, $\omega^2_{uv} \sim U(0, 4)$	$G_1(400, 600, \omega^1)$, $G_2(400, 500, \omega^2)$, $G_3(400, 700, \omega^3)$ $\omega^1_{uv} \sim U(0, 2)$, $\omega^2_{uv} \sim N(10, 2)$, $\omega^3_{uv} \sim N(7, 1.5)$
Cross-layer edge	$\lvert E^{12}\rvert = 200$ Edge weight: $\omega^{12}_{uv} \sim U(0, 1)$	$\lvert E^{12}\rvert = 150.\ \omega^{12}_{uv} \sim N(4, 1)$ $\lvert E^{23}\rvert = 200.\ \omega^{23}_{uv} \sim N(4.5, 2)$ $\lvert E^{13}\rvert = 250.\ \omega^{13}_{uv} \sim N(4.5, 2)$
Within-layer adjacency matrix	A_1, A_2	A_1, A_2, A_3
Cross-layer adjacency matrix	A_{12}	A_{12}, A_{23}, A_{13}

Table S2. Simulation studies – summary of the internal validation measures.

	Metric	CPD	Hierarchical	K-medoids	
				Wasserstein	Euclidean
Experiment 1	WSS	91.260	119.181	45.516	120.121
	BSS	44.096	14.097	8.387	8.048
	WB-ratio	2.069	8.454	5.42	19.667
Experiment 2	WSS	23.887	26.500	8.012	25.032
	BSS	12.209	2.210	2.387	3.242
	WB-ratio	1.956	11.992	3.351	7.721

We now apply the CPD and K-PaWM clustering algorithms to the two simulated multilayer networks. We evaluate the performance of topological clustering with respect to the standard clustering algorithms, i.e., hierarchical clustering and K-medoids, based on Euclidean distances. Table S2 shows the validation indices of the four clustering algorithms based on three validation metrics, i.e., WSS, BSS, and WB-ratio.

We find that for both multilayer networks ($\mathcal{G}^1$ and $\mathcal{G}^2$) the CPD algorithm outperforms hierarchical, K-medoids and K-PaWM clustering methods. In both Experiments 1 and 2, CPD delivers about 4 times better (lower) WB-ratio than hierarchical and K-medoids algorithms. CPD also yields approximately 2 times better WB-ratio than the K-PaWM algorithm. BSS of CPD is also noticeably higher (at least 3.5 times) than that of hierarchical, K-medoids, and K-PaWM algorithms. In turn, K-PaWM delivers better clustering performance than that yielded by the Euclidean distance based algorithms, i.e., hierarchical and K-medoids algorithm.

4. Climate-insurance network

For climate-insurance network, to profile the clusters from the four methods (i.e., CPD, hierarchical, K-medoids and K-PaWM), we study the differences in cluster means of the various attributes. Table S3 represents the profile of the clusters obtained using K-medoids with Wasserstein distance (i.e., K-PaWM). K-medoids with Euclidean distance based clusters are shown in Table S4. The profiles of the hierarchical clusters are shown in Table S5. Table S6, Table S7, Table S8, and Table S9 present within cluster variability of the attributes for CPD, hierarchical, K-medoids and K-PaWM, respectively.

Table S3. Profile (average) of the K-PaWM clusters.

Clusters	Count of FSA	Avg. credit score	Avg. precip.	Avg. claim amount ($)	Avg. # claims	Avg. house age
1	14	762	71	$ 3,386	0.25	67
2	92	756	75	$ 3,061	0.23	60
3	76	758	75	$ 1,811	0.15	68
4	46	756	74	$ 2,535	0.20	68
5	69	757	76	$ 2,135	0.18	58
6	77	756	77	$ 1,133	00.10	65
7	37	761	76	$ 2,560	0.19	45
8	45	755	74	$1,818	00.13	189
9	13	755	77	$ 2,686	0.17	69
10	35	754	76	$ 2,499	00.14	68
Total	**504**	**757**	**75**	**$ 2,216**	**0.17**	**74**

Table S4. Profile (average) of the K-medoids clusters.

Clusters	Count of FSA	Avg. credit score	Avg. precip.	Avg. claim amount ($)	Avg. # claims	Avg. house age
1	48	759	75	$ 2,549	0.20	86
2	60	756	75	$ 2,981	0.23	57
3	191	757	75	$ 1,363	0.12	78
4	22	768	86	$ 3,074	0.22	70
5	3	770	83	$ 4,016	0.30	32
6	78	755	75	$ 2,637	0.17	67
7	45	757	74	$ 1,258	0.11	94
8	9	765	80	$ 2,920	0.20	58
9	17	753	78	$1,580	0.18	60
10	5	763	76	$ 1,397	0.10	65
11	10	750	75	$ 3,103	0.19	83
12	6	762	59	$10,458	0.85	92
13	10	752	62	$ 5,937	0.34	47
Total	**504**	**757**	**75**	**$ 2,216**	**0.17**	**74**

Table S5. Profile (average) of the hierarchical clusters.

Clusters	Count of FSA	Avg. credit score	Avg. precip.	Avg. claim amount ($)	Avg. # claims	Avg. house age
1	471	757	75	$ 2,204	0.17	75
2	4	767	83	$ 3,688	0.29	29
3	2	769	78	$ 773	0.09	69
4	7	754	81	$ 4,298	0.32	52
5	9	754	78	$ 2,465	0.16	49
6	1	753	81	$ 748	0.08	53
7	6	764	77	$ 1,190	0.09	65
8	1	751	73	$ 1,659	0.12	15
9	2	751	66	$ 214	0.02	38
10	1	760	76	$ 290	0.03	93
Total	**504**	**757**	**75**	**$ 2,216**	**0.17**	**74**

Table S6. Profile (variance) of the CPD clusters.

Cluster	Count of FSA	Var. credit score	Var. Precip	Var. claim amount	Var. # claims	Var. house age
1	89	223.973	54.774	14057637.67	0.0498	3311.332
2	367	176.794	55.855	7563813.821	0.0437	1750.232
3	1	0.000	0.000	0.000	0.000	0.000
4	2	1 34.511	84.926	5635866.582	0.0238	269.125
5	35	1236.238	5.222	7009127.249	0.016	1351.172
6	1	0.000	0.000	0.000	0.000	0.000
7	1	0.000	0.000	0.000	0.000	0.000
8	1	0.000	0.000	0.000	0.000	0.000
9	1	0.000	0.000	0.000	0.000	0.000
10	1	0.000	0.000	0.000	0.000	0.000
11	1	0.000	0.000	0.000	0.000	0.000
12	1	0.000	0.000	0.000	0.000	0.000
13	1	0.000	0.000	0.000	0.000	0.000
14	1	0.000	0.000	0.000	0.000	0.000
15	1	0.000	0.000	0.000	0.000	0.000

Table S7. Profile (variance) of the hierarchical clusters.

Cluster	Count of FSA	Var. credit score	Var. Precip	Var. claim amount	Var. # claims	Var. house age
1	471	193.112	52.164	9220263.41	0.044	2079.903
2	4	44.515	1.34E-05	9452679.176	0.052	200.706
3	2	19.381	49.617	1101472.531	0.011	2902.621
4	7	33.946	106.072	2836225.904	0.018	699.647
5	9	52.867	59.457	8151427.825	0.019	625.109
6	1	0.000	0.000	0.000	0.000	0.000
7	6	105.521	0.586	431618.396	0.001	1578.694
8	1	0.000	0.000	0.000	0.000	0.000
9	2	32.044	260.001	5996.717	8.99E-05	79.036
10	1	0.000	0.000	0.000	0.000	0.000

Table S8. Profile (variance) of the K-medoids clusters.

Cluster	Count of FSA	Var. credit score	Var. Precip	Var. claim amount	Var. # claims	Var. house age
1	48	176.785	60.857	17853447.981	0.072	3840.781
2	60	128.099	72.982	9715823.582	0.045	2053.572
3	191	210.301	35.606	2187028.461	0.015	1240.892
4	22	88.271	26.435	5574418.679	0.034	863.386
5	3	14.682	1.78E-05	13531935.321	0.077	235.826
6	78	172.543	36.348	8987181.334	0.022	1432.463
7	45	220.080	36.274	969721.784	0.006	3557.068
8	9	101.398	15.965	4904832.972	0.012	2295.301
9	17	133.186	33.176	828403.254	0.014	816.820
10	5	127.473	0.126	219068.450	0.001	1971.725
11	10	233.283	22.706	981551.101	0.004	2133.698
12	6	145.868	78.826	101575481.102	0.809	2817.012
13	10	145.489	6.006	42672425.79	0.102	477.183

Table S9. Profile (variance) of the K-PaWM clusters.

Cluster	Count of FSA	Var. credit score	Var. Precip	Var. claim amount	Var. # claims	Var. house age
1	14	94.948	76.233	51216463.24	0.184	7681.152
2	92	113.676	75.233	10805130.75	0.054	1161.887
3	76	162.116	63.342	3739542.997	0.020	1884.616
4	46	225.653	77.127	12115967.33	0.065	2060.360
5	69	125.654	54.725	9869510.138	0.076	2217.150
6	77	284.125	26.162	1478103.348	0.008	1772.471
7	37	133.463	60.365	11783764.38	0.037	832.406
8	45	303.032	42.127	4849099.728	0.013	3419.479
9	13	157.623	3.052	13577092.01	0.020	2810.679
10	35	236.238	5.222	7009127.249	0.016	1351.172

References

1. Carlsson G (2009) Topology and data. *BAMS* 46(2):255–308.
2. Zomorodian A (2010) Fast construction of the Vietoris–Rips complex. *Computers and Graphics* 34(3):263–271.
3. Carlsson G (2019) Persistent homology and applied homotopy theory. *Handbook of Homotopy Theory.*