Topological data analysis of truncated contagion maps

View Online   Export Citation   CrossMark

Florian Klimm[1,2,a] (ID)

AFFILIATIONS

[1] Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestraße 63-73, D-14195 Berlin, Germany
[2] Department of Computer Science, Freie Universität Berlin, Arnimallee 3, D-14195 Berlin, Germany

Note: This article is part of the Focus Issue, Dynamics on Networks with Higher-Order Interactions.
[a] **Author to whom correspondence should be addressed:** klimm@molgen.mpg.de

ABSTRACT

The investigation of dynamical processes on networks has been one focus for the study of contagion processes. It has been demonstrated that contagions can be used to obtain information about the embedding of nodes in a Euclidean space. Specifically, one can use the activation times of threshold contagions to construct contagion maps as a manifold-learning approach. One drawback of contagion maps is their high computational cost. Here, we demonstrate that a truncation of the threshold contagions may considerably speed up the construction of contagion maps. Finally, we show that contagion maps may be used to find an insightful low-dimensional embedding for single-cell RNA-sequencing data in the form of cell-similarity networks and so reveal biological manifolds. Overall, our work makes the use of contagion maps as manifold-learning approaches on empirical network data more viable.

It is known that the analysis of spreading processes on networks may reveal their hidden geometric structures. These techniques, called contagion maps, are computationally expensive, which raises the question of whether they can be methodologically improved. Here, we demonstrate that a truncation (i.e., early stoppage) of the spreading processes leads to a substantial speedup in the computation of contagion maps. For synthetic networks, we find that a carefully chosen truncation may also improve the recovery of hidden geometric structures. We quantify this improvement by comparing the topological properties of the original network with the constructed contagion maps by computing their persistent homology. Finally, we explore the embedding of single-cell transcriptomics data and show that contagion maps can help us to distinguish different cell types.

I. INTRODUCTION

The study of spreading processes has a rich tradition in epidemiology,[1–3] the social sciences,[4–9] and applied mathematics.[10–12] Due to improved availability of interaction data, the description of contagion processes on networks has been a focus in recent years.[13] It is known that spreading phenomena are influenced by the geometry of a networks' underlying embedding,[14] for example, for the spread of pandemics[15] and the adaption of technologies.[16] By constructing *diffusion maps*, the interplay between dynamical processes and underlying geometry can be used to detect the manifold structure in high-dimensional data.[17–20] More recently, *contagion maps* have been proposed as a technique to detect the manifold structure in networks.[21] Contagion maps may reliably detect the manifold structure in noisy data, even if standard approaches, such as ISOMAP fail.[22] Generalization of these approaches, for example, to contagions on simplicial complexes[23] and Kleinberg-like networks,[24] is an active field of research. The construction of contagion maps, however, is computationally expensive, which makes the exploration of strategies for their speed-up a pressing issue. In the context of random walks on networks, it has been established that stopping a dynamical process and restarting it may improve the performance of machine-learning algorithms,[25–27] raising the question whether similar approaches might be able to enhance contagion maps.

In this work, we introduce the generalization of contagion maps that we call *truncated contagion maps* (see Fig. 1). The core
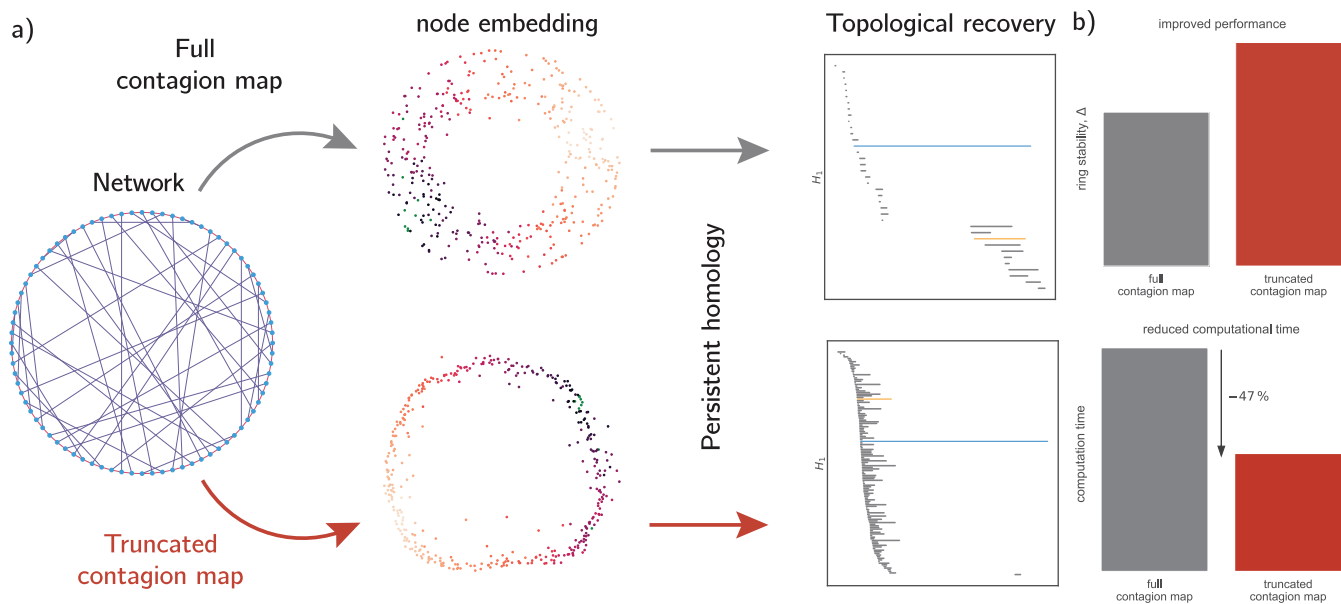
FIG. 1. Truncated contagion maps can improve recovery of topological features of a networks' underlying manifold while having reduced computational cost. (a) We show a schematic representation of the analysis pipeline in this paper. With contagion maps, we can compute a node embedding for a network. With persistent homology, we compute barcode diagrams to quantify the recovery of topological features in the learned embeddings. This approach is used for the full contagion map and the newly introduced truncated contagion map. (b) We compare the topological recovery of full contagion maps with that of truncated contagion maps. For appropriate parameter choices, truncated contagion maps show higher performance than full contagion maps while taking considerably less time to compute (numbers for a noisy ring lattice with $N = 1000$ nodes).

idea is that an early stoppage of the contagion processes that are used to detect the manifold may drastically reduce the computational cost of the embedding algorithm. In particular, our numerical observations indicate a quadratic scaling in the problem size for full contagion maps and a sub-quadratic scaling for truncated contagion maps. Furthermore, as we will show, the activations in the early stages of a contagion often follow the underlying manifold of the networks more clearly than the later stages. Omitting these late-stage dynamics results in improved manifold detection, which we verify with techniques from *topological data analysis.*

Topological data analysis is a field that aims to develop tools that allow the extraction of qualitative features in data that are hidden to standard approaches. A wide range of tools and methods for the topological analysis of data have been developed[28–30] and applied to problems in biology,[31,32] social sciences,[33–35] and many other fields.[36–38] *Persistent homology*[39,40] in particular is a topological data analysis method that received a lot of interest from practitioners because it enables the extraction of topological features in high-dimensional point-cloud data. In particular, topological data analysis is considering higher-order interactions between data points by constructing simplicial complexes.

Advances in *single-cell transcriptomics* enable the measurement and analysis of gene expression at a single-cell resolution.[41] One challenge in the analysis of single-cell transcriptomics data is

that experimental advances lead to an explosion in dataset sizes.[42] As complex organisms have thousands to ten thousands of genes, their gene-expression space is high-dimensional, making single-cell transcriptomics data notoriously hard to visualize. Yet, the vast majority of this space is empty and cells follow a few low-dimensional manifolds that represent, for example, cell types or differentiation trajectories. The presence of low-dimensional structures embedded in high-dimensional gene-expression spaces makes single-cell transcriptomics a promising field for the application of manifold-learning approaches.[43] In this work, we apply truncated contagion maps to single-cell transcriptomics data and show that the constructed embeddings are fruitful in representing the complex biological structure, which represents cell types and developmental trajectories.

The remainder of this article is organized as follows. In Sec. II, we summarize the existing methods and introduce truncated contagion maps. In Sec. III, we investigate the temporal development of the contagion processes. In Sec. IV, we demonstrate that truncated contagion maps may improve the recovery of topological features in contagion maps. In Sec. V, we discuss the improved computational complexity of truncated contagion maps vs full contagion maps. In Sec. VI, we apply truncated contagion maps to single-cell transcriptomics data of developing mouse oocytes. In Sec. VII, we discuss our findings.

II. METHODS

A. Noisy geometric networks

Noisy geometric networks[21] are geometric networks[44] to which "noisy" non-geometric edges have been added. They are defined as follows. Consider a manifold $\mathcal{M}$ that is embedded in an ambient space $\mathcal{A}$ such that $\mathcal{M} \subset \mathcal{A}$. First, we construct nodes V that are embedded in the manifold $\mathcal{M}$ such that $\mathbf{w}^{(i)} \in \mathcal{M}$ for nodes $i \in V$. Second, we construct edges which are of one of the two types, either *geometric* or *non-geometric*. The geometric edges $E^{(G)}$ follow the manifold such that $i, j \in E^{(G)}$ if the distance along the manifold between nodes i and j is below some distance threshold. The non-geometric edges $E^{(NG)}$ are constructed in a random process that does not consider the manifold. We obtain the noisy geometric network as $G = (V, E^{(G)} \cup E^{(NG)})$. Noisy geometric networks are a type of small-world network[13] that consists of long-range and short-range connections. The interplay of these two types of interactions has been discussed in the context of many types of networks, for example, social networks[45] and traffic networks.[14]

In this paper, we focus on noisy ring lattices $NRL(N, d^{(G)}, d^{(NG)})$ as prototypical examples of noisy geometric networks. Their parameters are the number N of nodes, the geometric degree $d^{(G)}$, and the non-geometric degree $d^{(NG)}$. The NRL networks are defined as follows. First, we place N nodes uniformly spaced along the unit circle $\mathcal{M} = \{(a, b) | a^2 + b^2 = 1\} \subset \mathbb{R}^2$ such that $\mathbf{w}^{(i)} = (\cos(2\pi i/N), \sin(2\pi i/N))$ is the location of node i. Second, we construct geometric edges such that each node is connected to its $d^{(G)}$ nearest-neighbor nodes. Third, we construct $N \times d^{(NG)}$ non-geometric edges uniformly at random such that each node has exactly $d^{(NG)}$ non-geometric edges. In practice, the non-geometric edges are constructed with *stub-matching*: We create a list of stubs that contains each node exactly $d^{(NG)}$ times. Then, we construct edges iteratively by drawing without replacement uniformly at random from this list while rejecting self-edges or parallel edges. In rare

cases, this procedure might fail to terminate (e.g., if only stubs of the same node are remaining), but this problem can be overcome by restarting the procedure. The ratio $\alpha = d^{(NG)}/d^{(G)}$ of non-geometric to geometric edges quantifies the "noisiness" of a noisy ring lattice with $\alpha = 0$, indicating that there are exclusively geometric edges and $\alpha = 1$ indicating that there are as many non-geometric edges as geometric edges. In Fig. 2, we show two example noisy geometric lattices with $\alpha = 0$ and $\alpha = 1$, respectively.

B. Watts threshold model

The Watts threshold model (WTM) is a well-established deterministic binary-state discrete-time model of social contagions on networks[46] and is a modification of Mark Granovetter's threshold model.[7] The WTM is defined as follows. Each node is in one of the two states, either active or inactive. At time $t = 0$, all nodes but a set of *seed nodes S* is inactive. All nodes have the same activation threshold $T \in [0, 1]$ (homogenous threshold), which determines how easily a node adapts to their neighbors' activation. An active node stays active but an inactive node is activated if more than a threshold fraction T of its d neighbors are active. We update all node states synchronously at each time point and stop the contagion once a steady state is reached (for a discussion of synchronous vs asynchronous update procedures, see Ref. 47).

C. Contagion maps

As introduced by Taylor *et al.*, we use the WTM to construct deterministic embeddings for networks. These embeddings are based on the nodes' activation times under different initial conditions. Specifically, we define a contagion map as $V \rightarrow \{\mathbf{x}^{(i)}\}_{i \in V}$, where $\mathbf{x}^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_j^{(i)}]$ with $x_j^{(i)}$ indicating the activation time of node i for contagion j. For the initial conditions, we use *cluster seeding* such that the neighborhood $\mathcal{N}(j)$ of node j is active at
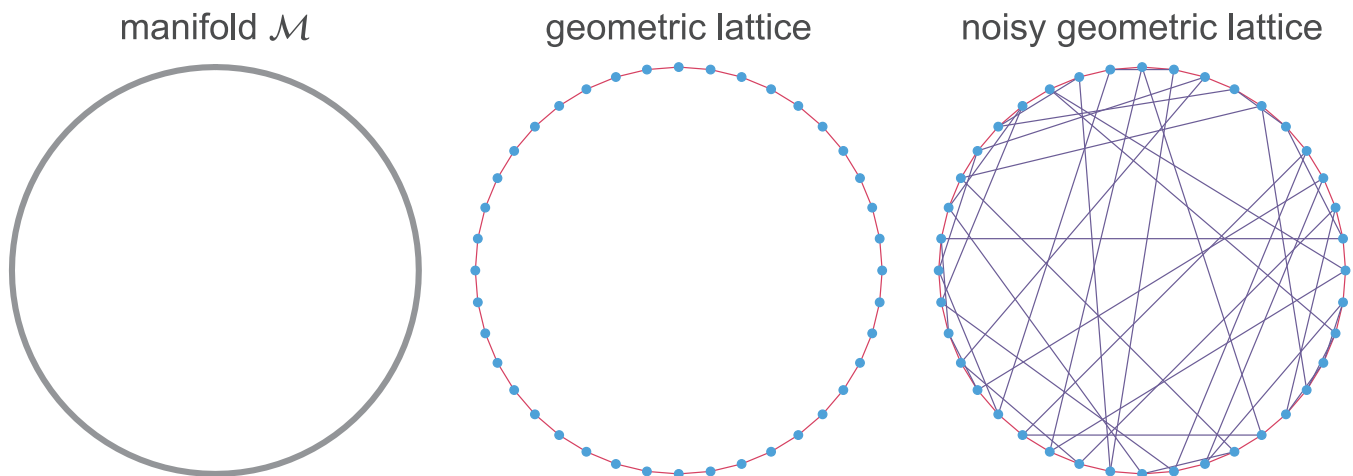

FIG. 2. Noisy geometric networks contain geometric edges that follow the underlying manifold and non-geometric edges that ignore the underlying manifold. Here, we show noisy geometric lattices of size $N = 50$, whose underlying manifold $\mathcal{M}$ is the unit circle $\{(a, b) | a^2 + b^2 = 1\}$ with ambient space $\mathcal{A} \subset \mathbb{R}^2$. The geometric lattice contains exclusively geometric edges and has $d^{(G)} = 2$. After adding $d^{(NG)} = 2$ non-geometric edges per node, we obtain a noisy geometric lattice with $\alpha = 1$.

ALGORITHM 1. Computing (truncated) contagion maps

---

input : Network $G = (V, E)$ with $N = |V|$ nodes
Threshold T
Number of steps s `// Ignored for full`
`contagion maps`
output: Node embedding $\mathbf{X}_{\text{symmetric}} \in \mathbb{N}^{N \times N}$
begin
    for *each node $j \in V = \{1, \ldots, N\}$* **do**
        Initialise cluster seeding by S infecting j and its
         neighbours;
        $\mathbf{X}[:, \mathbf{j}] \leftarrow$ Compute (truncated) contagion dynamics;
    end
    $\mathbf{X}_{\text{symmetric}} = \mathbf{X} + \mathbf{X}^\top$;
    return $\mathbf{X}_{\text{symmetric}}$
end

$T = 0$. After a steady state of the WTM is reached, some nodes might still be inactive. The infinite activation time of these inactive nodes is set to $2N \ll \infty$ because $2N$ is much larger than the largest possible activation time of $N - 1$ (An activation time of $N - 1$ would be reached by a single node if in each time step exactly one additional node is activated, for example, along a line graph.).

The obtained matrix $\mathbf{X}$ of activation times is not necessarily symmetric. To study an embedding that defines a semimetric, we investigate the symmetric contagion map $\mathbf{X}_{\text{symmetric}} = \mathbf{X} + \mathbf{X}^\top$. See Algorithms 1 and 2 for pseudocode of contagion maps and contagion dynamics, respectively.

For clarity, we will refer to the contagion maps as defined by Taylor *et al.* as "full contagion maps" in contrast to the "truncated contagion maps," which we will introduce now.

D. Truncated contagion maps

In this work, we introduce *truncated contagion maps* as a generalization of full contagion maps. Intuitively, they represent full contagion maps in which contagions are not run until a steady state is reached. Rather, we stop the contagion after $s \in \mathbb{N}_{>0}$ steps. Specifically, we define a truncated contagion map as $V \rightarrow \{\mathbf{x}^{(i)}\}_{i \in V}$, where $\mathbf{x}^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_j^{(i)}]$ with

$$x_j^{(i)} = \begin{cases} a_j^{(i)} & \text{for } a_j^{(i)} \leq s, \\ 2s & \text{else,} \end{cases} \tag{1}$$

where $a_j^{(i)}$ indicates the activation time of node i for contagion j. For the initial conditions, we use *cluster seeding* such that the neighborhood $\mathcal{N}(j)$ of node j is active at $T = 0$. We set the activation time of unactivated nodes to $2s$ such that full contagion maps are a special case of truncated contagion maps for $s = N$ but other choices for the handling of infinite activation times are possible. In practice, however, we anticipate approximately similar behavior of truncated and full contagion maps for $s \ll N$ because most contagions reach a steady state after a small number s of steps (in comparison to the network size N). As for full contagion maps, we study exclusively the

---

ALGORITHM 2. Compute (truncated) contagion dynamics

---

input : Network $G = (V, E)$ with $N = |V|$ nodes
seed nodes $S \subset V$
threshold T
max number of steps s
output: Activation time $\vec{x} \in \mathbb{N}^N$
begin
    Initialise activation times $\vec{x} = \vec{\text{NaN}}$
    Initialise step number $t = 0$
    while $\vec{x} \neq \vec{x}$ **do**
        for *each node $i \in V = \{1, \ldots, N\}$* **do**
            Check if node i activated
            Update $\vec{x}$
        end
        `// Break condition only checked for`
           `truncated contagion maps`
        if $t == s$ **then**
            Set activation time to $2 \times s$ for inactive nodes
            break while loop
        end
    end
    return $\vec{x}$
end

---

symmetric contagion maps $\mathbf{X}_{\text{symmetric}} = \mathbf{X} + \mathbf{X}^\top$ such that we obtain a semimetric.

E. Topological data analysis

Persistent homology is a widely adopted method from topological data analysis. It aims to identify stable (i.e., persistent) topological features of point-cloud data across a range of resolutions. To obtain these topological features, we must construct simplicial complexes from the point-cloud data. Specifically, we employ Vietoris–Rips simplicial filtration, which is a sequence of simplicial complexes that constructs a topological structure based on the data. Given a dataset X, a metric $d : Y \times Z \rightarrow \mathbb{R}$, and a scale parameter $\alpha \in \mathbb{R}$, we can construct a Vietoris–Rips complex as

$$\{VR(X, d, \alpha) := \{\sigma \subset X : d(p, q) \leq \alpha \ \forall \ p, q \in \sigma\}. \tag{2}$$

Computing the Vietoris–Rips complex for a selection of scale parameters $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_a$, then yields Vietoris–Rips simplicial filtration

$$\emptyset \subseteq \text{VR}(X, d, \alpha_1) \subseteq \text{VR}(X, d, \alpha_2) \subseteq \cdots \subseteq \text{VR}(X, d, \alpha_a), \tag{3}$$

which is a sequence of embedded simplicial complexes. This filtration can be thought of as a dynamical process on a simplicial complex in which faces are added stepwise. The most common way to visualize the persistence of topological features across a filtration is a *barcode*, which indicates in which filtration step a topological feature is born and in which step it ceases to exist. We use the RIPSER implementation[48] for the computation of Vietoris–Rips persistence barcodes and use the Euclidean norm. To summarize the 1D persistence barcodes (i.e., the presence of loops in the data), we

compute *ring stability* as $\Delta = l_1 - l_2$, where l_1 and l_2 are the persistence of the longest and second longest lifetime of loops in the data. This procedure has previously been used to identify whether contagions are predominantly spreading by wavefront propagation vs new cluster appearance.[21] In Fig. 1, we show two example Vietoris–Rips persistence barcodes constructed from contagion maps.

III. TEMPORAL DEVELOPMENT OF CONTAGIONS REVEALS SUITABLE TRUNCATION VALUES

In this section, we will describe the spread of WTM contagions on noisy ring lattices. In particular, we will demonstrate that in certain parameter regimes, the initial spread is mainly occurring along the manifold (called "wave front propagation," WFP) and at later stages of the contagion, the spread is less restricted to the manifold and leads to new contagion clusters at distant parts of the manifold ("appearance of new clusters," ANCs).

Taylor *et al.*[21] proved two critical thresholds for noisy ring lattices of noisiness α,

$$T^{(\text{WFP})} = 1/(2 + 2\alpha) \text{ and} \tag{4}$$

$$T^{(\text{ANC})} = \alpha/(1 + \alpha), \tag{5}$$

which indicate the thresholds above which no WFP and no ANC, respectively, occur. We show these critical thresholds, which intersect at $(\alpha, T) = (1/2, 1/3)$, in Fig. 3 and observe four qualitatively
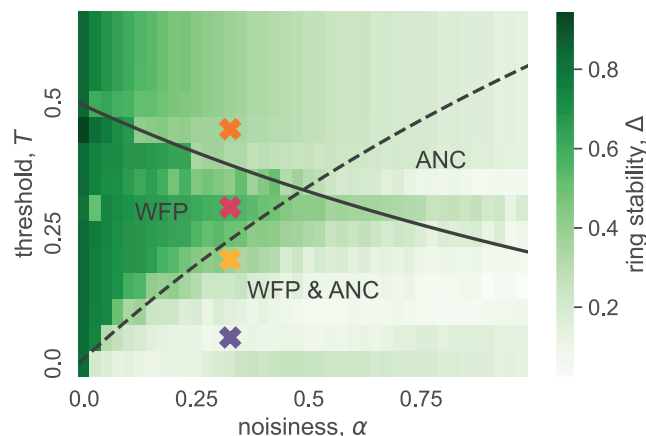

FIG. 3. Bifurcation diagram of WTM on noisy ring lattices matches well with the observed ring stability of truncated contagion maps. We plot the critical thresholds of WFP [solid line, Eq. (4)] and ANC [dashed line, Eq. (5)] vs the noisiness α. These curves divide the (T, α) parameter space into four qualitatively different contagion regimes: exclusively WFP, exclusively ANC, WFP and ANC, and neither. We compare the theoretical bifurcation diagram with the empirically observed ring stability Δ in truncated contagion maps with $s = 20$ of noisy ring lattices of size $N = 400$ with geometric degree $d^{(G)} = 40$ and varying non-geometric degrees $d^{(NG)} \in [0, 40]$ such that their noisiness $\alpha \in [0, 1]$. We find that only in the parameter regime of exclusive WFP, the ring stability Δ is high, indicating a good recovery of the underlying manifold's topology. We highlight the four parameter choices that we show in Figs. 4 and 5 with crosses.

different contagion regimes: exclusively WFP, exclusively ANC, WFP and ANC, and neither. As we will show, these regimes generalize from full contagion maps to truncated contagion maps. In the following, we investigate four parameter choices in more detail. Specifically, we summarize the Watts threshold model contagions for thresholds $T \in \{0.05, 0.2, 0.3, 0.45\}$ on a noisy ring lattice with $N = 400$ nodes and noisiness of $\alpha = d^{(NG)}/d^{(G)} = 2/6 = 1/3$ (see Fig. 4).

In Fig. 4(a), we show the size $q(t)$ of the contagion over time (i.e., the number of nodes that are active). First, we note that for $T \in \{0.05, 0.2, 0.3\}$, a global cascade is triggered as $q(t) \to 1$, whereas for $T = 0.45$, the contagion size remains constant. Second, a smaller threshold T results in a quicker contagion spread which, in turn, yields a smaller time until a global cascade is reached. By comparing the chosen thresholds with the critical thresholds of WFP [Eq. (4)] and ANC [Eq. (5)], we observe that $T = 0.05$ and $T = 0.2$ fall in the regime in which WFP and ANC occur, $T = 0.3$ falls in the regime in which exclusively WFP occurs, and $T = 0.45$ results in no contagion (see colored crossed in Fig. 3).

To study whether the contagion spreads along the manifold or independent of it, we investigate along which type of edges the contagion spreads predominantly. In Fig. 4(b), we show the number $E(t)$ of activating edges (i.e., edges incident to a node that is active at t and a node that becomes active at $t + 1$). We distinguish between geometric edges (solid lines) and non-geometric edges (dashed lines). We observe that for thresholds $T \in \{0.05, 0.2\}$, the contagion spreads quickly and to a similar extent along geometric and non-geometric edges. For a threshold of $T = 0.3$ in contrast, we find a slow contagion process. We also find that initially the contagion process is strongly dominated by contagion along geometric edges (see inlay). After time $t \approx 20$, however, the influence of non-geometric edges increases. Our observation indicates that for slow contagion processes, the contagion initially spreads along the manifold as WFP. Yet, a spreading along non-geometric edges becomes more important over time. This highlights that the derived critical thresholds [Eqs. (4) and (5)] are strictly valid only for the early contagion steps and become less appropriate the longer the contagion progresses. These observations indicate that a truncation choice that removes these late-stage contagion steps might improve the contagion map. In this case, for example, a truncation parameter $s \in [20, 60]$ might be appropriate. As a simple, heuristic choice, we suggest $s^* = N/10 = 40$, which we will explore in more detail in Sec. IV. As we will show in Sec. IV, such truncations allow us to improve the recovery of topological features with truncated contagion maps in comparison with full contagion maps.

IV. TOPOLOGICAL DATA ANALYSIS OF TRUNCATED CONTAGION MAPS

In this section, we analyze the topological properties of contagion maps on noisy ring lattices. This allows us to identify parameter regimes in which the contagion processes follow the underlying manifold. In particular, we compare the behavior of truncated contagion maps with that of full contagion maps.

In Fig. 6, we study contagion maps for noisy ring lattices with $N = 400$ nodes and noisiness of $\alpha = d^{(NG)}/d^{(G)} = 2/6 = 1/3$. Each column represents a distinct threshold $T \in \{0.05, 0.2, 0.3, 0.45\}$,
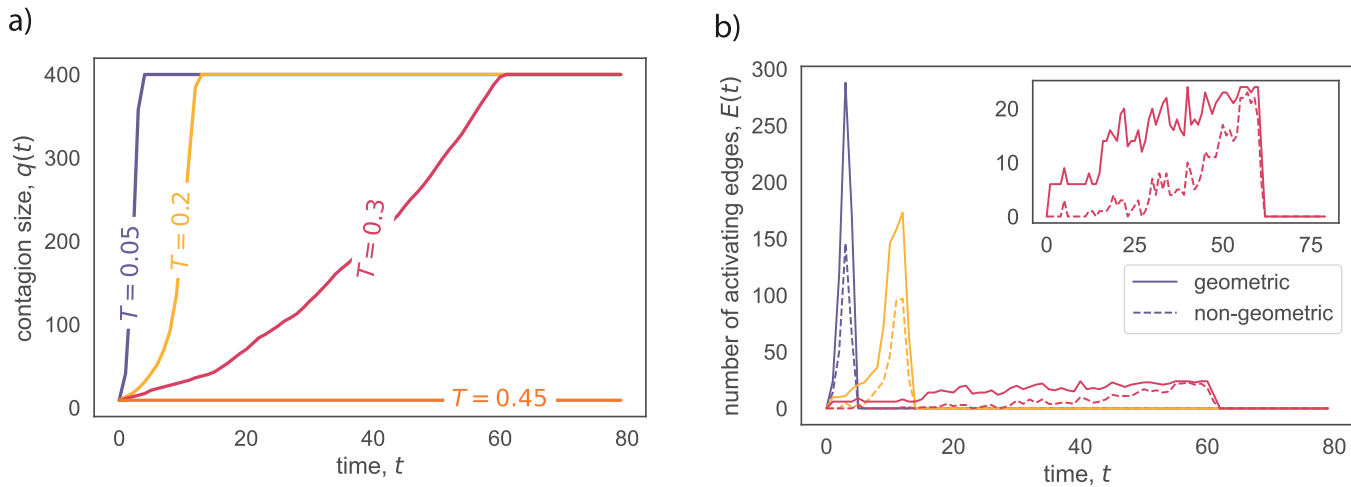
FIG. 4. The temporal development of the contagion is strongly dependent on the threshold T. (a) The size $q(t)$ of the contagion is growing over time for thresholds $T \in \{0.05, 0.2, 0.3\}$ and reaches $q(t) = 1$, which indicates a global contagion (i.e., all nodes are active). The size $q(t)$ of the contagion remains constant for $T = 0.45$, indicating that no cascade is triggered. (b) We show the number $E(t)$ of edges involved in the activation process and distinguish between geometric edges (solid lines) and non-geometric edges (dashed lines). The inlay shows a zoom onto the contagion with threshold $T = 0.3$ in which we observe that for the early (i.e., $t < 20$) contagion steps almost no non-geometric edges are involved.

which corresponds to the example thresholds in Fig. 4 and distinct regimes in the bifurcation analysis of the critical thresholds [Eqs. (4) and (5)]. In the top row, we show the noisy ring lattice with nodes placed at their locations $\mathbf{w}^{(i)} = (\cos(2\pi i/N), \sin(2\pi i/N))$. The color of each node indicates the node's activation time during one realization of the WTM in which the green nodes are the seed nodes. For all other nodes, the color map ranges from blue (small activation time) to yellow (large activation time). Gray nodes never adopt the contagion. As expected, we observe that for two smallest thresholds $T \in \{0.05, 0.2\}$, the contagion quickly spreads to the whole network and that for the largest threshold $T = 0.45$, no contagion beyond the seed nodes occurs. For intermediary threshold $T = 0.3$, however, we observe that the contagion spreads approximately along the circular manifold.

In the remaining rows of Fig. 6, we show the point clouds $\mathbf{x}^{(i)}$ obtained from contagion maps. Specifically, rows two, three, and four show truncated contagion maps for $s = 10$ steps, $s = 40$ steps, and $s = 60$ steps, respectively, and the last row shows the full contagion map, which is equivalent to $s = N$. To visualize the N-dimensional point clouds, we use a two-dimensional projection through principal component analysis (PCA). Similar to the networks in the top row, we color each point with the activation time under one realization of the WTM. Yet, as we truncate the contagion after s steps a node might be gray (i.e., never activated) either because a steady state is reached or because the contagion stopped earlier. Therefore, the number of gray (i.e., never activated) nodes varies with the number s of considered contagion steps, although the seed nodes were identical.

In general, we find that the contagion maps $\mathbf{x}^{(i)}$ for threshold $T = 0.3$, which is in a regime in which there is exclusively WFP

predicted, most closely resemble the two-dimensional ring topology of the original network (up to rotation). Furthermore, in this regime, the truncation parameter s has a strong influence on the obtained embedding. For a small number $s = 10$ of contagion steps, the ring structure is perturbed and appears folded. Increasing the number of contagion steps to $s = 20$ makes the ring structure more clearly visible. Increasing the number s of contagion steps further, however, results in a "broadening" of the circular point clouds. Our observation suggests that for the threshold $T = 0.3$, a truncated contagion map might better recover the underlying circular manifold than the full contagion map.

After the bifurcation analysis in Sec. III and the qualitative investigation in Fig. 5, we now quantify the recovery of the manifold's topological features with persistent homology. In particular, we know that the underlying manifold $\mathcal{M}$ of a noisy ring lattice is the unit circle. By comparing how closely the persistent homology of the contagion maps resembles the topological properties of the unit circle, we can quantify the recovery of its topological properties. A unit circle has one circular hole and thus its first Betti number is $b_1 = 1$. Therefore, we compute the ring stability Δ as a quantification of how "ring-like" a contagion map is. In Fig. 6(a), we show the ring stability Δ of contagion maps on noisy ring lattices of size $N = 400$ and noisiness $\alpha = 1/3$. We vary the threshold $T \in [0, 0.6]$ and highlight the critical thresholds $T^{(\text{ANC})} = 0.25$ and $T^{(\text{WFP})} = 0.375$ [Eqs. (4) and (5)] as vertical dotted lines. All contagion maps have their largest ring stability Δ in the regime between these thresholds (i.e., when only WFP is possible). Furthermore, the truncated contagion maps with $s = 10$ and $s = 20$ have a higher ring stability Δ than the full contagion map, which matches our observations in Fig. 5 that an appropriate truncation of the contagion map
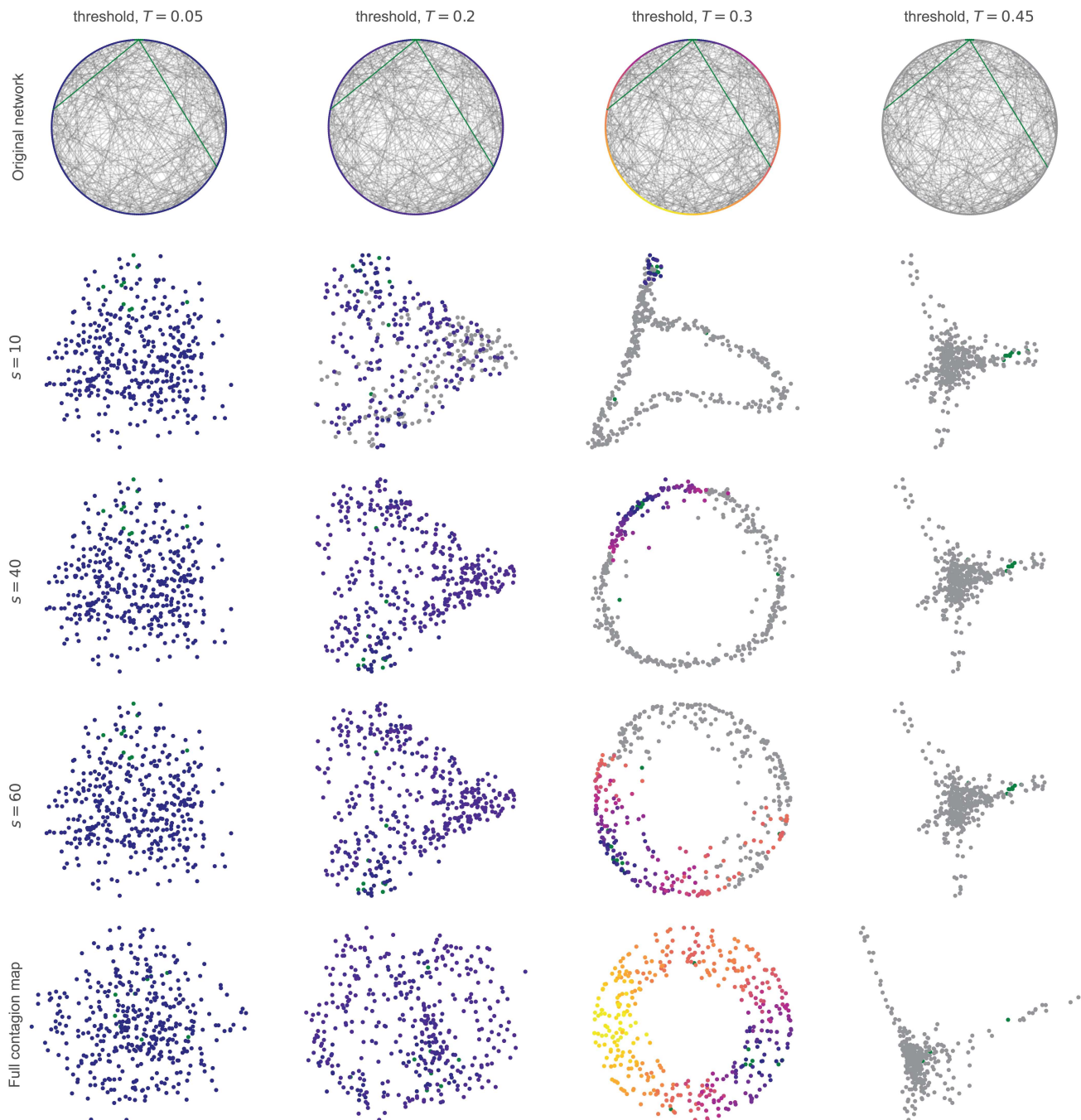
FIG. 5. Truncated contagion maps applied to noisy ring lattices can recover the ring manifold for appropriate thresholds T. We applied contagion maps to noisy ring lattices with $N = 400$ nodes degrees $= (d^{(G)}, d^{(NG)}) = (6, 2)$. In each column, we show the results for one of the four thresholds $T \in \{0.05, 0.2, 0.3, 0.45\}$. (top row) The network with nodes embedded on the ring manifold. Each node's color indicates the activation time from small (blue) to large (yellow) in one realization of the WTM. Green nodes are seed nodes and gray nodes are inactive at the steady state (other rows). We show two-dimensional projections of the N-dimensional contagion maps. A node's color indicates its activation time as in the top row. The rows represent truncated contagion maps with $s = 10$, $s = 40$, and $s = 60$, and a full contagion map, respectively. We observe that with threshold $T = 0.3$, the contagion maps most closely resemble the circular embedding manifold. In this regime, intermediate step numbers $s = 40$ lead to the best recovery of the circular manifold (up to rotation).
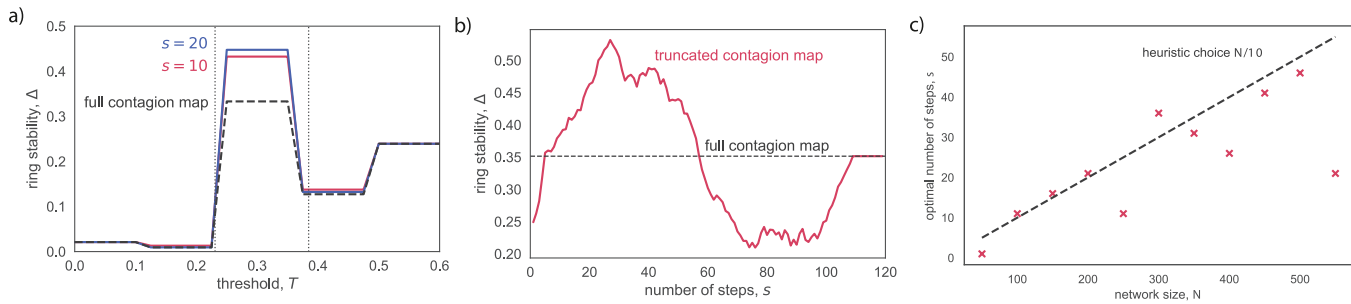
FIG. 6. Truncated contagion maps enable an improved recovery of topological features, in comparison with a full contagion map. We quantify the recovery of the topological features by computing the ring stability Δ of contagion maps on noisy ring lattices with $N = 400$. (a) For all contagion maps, the ring stability is the largest in the threshold regime $\frac{\alpha}{1+\alpha} < T < \frac{1}{2+2\alpha}$ in which WFP dominates the spreading process. The truncated contagion maps with $s = 10$ (red line) and $s = 20$ (blue line) have larger ring stability Δ than the full truncation map (dashed black line). (b) The ring stability Δ varies with the number s of included contagion steps. For small number s of contagion steps, the truncated contagion map results in improved ring stability Δ in comparison to the full contagion map (dashed black line). For $s > 60$, when ANC occurs, the ring stability Δ diminishes. As expected the performance of the full contagion map matches the truncated contagion map for large s. Results for the threshold $T = 0.3$ (c) We investigate the truncation parameter s^* that yields the best recovery of the ring manifold for noisy ring lattice networks of sizes $N \in [50, 550]$. We find that the optimal truncation increases approximately linearly with network size N, which suggests a simple heuristic choice of $s^* = N/10$ (dashed line).

may improve the recovery of the ring-like topology. In other threshold regimes, the ring stability Δ is almost indistinguishable between truncated and full contagion maps.

In Fig. 6(b), we investigate the influence of the number s of contagion steps on the ring stability Δ of contagion maps with a threshold of $T = 0.3$. We find that for a large number $s \to N$ of steps, the ring stability is the same as for the full contagion map (dashed line), which matches our theoretical expectation (see Sec. II D). For a small number $s < 60$ of steps, the truncated contagion map has larger ring stability Δ than the full contagion map, whereas for $s > 60$ steps, the full contagion map outperforms the truncated contagion map. We explain this behavior as a trade-off between data quality and data quantity. The early-stage contagions more closely follow the underlying manifold, yielding better embeddings for the truncated contagion maps. Yet, full contagions have more data as less entries in the contagion map have infinite activation times.

The dependence of the ring stability Δ on the truncation parameter s raises the question whether the optimal choice varies with the network size N. In Sec. III, we suggested $s^* = N/10$ as a simple, heuristic choice because, in this regime, the contagion is predominantly spreading along then manifold. In Fig. 6(c), we compare this heuristic with the actual optimal s^* for noisy ring-lattice networks of sizes $N \in [50, 550]$. We find that the optimal truncation parameter s^* is increasing with the network size N and is in a decent agreement with the suggested heuristic.

Thus far, we fixed the noisiness α of the networks. In Fig. 3, we compare the analytically predicted bifurcation diagram with the ring stability Δ of truncated contagion maps with $s = 20$ steps for various thresholds $T \in [0, 0.6]$ for noisy ring lattices with $N = 400$ and various noisiness values $\alpha \in [0, 1]$. Parameter combinations (α, T) with large ring stability Δ closely match the regime in which the bifurcation analysis predicts exclusively WFP. This indicates that in regimes that are dominated by WFP, the truncated contagion map consistently recovers the underlying topology of the unit circle, whereas in other parameter regimes this is not possible.

V. COMPUTATIONAL COMPLEXITY OF (TRUNCATED) CONTAGION MAPS

Taylor et al.[21] established that the typical computational cost of full contagion maps is $\mathscr{O}(N^2 d)$, where N is the number of nodes in the input network and d the average degree of its nodes. As truncated contagion maps omit some activation steps, we anticipate that less computational time is needed than for the full contagion map.

We empirically verify the reduction of computational cost for contagion maps with thresholds $T = 0.3$ on noisy ring lattices with noisiness $\alpha = 1/3$ and consider $s = 20$ steps in the truncated contagion maps, which is a parameter combination that yields a good recovery of the ring manifold (see Sec. IV). We vary the network sizes $N \in \{32, \ldots, 10^4\}$ to identify the computational complexity [see Fig. 7(a)]. We find that for small network sizes $N \leq 100$, the computation time δt of truncated and full contagion maps are indistinguishable because a steady state is reached before a truncation after $s = 20$ influences the contagions. Yet, for networks of sizes $N > 100$, the truncated contagion map is quicker to compute than the full contagion map. Furthermore, we identify that the gap between computational costs widens with increasing network size N. To obtain estimates of the computational complexity of the form $\delta t = \zeta N^\gamma$, we fit its linearized form $\log(\delta t(\log(N))) = \log \zeta + \gamma \log(N)$ to the logarithm of the data. We find that the computational time of the full contagion map and the truncated contagion map scale superquadratic ($\delta t \propto N^{2.24}$) and subquadratic ($\delta t \propto N^{1.79}$) with the problem size N, respectively, yielding an improved scaling of almost $\sqrt{N}$ for the truncated contagion maps. The reduced computational complexity of the truncated contagion maps can be understood by investigating the time until a steady state is reached. If comparing networks of the same class (i.e., similar characteristics, such as characteristic path length and mean degree), in larger networks the time until a global cascade has been reached tends to be longer than in smaller networks. Thus, a restriction to a fixed number s of steps by the truncated contagion maps omits a larger number of
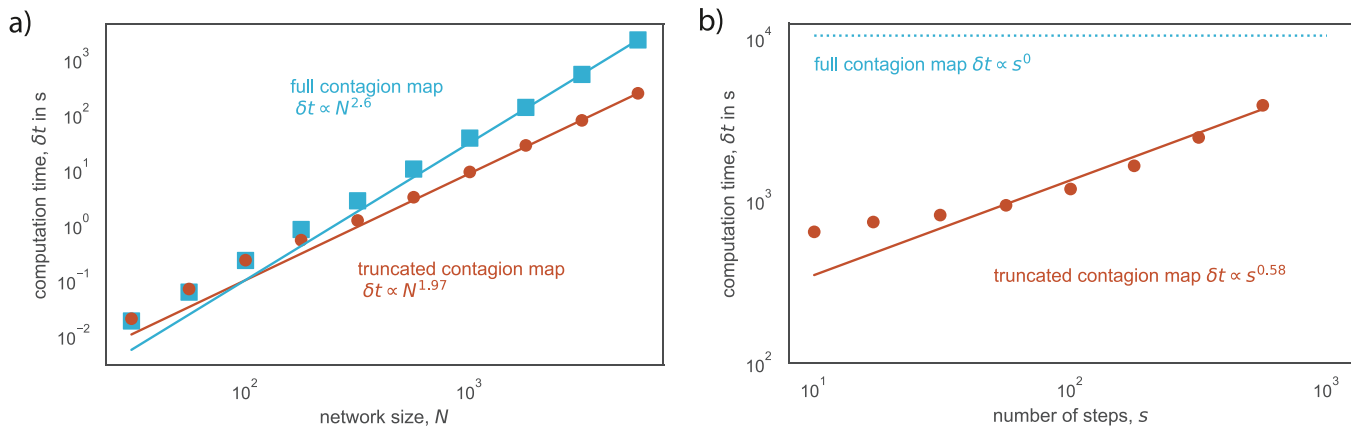
FIG. 7. Truncated contagion maps have a smaller computational complexity than full contagion maps. (a) We show the computation time δt for contagion maps of noisy ring lattices of various sizes $N \in \{32, 10^4\}$, noisiness $\alpha = d^{(NG)}/d^{(G)} = 6/2$, threshold $T = 0.3$. The truncated contagion map (red disks; $s = 20$ steps considered) is for small network sizes $N < 100$ of similar computational cost as the full contagion map (blue squares) but is considerably faster for large networks. We estimate the computational complexity with least-squares fits of a linear equation to the logarithm of the computation time. The computation time δt for full and truncated contagion maps scale super-quadratic ($\gamma \approx 2.24$) and sub-quadratic ($\gamma \approx 1.79$) with the network size N, respectively. (b) We show the computation time δt for contagion maps of noisy ring lattices of size $N = 10\,000$, noisiness $\alpha = d^{(NG)}/d^{(G)} = 6/2$, and threshold $T = 0.3$. We vary the number s of steps and identify an approximately $\gamma \approx 0.58$, with truncated contagion maps approaching the computational time of full contagion maps for $s \to N$.

steps for larger networks. This increasing omission reduces the computational time for larger networks, leading to a substantial reduction of the computation time for truncated contagion maps.

We further investigate the influence of the number s of steps on the computational time δt in Fig. 7(b). Empirically, we find that the computational time grows sublinear with s such that $\delta t \propto s^{0.58}$. As expected, for large number s of steps, the computational time of the truncated contagion maps approaches the computational time of the

full contagion maps, highlighting the need for truncation to reduce the computational complexity.

VI. CONTAGION MAPS OF SINGLE-CELL TRANSCRIPTOMICS DATA

In this section, we investigate a single-cell transcriptomics dataset of growing mouse oocytes,[49,50] in which the oocyte undergo several key stages to generate mature oocytes. Specifically, Gu *et al.*
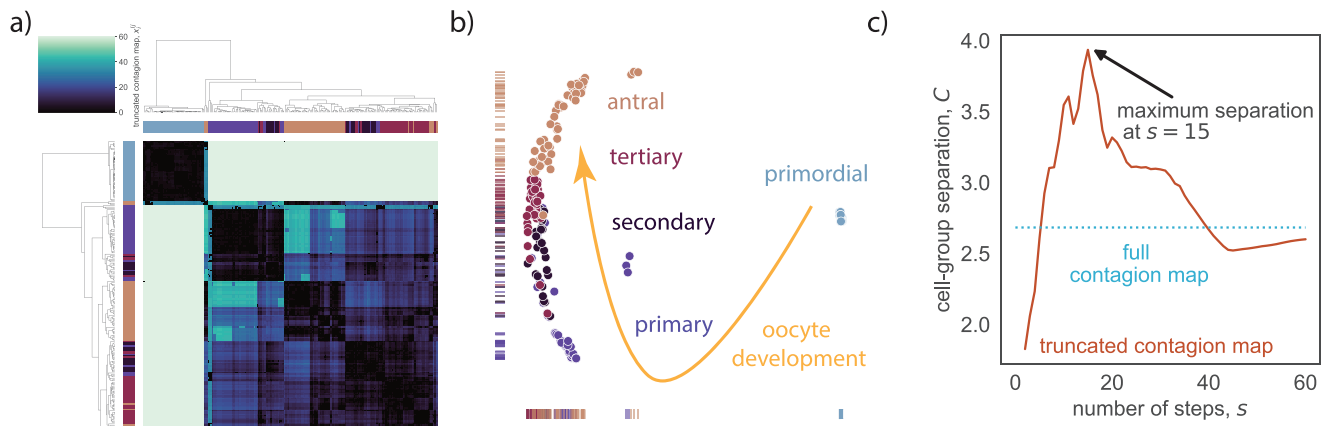

FIG. 8. Truncated contagion map uncovers differentiation trajectory in single-cell transcriptomics data of mouse oocytes. (a) We show a truncated contagion map $\mathbf{x}^{(i)}$ with threshold $T = 0.3$ and truncation $s = 10$ for a cell-similarity network obtained from single-cell transcriptomics data.[49] The UPGMA method with Euclidean distance was used for the hierarchical clustering. (b) A two-dimensional embedding of the truncated contagion map reveals the oocyte developmental trajectory from primordial oocytes over primary, secondary, and tertiary oocytes to antral oocytes (color indicates cell identity). (c) The cell-group separation C is the highest for a truncation parameter $s = 15$ and outperforms a full contagion map (dashed, horizontal line). As expected, for large number s of steps, the performance of the truncated contagion map approaches the performance of the full contagion map.

measure gene expression in five successive stages: primordial follicle, primary oocytes, secondary oocytes, tertiary oocytes, and antral oocytes. The raw gene-expression matrix **Count** $\in \mathbb{N}^{(c \times g)}$ in which entry **Count**$_{i,j}$ indicates the strength of expression of gene i in cell j is available from GEO under GSE114822. For further analysis, we use SCANPY[51] and follow current best practices.[52] For example, we filter out 5937 genes that are detected in less than 3 cells (for details, see online material). We construct a cell-similarity graph as a k-nearest-neighbor graph $G = (V, E)$ in which the $N = 223$ nodes represent the cells $V \in \{1, \ldots, c\}$ and each node is connected to its $k = 20$ nearest neighbors in gene-expression space.

Then, we construct a truncated contagion map with threshold $T = 0.3$ and truncation $s = 15$ from the cell-similarity graph (see Fig. 8). The contagion map $\mathbf{x}^{(i)}$ in Fig. 8(a) has a block-diagonal structure in which the non-growing oocytes (i.e., primordial) are clearly separated from the growing oocyte (primary, secondary, and tertiary) and the fully-grown oocyte (antral). A two-dimensional PCA projection of the contagion map $\mathbf{x}^{(i)}$ in Fig. 8(b) reveals the oocyte development trajectory in gene-expression space. The primary oocytes are developing into secondary and tertiary oocytes and, finally, into antral oocytes. The separation of the developmental stages is not sharp but rather a continuum, yet a pseudo-temporal ordering that matches the cell identities is observable.

To investigate the performance of the truncated contagion maps in separating the different cell types, we compute the *cell-group separation C*, as the ratio of the median distance between cells of different cell types to the median distance between cells of the same cell type. We investigate this ratio C in dependence on the number s of steps in Fig. 8(c). We find that for most truncation parameters s, the cell-group separation C is improved for the truncated contagion maps vs the full contagion map (dashed line). The maximum separation (i.e., best performance) is reached for $s = 15$. We note that this optimal choice is close to the heuristic choice of $s^* = N/10 \approx 22$.

Our study on the single-cell transcriptomics data shows that contagion maps can reveal a manifold structure in a high-dimensional gene-expression space by constructing a cell-similarity graph. In particular, we observe that primordial oocytes are clearly separated from developing oocytes and within the developing oocytes, we observe an ordering in gene-expression space.

VII. CONCLUSION

In this work, we introduced truncated contagion maps, an extension of contagion maps that has improved computational complexity and are thus computationally less expensive. This is achieved by truncating (i.e., stopping) the contagion processes before they reach a steady state, a technique that is well-established in the context of random walks.[25] Furthermore, we demonstrate by quantifying the persistent homology that, for appropriate parameter choices, truncated contagion maps might improve the recovery of the embedding manifold of noisy geometric networks. We explain this behavior with the dominance of spreading along the manifold in early steps of a contagion, whereas at later steps, the spread along non-geometric edges increases. Finally, we demonstrated that contagion maps are a manifold-learning technique that can reveal low-dimensional structures in a cell-similarity network constructed from single-cell transcriptomics data.

Our findings suggest various avenues for further studies. For example, it is an open question whether these truncation approaches can also be extended to the simplicial complex variant of the contagion maps,[23] which is particularly relevant as higher-order interactions are important in cellular interactions.[53,54] Furthermore, we hypothesize that a subsampling approach with selected seeding, as used in the k-means++ clustering algorithm,[55] or landmark selection, as used in L-Isomap,[56] might further reduce the computational complexity of contagion maps. The establishment of more sophisticated heuristics for the choice of the truncation parameter s will be fruitful, as its optimal choice is likely to depend on the structure of the analyzed network in a complex, non-linear way, which is not reflected in the simple heuristic we proposed in this paper.

AUTHOR DECLARATIONS
Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Florian Klimm: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review and editing (equal).

DATA AVAILABILITY

The data that support the findings of the study are openly available in a public repository that does not issue DOIs. PYTHON code to reproduce the results in this paper are available in GitHub at https://github.com/floklimm/contagionMap, Ref. 57.

REFERENCES

[1]W. O. Kermack and A. G. McKendrick, "Contributions to the mathematical theory of epidemics–I. 1927," Bull. Math. Biol. **53**, 33–55 (1991).
[2]M. J. Keeling and P. Rohani, *Modeling Infectious Diseases in Humans and Animals* (Princeton University Press, 2011).
[3]F. Schlosser, B. F. Maier, O. Jack, D. Hinrichs, A. Zachariae, and D. Brockmann, "COVID-19 lockdown induces disease-mitigating structural changes in mobility networks," Proc. Natl. Acad. Sci. U.S.A. **117**, 32883–32890 (2020).
[4]T. W. Valente, "Social network thresholds in the diffusion of innovations," Social Networks **18**, 69–89 (1996).
[5]N. A. Christakis and J. H. Fowler, "Social contagion theory: Examining dynamic social networks and human behavior," Stat. Med. **32**, 556–577 (2013).
[6]D. Guilbeault, J. Becker, and D. Centola, "Complex contagions: A decade in review," in *Complex Spreading Phenomena in Social Systems* (Springer, 2018), pp.3–25.
[7]M. Granovetter, "Threshold models of collective behavior," Am. J. Sociol. **83**, 1420–1443 (1978).

[8]M. Wiedermann, E. K. Smith, J. Heitzig, and J. F. Donges, "A network-based microfoundation of Granovetter's threshold model for social tipping," Sci. Rep. **10**, 1–10 (2020).

[9]S.-W. Oh and M. A. Porter, "Complex contagions with timers," Chaos **28**, 033101 (2018).

[10]J. Borge-Holthoefer, R. A. Banos, S. González-Bailón, and Y. Moreno, "Cascading behaviour in complex socio-technical networks," J. Complex Networks **1**, 3–24 (2013).

[11]M. A. Porter and J. P. Gleeson, "Dynamical systems on networks: A tutorial," arXiv:1403.7663 (2014).

[12]J. P. Gleeson, "Cascades on correlated and modular random networks," Phys. Rev. E **77**, 046117 (2008).

[13]D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," Nature **393**, 440–442 (1998).

[14]D. Brockmann and D. Helbing, "The hidden geometry of complex, network-driven contagion phenomena," Science **342**, 1337–1342 (2013).

[15]S. A. Marvel, T. Martin, C. R. Doering, D. Lusseau, and M. E. Newman, "The small-world effect is a modern phenomenon," arXiv:1310.2636 (2013).

[16]A. J. Ammerman and L. L. Cavalli-Sforza, "Measuring the rate of spread of early farming in Europe," Man **6**, 674–688 (1971).

[17]M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, "Determination of reaction coordinates via locally scaled diffusion map," J. Chem. Phys. **134**, 124116 (2011).

[18]A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis, and P. G. Debenedetti, "Nonlinear dimensionality reduction in molecular simulation: The diffusion map approach," Chem. Phys. Lett. **509**, 1–11 (2011).

[19]R. R. Coifman, I. G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler, "Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems," Multiscale Model. Simul. **7**, 842–864 (2008).

[20]H. Kim, R. Eykholt, and J. Salas, "Nonlinear dynamics, delay times, and embedding windows," Phys. D **127**, 48–60 (1999).

[21]D. Taylor, F. Klimm, H. A. Harrington, M. Kramár, K. Mischaikow, M. A. Porter, and P. J. Mucha, "Topological data analysis of contagion maps for examining spreading processes on networks," Nat. Commun. **6**, 1–11 (2015).

[22]B. I. Mahler, "Contagion dynamics for manifold learning," arXiv:2012.00091 (2020).

[23]B. U. Kilic and D. Taylor, "Simplicial cascades are orchestrated by the multidimensional geometry of neuronal complexes," arXiv:2201.02071 (2022).

[24]B. I. Mahler, "Analysis of contagion maps on a class of networks that are spatially embedded in a torus," SIAM J. Appl. Math. **81**, 1416–1440 (2021).

[25]H. Tong, C. Faloutsos, and J.-Y. Pan, "Fast random walk with restart and its applications," in *Sixth International Conference on Data Mining (ICDM'06)* (IEEE, 2006), pp. 613–622.

[26]T. H. Kim, K. M. Lee, and S. U. Lee, "Generative image segmentation using random walks with restart," in *European Conference on Computer Vision* (Springer, 2008), pp. 264–275.

[27]J. Shun, F. Roosta-Khorasani, K. Fountoulakis, and M. W. Mahoney, "Parallel local graph clustering," arXiv:1604.07515 (2016).

[28]N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, and H. A. Harrington, "A roadmap for the computation of persistent homology," EPJ Data Sci. **6**, 1–38 (2017).

[29]G. Bianconi, "Interdisciplinary and physics challenges of network theory," Europhys. Lett. **111**, 56001 (2015).

[30]C. Bick, E. Gross, H. A. Harrington, and M. T. Schaub, "What are higher-order networks?," arXiv:2104.11329 (2021).

[31]B. J. Stolz, H. A. Harrington, and M. A. Porter, "Persistent homology of time-dependent functional networks constructed from coupled time series," Chaos **27**, 047410 (2017).

[32]C. Giusti, R. Ghrist, and D. S. Bassett, "Two's company, three (or more) is a simplex," J. Comput. Neurosci. **41**, 1–14 (2016).

[33]B. J. Stolz, H. Harrington, and M. A. Porter, see https://ssrn.com/abstract=2843662 for "The Topological 'Shape' of Brexit."

[34]M. Feng and M. A. Porter, "Persistent homology of geospatial data: A case study with voting," SIAM Rev. **63**, 67–99 (2021).

[35]M. Feng and M. A. Porter, "Spatial applications of topological data analysis: Cities, snowflakes, random structures, and spiders spinning under the influence," Phys. Rev. Res. **2**, 033426 (2020).

[36]B. J. Stolz, J. Tanner, H. A. Harrington, and V. Nanda, "Geometric anomaly detection in data," Proc. Natl. Acad. Sci. U.S.A. **117**, 19664–19669 (2020).

[37]V. Salnikov, D. Cassese, R. Lambiotte, and N. S. Jones, "Co-occurrence simplicial complexes in mathematics: Identifying the holes of knowledge," Appl. Network Sci. **3**, 1–23 (2018).

[38]A. E. Sizemore, E. A. Karuza, C. Giusti, and D. S. Bassett, "Knowledge gaps in the early growth of semantic feature networks," Nat. Hum. Behav. **2**, 682–692 (2018).

[39]H. Edelsbrunner and J. Harer, "Persistent homology—A survey," Contemp. Math. **453**, 257–282 (2008), https://www.maths.ed.ac.uk/ v1ranick/papers/edelhare.pdf.

[40]H. Edelsbrunner and D. Morozov, "Persistent homology: Theory and practice" in *European Congress of Mathematics Kraków*, 2–7 July, 2012 (EMS Press, 2012).

[41]S. Aldridge and S. A. Teichmann, "Single cell transcriptomics comes of age," Nat. Commun. **11**, 1–4 (2020).

[42]D. Lähnemann, J. Köster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel, A. Mahfouz *et al.*, "Eleven grand challenges in single-cell data science," Genome Biol. **21**, 1–35 (2020).

[43]K. R. Moon, J. S. Stanley III, D. Burkhardt, D. van Dijk, G. Wolf, and S. Krishnaswamy, "Manifold learning-based methods for analyzing single-cell RNA-sequencing data," Curr. Opin. Syst. Biol. **7**, 36–46 (2018).

[44]M. Barthélemy, "Spatial networks," Phys. Rep. **499**, 1–101 (2011).

[45]M. S. Granovetter, "The strength of weak ties," Am. J. Sociol. **78**, 1360–1380 (1973).

[46]D. J. Watts, "A simple model of global cascades on random networks," in *The Structure and Dynamics of Networks* (Princeton University Press, 2011), pp. 497–502.

[47]M. A. Porter and J. P. Gleeson, *Dynamical Systems on Networks*, Frontiers in Applied Dynamical Systems: Reviews and Tutorials Vol. 4 (Springer, 2016).

[48]C. Tralie, N. Saul, and R. Bar-On, "Ripser.py: A lean persistent homology library for python," J. Open Source Software **3**, 925 (2018).

[49]C. Gu, S. Liu, Q. Wu, L. Zhang, and F. Guo, "Integrative single-cell analysis of transcriptome, DNA methylome and chromatin accessibility in mouse oocytes," Cell Res. **29**, 110–123 (2019).

[50]H. Zhang, M. Esposito, M. G. Pezet, J. Aryaman, W. Wei, F. Klimm, C. Calabrese, S. P. Burr, C. H. Macabelli, C. Viscomi, M. Saitou, M. R. Chiaratti, J. B. Stewart, N. Jones, and P. F. Chinnery, "Mitochondrial DNA heteroplasmy is modulated during oocyte development propagating mutation transmission," Sci. Adv. **7**, eabi5657 (2021).

[51]F. A. Wolf, P. Angerer, and F. J. Theis, "Scanpy: Large-scale single-cell gene expression data analysis," Genome Biol. **19**, 1–5 (2018).

[52]M. D. Luecken and F. J. Theis, "Current best practices in single-cell RNA-seq analysis: A tutorial," Mol. Syst. Biol. **15**, e8746 (2019).

[53]S. Klamt, U.-U. Haus, and F. Theis, "Hypergraphs and cellular networks," PLoS Comput. Biol. **5**, e1000385 (2009).

[54]F. Klimm, C. M. Deane, and G. Reinert, Hypergraphs for predicting essential genes using multiprotein complex data," J. Complex Networks **9**, cnaa028 (2021).

[55]S. Vassilvitskii and D. Arthur, "k-means++: The advantages of careful seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (ACM, 2006), pp. 1027–1035.

[56]V. De Silva and J. B. Tenenbaum, "Global versus local methods in nonlinear dimensionality reduction," in *NIPS'02: Proceedings of the 15th International Conference on Neural Information Processing Systems* (MIT Press, 2002), Vol. 15, pp. 705–712; available at https://dl.acm.org/doi/10.5555/2968618.2968708.

[57]GitHub Repository available at https://github.com/floklimm/contagionMap.