Contagions and Manifolds



Barbara Ilse Mahler
Hertford College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2019

Abstract

A spreading process on a network can be influenced by the network's underlying spatial structure. In particular, if a network's nodes are embedded in a manifold and connected by both 'geometric edges', which respect the geometry of the underlying manifold, and 'non-geometric edges' that are not constrained by that geometry, a threshold contagion can either propagate as a wave along the network or jump via long non-geometric edges to remote areas of the network. In this thesis, we study so-called 'contagion maps', which are defined via spreading dynamics on such 'noisy geometric networks'. We take two complementary perspectives: In the first part of this thesis, we use contagion maps to examine the spreading of a threshold contagion on a class of networks that are embedded in a torus. We analyse the dimensionality, geometry, and topology of point clouds produced by contagion maps to examine qualitative properties of this spreading process. We conduct a bifurcation analysis to identify a region in parameter space in which the contagion propagates predominantly via wavefront propagation and find that this region aligns with our computational analysis of the point clouds produced by contagion maps. We also consider different probability distributions for constructing non-geometric edges — reflecting different decay rates with respect to the distance between nodes in the underlying manifold — and examine how qualitative properties of the spreading dynamics are affected by these choices. Moreover, we incorporate an approach for calibrating persistent-homology barcodes of different scales. This approach is applicable in various other applications that use persistent homology, and the Wasserstein distance between barcodes in particular. In the second part of this thesis, we test contagion maps as a manifold learning tool on a variety of data sets. In particular, we compare the performance of contagion maps on these data sets to that of Isomap and find that, under certain conditions, contagion maps are able to reliably detect underlying manifold structure in noisy data, when Isomap is prone to noise-induced error. Our work consolidates contagion maps both as a tool for investigating spreading behaviour on spatial networks and as a technique for manifold learning.

Contents

1 Introduction **1**

1.1 Spreading dynamics on spatial networks 1

1.2 Spreading dynamics for manifold learning 5

1.3 Overview and structure of this thesis 10

2 Background Material from Algebraic Topology **14**

2.1 Simplicial homology 15

2.2 Persistent homology 19

2.3 Barcodes, persistence diagrams, and the Wasserstein distance 22

2.4 Persistent homology on point-cloud data 25

3 Contagions on Noisy Geometric Networks **30**

3.1 Background 30

3.1.1 Noisy geometric networks 30

3.1.2 Contagions on networks 32

3.2 Our network and contagion models 33

3.2.1 Network model 34

3.2.2 Contagion model 37

4 Contagion Maps for Spreading Dynamics **38**

4.1 Methods 38

4.1.1 Contagion maps 39

4.1.2 Geometry . . . 40
4.1.3 Topology . . . 42
4.1.4 Dimensionality . . . 47
4.2 Results . . . 47
4.2.1 Explorations on various small networks . . . 48
4.2.1.1 Non–noisy periodic lattice . . . 48
4.2.1.2 Noisy networks based on a periodic lattice . . . 54
4.2.2 Numerical results for 50-by-50 networks . . . 63
4.2.3 Effect of the distance-decay parameter γ on contagion maps . . . 67
4.2.4 Bifurcation analysis . . . 70
4.2.4.1 Wavefront Propagation (WFP) . . . 71
4.2.4.2 Appearance of New Clusters (ANC) . . . 73
4.3 Conclusions and future work . . . 83

5 Contagion Maps for Manifold Learning 87

5.1 Algorithms . . . 87
5.1.1 Isomap . . . 88
5.1.2 Contagion maps . . . 90
5.1.3 Workflow . . . 91
5.2 Data sets . . . 92
5.3 Results . . . 93
5.3.1 Noisy regularly spaced points from Swiss roll . . . 93
5.3.2 Swiss roll data set from Tenenbaum *et al.* . . . 102
5.3.3 Torus-based networks . . . 108
5.3.3.1 Dimensionality . . . 110
5.3.3.2 Topology . . . 115
5.3.3.3 Geometry . . . 119

5.3.4 Conformation space of the cyclo-octane molecule 121

5.4 Conclusions and future work 127

A Persistent Homology of a Torus with various Choices of Metric 129

A.1 The $d_{2,1}$ metric 130

A.2 The $d_{2,2}$ metric 136

A.3 The $d_{\text{arc},1}$ metric 138

A.4 The $d_{\text{arc},2}$ metric 140

Bibliography 142

Chapter 1

Introduction

Many complex systems involve discrete sets of entities that may interact with and be related to one another in some way. The structure and behaviour of such systems can often be captured in the form of point clouds or networks. Point clouds and networks have topological structure, suggesting that it should be insightful to study such systems with an algebraic-topological approach. We will use topological methods, in particular persistent homology, as well as more established statistical techniques to (1) study spreading on spatially embedded networks, and (2) infer underlying structure in point-cloud and network data.

1.1 Spreading dynamics on spatial networks

Spreading dynamics are ubiquitous in many areas of life, including social settings and biological processes. The spreading of a contagious disease or of an idea between people are two obvious examples, but there are various other phenomena that give rise to spreading dynamics in less obvious ways [59].

The spreading of real-world contagions is often guided by the geometry of the underlying domain [22, 40, 67, 80]. One example is given by contagious diseases that, historically, spread gradually along part of the earth's surface. Similarly, in a world where means of transport and communication are limited, information typically dis-

seminates via actors who are physically close. In such a world, contagions often propagate as a wavefront, passing between geometrically close entities. However, with modern transport and communication technology, there are now many scenarios where — even in the presence of a well-defined underlying geometry, such as the spherical surface of the earth — a contagion can also spread via connections that are not intrinsically geometric [1, 8]. Examples of such scenarios include the spreading of an infectious disease via passengers traveling on a long-distance flight and the dissemination of information via internet communication. In these examples, a contagion jumps across space to distant locations, rather than following the geometry of an underlying domain.

One way to study such phenomena is to consider contagion models on networks that are embedded in some underlying geometric space [62, 64]. In particular, one can consider networks that have both *geometric edges* that respect the geometry of the underlying space, in the sense that they can only connect nodes that are close according to the space's metric, and *non-geometric* edges, which are not constrained by the underlying geometry and can connect nodes across a long range. Following terminology from [71], we refer to such networks as *noisy geometric networks*. It is interesting and important to ask [49, 59] what propagation pattern(s) a contagion follows and how much such patterns are influenced by the structure of the underlying space. Two fundamental spreading mechanisms that can occur on a noisy geometric network are *wavefront propagation* (WFP) and the *appearance of new clusters* (ANC) [71]. Wavefront propagation is the spreading of a contagion along the structure of the underlying space via geometric edges. The appearance of new clusters occurs when long-range, non-geometric edges connect activated nodes with nodes in a region of the network that has been unaffected by the contagion and thereby lead to a new cluster of activated nodes in this previously unaffected region.

Consider, for instance, a *threshold contagion model* [49] that starts with a cluster of nodes that are active initially and in which a node becomes activated at a time step if and only if it has at least T neighbours that were activated at the previous time step[1]. If $T = 1$, a node becomes activated at a time step exactly when it has any neighbours that were activated at the previous time step and the model behaves as a *simple contagion*. On a purely geometric network (i.e. a geometric network without non-geometric edges), such a simple contagion typically spreads evenly around the seed clusters, following the spatial structure of the underlying space. In other words, it spreads via WFP. This is the basic spreading mechanism of diseases in a world that does not include connections between remote places (e.g. in the form of airline traffic).

The addition of noisy, non-geometric edges to a geometric network can dramatically change the spreading behaviour of such a simple contagion. The non-geometric edges can connect clusters of active nodes with distant inactive nodes and thereby lead to the appearance of new clusters of active nodes in regions of the network that have not yet been reached by wavefront propagation. We can view the non-geometric edges as 'bridges' that accelerate the spreading process considerably, especially if they are 'long' (see Figure 1.1).

Now consider a threshold contagion model that (again) starts with a cluster of nodes being activated initially, but where for a node to become activated at a time step, it requires $T \geq 2$ of its neighbours to be activated at the previous time step. On a purely geometric network, such a contagion also spreads by WFP, although the expansion around the seed clusters may be somewhat less even (on an irregular network). However, the addition of non-geometric edges usually has much less of an

[1]It is also common to put a threshold on the fraction (rather than the number) of infected neighbours in the update rule, and we will work with a model that uses the fractional threshold in the main part of this thesis. We consider an update rule putting a threshold on the number of neighbours in this section for ease of explaining various phenomena in an intuitive way. Analogous phenomena apply to the fractional case and to more complicated update rules.

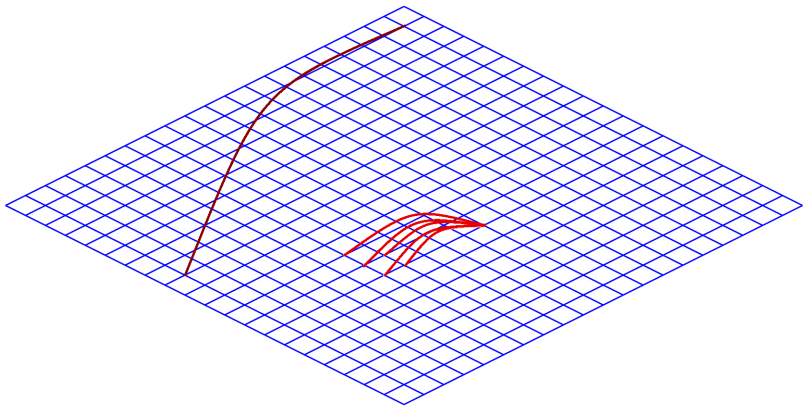


Figure 1.1: Types of non-geometric edges illustrated on a two-dimensional lattice. A 'long bridge' (dark red) is formed by a non-geometric edge that connects nodes that are far away from each other in the underlying space. A 'wide bridge' (bright red) is formed by several non-geometric edges that are 'parallel'. Here, 'parallel' means that they all connect a node with members of a cluster of nodes that are adjacent via geometric edges.

effect on the propagation of a contagion that follows a threshold model with $T \geq 2$ than on that of a contagion that follows a threshold model with $T = 1$. The former cannot spread via a single edge; it requires several edges to act collectively to activate a node. Therefore, bridges of non-geometric edges need to consist of multiple edges to facilitate the propagation of such a complex contagion. In other words, for non-geometric edges to be effective at spreading such a complex contagion via ANC, they must form bridges that are 'wide' as well as 'long', which is a much stricter requirement. The fact that non-geometric edges connect distant nodes and tend to be fewer in number implies that they are less likely to be 'parallel' (see Figure 1.1). This phenomenon is known as the 'weakness of long ties' and was examined in depth by Centola *et al.* [16] and Ghasemiesfeh *et al.* [35]. Of course, non-geometric edges can contribute to the spreading of complex contagions via WFP, but this contribution is normally a minor one.

Taylor *et al.* [71] used methods from nonlinear dimensionality reduction and topological data analysis to study spreading behaviour of a threshold contagion model on

noisy geometric networks. Specifically, they explored the occurrence of the two competing propagation mechanisms of WFP and ANC on a noisy ring lattice, and they examined the extent to which the spreading process follows the ring structure. To investigate the extent to which a complex contagion on a network adheres to the structure of the underlying space, they introduced the notion of a *contagion map*, which maps each node of a network to a point in $\mathbb{R}^n$ based on its activation times in different realizations of a contagion process. It thereby produces a point cloud that can be viewed as a geometrical distortion of the network that reflects the contagion's spreading behaviour. To see if they could identify the structure of the underlying space, Taylor *et al.* examined the geometry, dimensionality, and topology of such point clouds. They compared their results with a bifurcation analysis of the contagion on noisy ring lattices and found that the contagion map successfully recovers the geometry, dimensionality, and topology of the underlying space exactly when the contagion propagates predominantly by WFP. This illustrates that one can use contagion maps to illuminate propagation patterns of spreading processes on noisy geometric networks whose underlying space is known, an approach on which we build our work in Chapter 4. Moreover, they found that on noisy ring lattices WFP occurs for a wide range of the network and contagion parameters, suggesting that contagion maps are a viable tool for inferring the structure of the underlying space of a noisy geometric network from contagion dynamics on it. That is, one may be able to use contagion maps as a technique for manifold learning.

1.2 Spreading dynamics for manifold learning

Data-mining algorithms seek to extract information from data. Typically, data captures a number of attributes (so-called 'features'[2]) of a collection of objects (so-called

[2]Note that the use of the term 'feature' in this context differs from its use for topological attributes of a space that we refer to heavily in large parts of this thesis.

'instances'), each of which can be specified by a real-valued number. Data of this form can be represented as a point cloud in Euclidean space[3], known as a *feature space*. If the number of features is n, then each object is represented by a vector in $\mathbb{R}^n$ whose Cartesian coordinates are the values of its features. Data is often very high-dimensional, meaning that it records a very high number of features (e.g. DNA microarray data that measures expression levels of a large number of genes). Often the number of features (i.e. the dimension of the feature space) is much larger than the number of objects (i.e. the number of points in the point cloud). Large point clouds can usually be handled efficiently. Managing high dimensionality, on the other hand, tends to be difficult and computationally intensive — a phenomenon that is sometimes called 'the curse of dimensionality' [5]. However, it is often the case that some features are redundant (e.g. the same measurement in both feet and meters) or irrelevant to the information to be extracted (e.g. irrelevant genes on a DNA microarray). In other words, the actual intrinsic dimension of a data set may be much lower than the dimension of the feature space. This suggests that it may be possible to reduce the dimension of the data without losing relevant information, thereby improving the ability to extract relevant, non-redundant information. This is precisely what dimensionality-reduction techniques are designed to do: to find relevant low-dimensional structure in high-dimensional data sets. In particular, manifold-learning techniques aim to identify lower-dimensional manifold structure in high-dimensional data sets [47]. High-dimensional point-cloud data may represent a large number of features on a collection of objects as described above. Another way for high-dimensional point-cloud data with low-dimensional intrinsic structure to arise is as a sample of points from a low-dimensional manifold that is embedded in a high-dimensional space. Manifold-learning techniques are intended to identify such an underlying manifold.

[3]Note, however, that there are data types that cannot be represented in Euclidean space.

Consider, for instance, data points that lie on a plane in three-dimensional space. Principal component analysis (PCA), a classical dimensionality-reduction technique [69], can find the directions along which the data has maximum variance, as well as the relative importance of these directions. In the case of the plane embedded in three-dimensional space, PCA returns three vectors: two of positive weight spanning the plane and one vector of zero weight that is orthogonal to the plane. PCA can thus identify the plane underlying the ostensibly three-dimensional data. More generally, consider data points that are concentrated around a low-dimensional manifold (reflecting the underlying information) that is embedded in a high-dimensional space. PCA is a *linear* dimensionality-reduction method: If the manifold is nonlinear, PCA is unable to detect the low-dimensional space underlying the data set. This is where manifold-learning techniques (as a type of *nonlinear* dimensionality reduction) can be effective. The purpose of manifold learning is to uncover low-dimensional manifold structure of a data set in a high-dimensional feature space, even if the structure of the data is curved. Some established manifold-learning algorithms are Isomap [72], locally linear embedding (LLE) [65], Laplacian eigenmaps [4], semidefinite embedding [78], and diffusion maps [20].

A common procedure for nonlinear dimensionality reduction is the following:

(1) Create a network on the data points, such as by defining an edge between any two nodes within some distance ϵ (producing the *ϵ-neighbourhood graph*) or by connecting each point to its k closest neighbours (producing the *k-nearest-neighbour graph*).

(2) Define some notion of pairwise distance between data points based on this network (e.g. shortest-path length for Isomap). The aim is to approximate the actual geodesic distance on the underlying manifold.

(3) Map points to some space based on the pairwise distances. Possible ways of

doing this include the following: (a) use a multidimensional scaling algorithm, which finds an embedding that preserves pairwise distances as well as possible; or (b) take distances to be coordinates in a space of dimension equal to the number of data points (the approach that contagion maps take as a manifold-learning technique [71]).

Questions that arise naturally in relation to manifold learning include the following. How can we measure the performance of a manifold-learning technique? What are some challenges for manifold-learning techniques? What conditions of a data set should a 'good' manifold-learning algorithm be able to handle effectively? How can we compare different manifold-learning algorithms and determine which algorithms are most suitable for various types of data sets? Particular challenges for manifold-learning techniques include curvature, non-uniform sampling, and noise of data. Many well established manifold-learning techniques perform poorly when faced with some of these challenges. Isomap, for instance, can be very sensitive to noise. Consider, for example, a noisy point sample on the Swiss roll (see Figure 1.2). Noise can lead to two points on adjacent sheets lying close together. The k-nearest-neighbour graph might then have an edge that connects the corresponding nodes (i.e. a 'short-circuit error'), although the points lie far apart in the intrinsic geometry. Consequently, Isomap falsely considers the two points to be close, and thus fails as a manifold-learning technique in this case.

As a manifold-learning technique, contagion maps [71] (see the final paragraph of Section 1.1) circumvent this issue by exploiting the 'social reinforcement' phenomenon that characterizes threshold contagions. When the threshold of a contagion is small enough to allow spreading via a single edge, the contagion map can be viewed as a variant of Isomap and is similarly sensitive to the type of noise described above. For larger thresholds, however, a single errant edge in the k-nearest-neighbour graph

cannot carry a contagion, and, as a result, the contagion map does not view the two points as close and performs well as a manifold-learning technique.

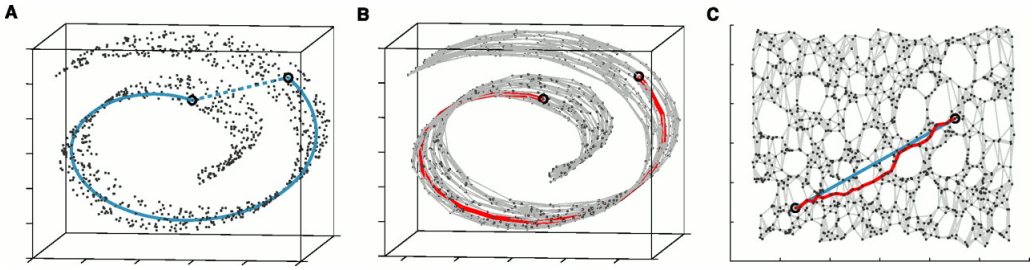


Figure 1.2: (A) Points sampled from a Swiss roll. The distance between two points according to the ambient space can differ significantly from the geodesic distance between the same two points along the underlying two-dimensional manifold. (B) The Isomap algorithm first builds a neighbourhood graph on the point cloud and then finds the length of a shortest path between any pair of nodes in this network. (C) The length of a shortest path between two nodes in the network approximates the geodesic distance between the corresponding two points under favourable conditions [6]. Isomap performs classical multidimensional scaling based on the set of shortest-path distances. (Figure taken with permission from [72].)

Tools from applied algebraic topology, in particular the theory of persistent homology, have been successfully applied in a wide variety of settings ranging from the discovery of a new subtype of breast cancer [56] and the analysis of the neural activities of several animal species [9, 31, 68] to the quantification of pore-geometry in nanoporous materials [48]. *Homology* is an algebraic tool for defining and identifying properties of topological spaces. Given a topological space X, homology associates a family of *homology groups*, which capture the topological features of X at different dimensions. For a natural number n, the nth homology group $H_n(X)$ provides information about the n-dimensional 'holes' of the space X. The rank of $H_0(X)$ counts the number of connected components of X, the rank of $H_1(X)$ counts the number of 1-dimensional 'holes', or 'loops', of X, the rank of $H_2(X)$ counts the number of 2-dimensional 'holes', or 'voids', of X, etc. *Persistent homology* is an extension of homology that enables one to detect topological features of a space across a range of scales or resolutions [27, 36]. The input for persistent homology is a *filtration* (i.e. a

nested sequence) of topological spaces. Persistent homology tracks the evolution of the topological features of this sequence. There are various techniques to obtain a filtration based on a data set and the choice of filtration depends both on the type of data and on the information that one wishes to infer through persistent homology. One of the most common applications of persistent homology is the task of recovering a manifold from a random, potentially noisy sample of points from it [14]. This application illuminates the natural overlap of topological data analysis with manifold learning. Both are designed to find shape regardless of exact geometry (including measures of curvature and length), and both aim to be robust to noise. Traditionally, they differ in their respective approaches, and, as a result, in their ability to identify different types of structural features. Persistent homology can, for example, identify a sphere (by its topological features in dimensions 1 and 2), but not a Swiss roll (as it has no non-trivial topological features). A manifold-learning algorithm like Isomap, on the other hand, can 'unroll' the Swiss roll (under favourable conditions), but cannot see that a sphere is a two-dimensional manifold: Isomap detects a sphere's lowest embedding dimension 3, but cannot see its intrinsic 2-dimensional structure. We combine the two approaches in the work contained in Chapter 5 by processing our data via a manifold-learning-type procedure first, and then computing persistent homology (along with some other measures) based on this processed data.

1.3 Overview and structure of this thesis

In the first part of this thesis, we study contagion maps by building on the approach of [71] and investigating a new model. We still use a threshold contagion model, but we consider a more complicated family of noisy geometric networks. Our networks, which can be interpreted as geometrically embedded in a flat torus, are similar to the Kleinberg small-world model [46]. We use a contagion map to construct a point

cloud that represents the dynamics of the contagion from a set of different initial conditions. We then examine the structure of this point cloud in three different ways: topologically (via the homology of a space built on the point cloud), geometrically (via distances between pairs of points), and with regard to dimensionality (via the approximate embedding dimension). We compare our findings to the topological and geometric structure, as well as the embedding dimension, of the torus. If the point cloud's structure resembles that of the torus geometrically, topologically and in terms of its embedding dimension, it is suggestive that the contagion propagates via wavefront propagation along the structure of the underlying torus.

The motivation for our choice of network model is fourfold: (1) it can be viewed as a two-dimensional analogue of the noisy ring lattice that was studied in [71]; (2) the Kleinberg small-world model includes nontrivial and adjustable spatial scaling from its probability distribution for constructing non-geometric edges; (3) the flat torus has locally Euclidean geometry and is entirely homogeneous (the local geometry at one point is the same as that at any other); and (4) the embedding of the network in the torus entails nontrivial topological features to take into account when comparing the topological structure of the contagion map to that of the underlying space. In our investigation, we find, for a certain region of the parameter space, that the structure of the contagion map resembles that of a torus in terms of topology, geometry, and dimensionality. Furthermore, this region corresponds to scenarios in which we can predict analytically that the contagion spreads predominantly via WFP rather than ANC. This consolidates the approach of [71] as a way of determining spreading behaviour of contagions on noisy geometric networks. Moreover, the presence of this region indicates the use of contagion maps as a manifold learning technique that is robust to noise.

We explore this application of contagion maps in the second part of this thesis. In

particular, we compare the performance of contagion maps to that of the well-known dimensionality reduction technique Isomap when applied to a number of artificial and real-world data sets. We find that, under certain circumstances, contagion maps successfully identify underlying structure of data while Isomap fails. This demonstrates that contagion maps are useful for dimensionality reduction and should be a valuable addition to a toolkit of dimensionality-reduction techniques.

This thesis is organized in four additional chapters. In Chapter 2, we present some of the mathematical background for the methodology used in later chapters. In particular, we introduce the concept of persistent homology and the Wasserstein distance between two barcodes, one of the central tools in our analysis of the spatial structure of data. In Chapter 3, we provide background on noisy geometric networks and contagions on networks, and define the particular network model as well as the threshold contagion model that we will study.

In Chapter 4, we describe our methodology for studying the spreading behaviour of our contagion model on our networks, and present our results. In Section 4.1, we define the contagion map, which maps the nodes of a network to points in high-dimensional space by recording the activation times of the nodes in several realizations of a contagion model on the network. We describe how we analyse this point cloud by examining its geometry, topology and dimensionality. We address the issue of the scale of a barcode affecting the Wasserstein distance, and we propose and incorporate a way of compensating for this effect. In Section 4.2, we perform numerical experiments to explore how different values of the various network and contagion parameters affect the shape of the contagion map. We also conduct a bifurcation analysis to compare to our numerical findings. Finally, in Section 4.3, we summarize and discuss our findings; and we suggest some possible directions for future work.

In Chapter 5, we compare contagion maps to Isomap as a manifold learning

technique. We describe both algorithms in Section 5.1, and introduce the data sets that we test them on in Section 5.2. Section 5.3 contains our results for using various versions of Isomap and contagion maps on these data sets. We summarize our findings in Sections 5.4, and list open questions raised by the results in this chapter as well as avenues for future work.

We explore the effects of different choices of metric on the persistent homology of a regularly spaced point sample on a torus in Appendix A.

Chapter 2

Background Material from Algebraic Topology

In later chapters, we will use methods from algebraic topology to study contagion processes on networks whose underlying spatial structure is known, and (conversely) to infer underlying spatial structure of network and point-cloud data when this structure in unknown. In particular, we will use *persistent homology*, one of the most important concepts in applied algebraic topology. The idea of persistent homology was introduced independently by Frosini, Ferri, Robins, and Edelsbrunner [13, 29, 33, 63]. It is an algebraic tool for detecting topological features of a space at different scales and determining their 'persistence'. Features that persist for a short amount of time are often classified as noise and not considered characteristic features of the space. In this chapter, we first give an account of the necessary material form classical algebraic topology that persistent homology builds on. We then present persistent homology, as well as the notion of a *barcode*, an algebraic object that represents the persistent homology of a filtration, and the *Wasserstein distance* between two barcodes. Finally we describe two ways of constructing a simplicial complex on a point cloud: the Čech complex and the Vietoris–Rips complex. The Čech complex provides a topologically faithful representation of the space underlying a point cloud. The Vietoris–Rips complex is a reasonable and easily computable approximation to the Čech complex, which

is why we will use it in later chapters to examine the shape of point-cloud data.

2.1 Simplicial homology

The material in this section consists of classical result, which can be found e.g. in Chapter 2 of [39].

Definition 1. An *abstract simplicial complex* is a finite collection Σ of finite ordered sets that is closed under inclusion: whenever $\alpha \in \Sigma$ and $\beta \subseteq \alpha$, it follows that $\beta \in \Sigma$.

The elements of Σ are called *simplices*. Two (ordered) simplices are considered to be equal if they are an even permutation of each other, and to have opposite orientation if they are an odd permutation of each other. A *face* of a simplex α is a non-empty proper subset $\beta \subset \alpha$. The dimension of a simplex $\alpha \in \Sigma$ is $|\alpha| - 1$. The 0-dimensional simplices are called *vertices*. The dimension of a simplicial complex is the maximum of the dimensions of the simplices that it contains. A *simplicial subcomplex* $\Omega \subseteq \Sigma$ of a simplicial complex Σ is a subcollection of simplices that is itself a simplicial complex. For $n \in \mathbb{N}$, the *n-skeleton* of a simplicial complex is the union of its simplices of dimensions $m \leq n$.

To each simplex, we can assign a polytope, which is called its *geometric realization*: A 0-simplex corresponds to a vertex, a 1-simplex corresponds to an edge, a 2-simplex corresponds to a triangle, a 3-simplex corresponds to a tetrahedron, and so on. A simplicial complex Σ can thereby be represented as a subset of the simplex spanned by its vertices. See Figure 2.1 for examples of geometric realizations.

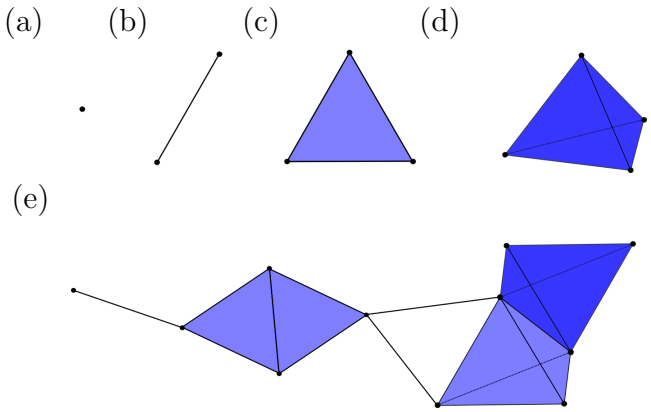


Figure 2.1: Geometric realizations of (a) a 0-simplex, (b) a 1-simplex, (c) a 2-simplex, (d) a 3-simplex, and (e) a 3-dimensional simplicial complex.

Let Σ be a simplicial complex, let k be an integer, and let R be some Abelian group. A *k-chain* is a formal linear combination of k-simplices in Σ over R, such that $a \cdot \alpha = -a \cdot p_o(\alpha)$, such that $a \in R$, $\alpha \in \Sigma$, and p_o is an odd permutation. We can turn the set C_k of k-chains into an Abelian group by defining the binary operation to be componentswise addition. In topological data analysis, the most used group is $\mathbb{Z}/2\mathbb{Z}$; in this case, a k-chain can viewed as a collection of k-simplices in Σ, and addition is equivalent to taking the symmetric difference. It is straightforward to check that C_k satisfies the group axioms with this binary operation and that the identity element is the empty set.

The *boundary* $\partial_k(\alpha)$ of a k-simplex α is the alternating sum of its $(k-1)$-dimensional faces. The *boundary* of a k-chain is the sum of the boundaries of its simplices (see Figure 2.3 for an example). The boundary of a k-chain is a $(k-1)$-chain and hence the boundary operator ∂_k defines a function $\partial_k : C_k \longrightarrow C_{k-1}$. This function commutes with the binary operation defined on C_k. That is, ∂_k is a homomorphism, called the *boundary homomorphism*. We thus have a sequence of Abelian groups connected by boundary homomorphisms:

$$\cdots \xrightarrow{\partial_{k+2}} C_{k+1} \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \xrightarrow{\partial_{k-1}} \cdots \ . \tag{2.1}$$

A k-chain that is in the image B_k of ∂_{k+1} is called a *k-boundary*. A k-chain that is in the kernel Z_k of ∂_k is called a *k-cycle* (see Figure 2.2). A fundamental property of the boundary operator is that the boundary of a boundary is empty (see Lemma 1 below). Any sequence (2.1) of Abelian groups with this property is called a *chain complex*.

Lemma 1. (Fundamental Lemma of Homology) *For any integer k and k-chain $d \in C_k$ we have that $\partial_{k-1}\partial_k(d) = 0$.*

Proof. It suffices to show that $\partial_{k-1}\partial_k(\alpha) = 0$ for $\alpha = [v_0, v_1, \ldots, v_k]$ a k-simplex of Σ. Indeed,

$$\partial_k(\alpha) = \sum_i (-1)^i [v_0, v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_k] \ ,$$

and hence

$$\begin{aligned}\partial_{k-1}\partial_k(\alpha) = &\sum_{j<i} (-1)^i (-1)^j [v_0, v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_{i-1}, v_{i+1}, \ldots, v_k] \\ &+ \sum_{j>i} (-1)^i (-1)^{j-1} [v_0, v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_{j-1}, v_{j+1}, \ldots, v_k] = 0 \ .\end{aligned}$$

□

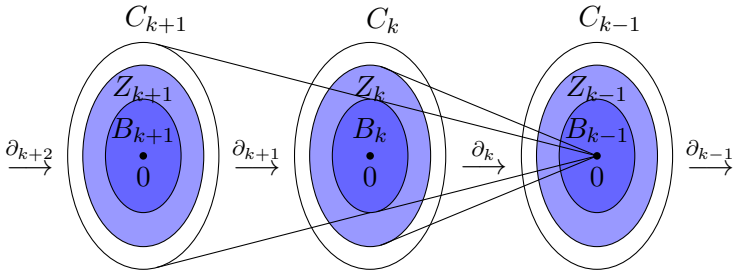


Figure 2.2: The chain complex consisting of a sequence of chain, cycle, and boundary groups connected by boundary homomorphisms.

That is, the kth boundary group B_k is a subgroup of the kth cycle group Z_k, so the following definition makes sense.

Definition 2. Given a simplicial complex Σ and an integer k, the *kth homology group* $H_k(\Sigma)$ is the quotient group of the kth cycle group $Z_k(\Sigma)$ by the kth boundary group $B_k(\Sigma)$:

$$H_k(\Sigma) = Z_k(\Sigma)/B_k(\Sigma) \ .$$

The kth *Betti number* $\beta_k(\Sigma)$ is the rank of the kth homology group of Σ:

$$\beta_k(\Sigma) = \operatorname{rank} H_k(\Sigma) \ .$$

Two k-cycles represent the same element of the kth homology group H_k if they differ only by k-boundaries. Roughly speaking, $\beta_n(\Sigma)$ is the number of n-dimensional 'holes' of the space Σ. For example, $\beta_0(\Sigma)$ is the number of connected components, $\beta_1(\Sigma)$ is the number of 'loops', and $\beta_2(\Sigma)$ is the number of 'voids' of the geometric realization of Σ. See Figure 2.3 for an example that illustrates these ideas.

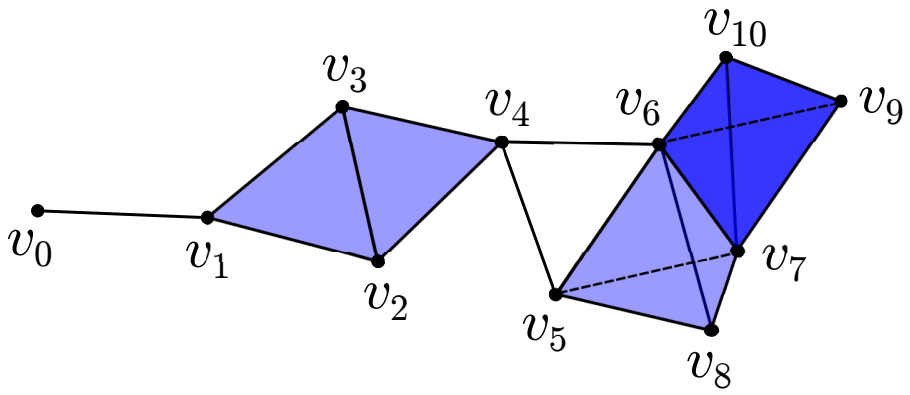


Figure 2.3: A 3-dimensional simplicial complex with Betti numbers $\beta_0 = 1$, $\beta_1 = 1$, and $\beta_2 = 1$. The boundary of the 3-simplex $\{v_6, v_7, v_9, v_{10}\}$ is $\partial_3(\{v_6, v_7, v_9, v_{10}\}) = \{v_7, v_9, v_{10}\} - \{v_6, v_9, v_{10}\} + \{v_6, v_7, v_{10}\} - \{v_6, v_7, v_9\}$. The boundary of the 2-chain $\{v_1, v_2, v_3\} + \{v_2, v_4, v_3\}$ is $\partial_2(\{v_1, v_2, v_3\} + \{v_2, v_4, v_3\}) = \{v_2, v_3\} - \{v_1, v_3\} + \{v_1, v_2\} + \{v_4, v_3\} - \{v_2, v_3\} + \{v_2, v_4\} = -\{v_1, v_3\} + \{v_1, v_2\} + \{v_4, v_3\} + \{v_2, v_4\}$. The 1-chain $\{v_4, v_5\} + \{v_5, v_6\} + \{v_6, v_4\}$ is a 1-cycle, and it is not a 1-boundary. It represents a non-trivial element of H_1 and accounts for the hole that it 'encloses'. The 1-chain $\{v_2, v_4\} + \{v_4, v_3\} + \{v_3, v_2\}$ is a 1-cycle, but it does not account for a 1-dimensional hole because it is also the boundary of the 2-chain $\{v_2, v_4, v_3\}$; the hole is 'filled in'. The 1-chain $\{v_5, v_8\} + \{v_8, v_6\} + \{v_6, v_4\} + \{v_4, v_5\}$ is a 1-cycle. It represents the same element in H_1 as the 1-cycle $\{v_4, v_5\} + \{v_5, v_6\} + \{v_6, v_4\}$ because these two cycles differ only by the boundary of the 2-chain $\{v_5, v_8, v_6\}$.

2.2 Persistent homology

Definition 3. A finite *filtration* of a finite simplicial complex Σ is a nested sequence of simplicial subcomplexes of Σ, of which the 0th member is the empty complex and the last member is all of Σ. That is,

$$\emptyset = F_0 \subseteq F_1 \subseteq \cdots \subseteq F_n = \Sigma.$$

A filtration of a simplicial complex Σ can be viewed as the construction of Σ from the empty set by sequentially adding collections of simplices (see Figure 2.4).

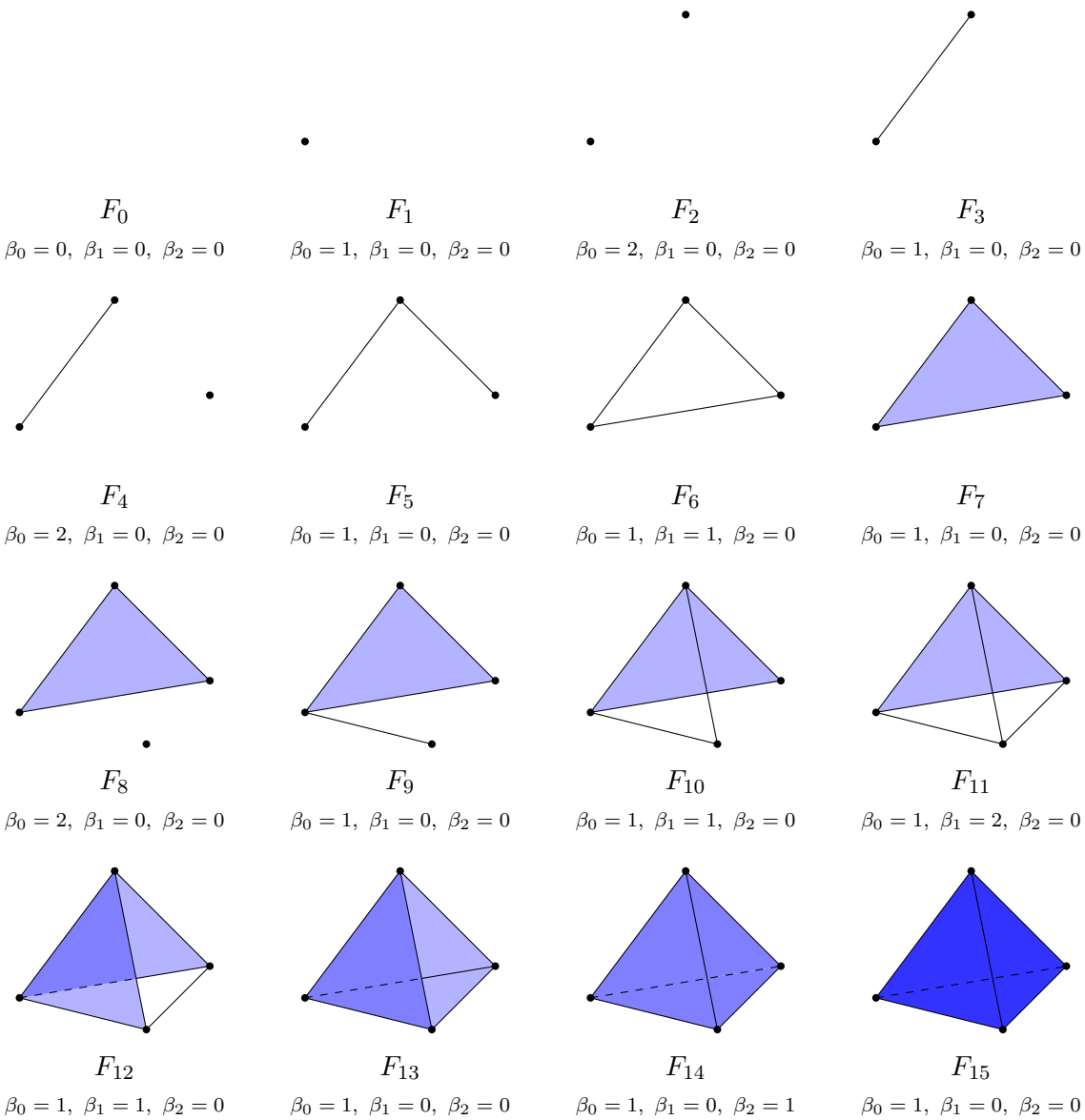


Figure 2.4: Example of a filtration $F_0 \subseteq F_1 \subseteq \cdots \subseteq F_{15}$ of the tetrahedron. In this example, exactly one simplex is added at each filtration step. The increase and decrease of β_0, β_1, and β_2 indicate the appearance and disappearence of connected components, 'loops', and 'voids', respectively. For example, the 'loop' appearing at filtration step F_{10} corresponds to the 1-dimensional homology class that consists of the chain of the three 1-simplices 'around' it. Its birth is indicated by the increase of β_1. The 'loop' persists until filtration step F_{12}, where it disappears as the corresponding homology class now merges with the trivial homology class, because its representative chain becomes the boundary of the 2-simplex that is added. Its death is indicated by a decrease of β_1.

Given a filtration $F_0 \subseteq F_1 \subseteq \cdots \subseteq F_n$ of a simplicial complex Σ, for every $i \leq j$

and dimension k, the inclusion map from F_i to F_j induces a homomorphism

$$f_k^{i,j} : H_k(F_i) \longrightarrow H_k(F_j) \ .$$

Therefore, there is a sequence of homology groups that are related via these homomorphisms:

$$0 = H_k(F_0) \xrightarrow{f_k^{0,1}} H_k(F_1) \xrightarrow{f_k^{1,2}} \cdots \xrightarrow{f_k^{n-1,n}} H_k(F_n) = H_k(\Sigma) \ .$$

One can track the evolution of Σ simplicial complex along the filtration by observing the algebraic structures of the homology groups in this sequence.

We can generalize the notion of a homology group in the setting of a filtration of a simplicial complex.

Definition 4. For an integer k, the *kth persistent homology groups* are the images of the homomorphisms induced by inclusion:

$$H_k^{i,j} = \operatorname{im} f_k^{i,j} \ , \quad 0 \leq i \leq j \leq n \, , \quad i, j \in \mathbb{N} \, .$$

The *kth persistent Betti numbers* are the ranks of these groups: $\beta_k^{i,j} = \operatorname{rank}(H_k^{i,j})$.

Using Definition 4, we can formalize the notion of birth and death of a homology class.

Definition 5. A non-zero homology class $\xi \in H_k(F_i)$ is *born at* F_i if $\xi \notin H_k^{i-1,i}$, and it *dies at* F_j if $f_k^{i,j-1}(\xi) \neq 0$ but $f_k^{i,j}(\xi) = 0$.

For $0 \leq i \leq j \leq n$, the kth persistent homology group $H_k^{i,j}$ can be interpreted as the group that consists of all homology classes that are born at or before F_i and are still alive at F_j.

We can now finally give a mathematically rigorous definition of persistence.

Definition 6. Let $\emptyset = F_0 \subseteq F_1 \subseteq \cdots \subseteq F_n = \Sigma$ be a filtration of a simplicial complex Σ, and let k be an integer. If $\xi \in H_k(F_i)$ is a homology class that is born at F_i and dies entering F_j, then its *persistence* (or 'lifespan') is the difference $\text{pers}(\xi) = \text{j} - \text{i}$. If ξ never dies, we note its death time as infinite (i.e. $j = \infty$), making its persistence infinite as well (i.e. $\text{pers}(\xi) = \infty$). The interval $[i, j)$ is referred to as a *persistence interval*.

2.3 Barcodes, persistence diagrams, and the Wasserstein distance

If we work with field coefficients (i.e. R is a field), then $H_k(F_i)$ is a vector space and $f_k^{i,j}$ is a linear map for all $k \in \mathbb{Z}$ and $0 \leq i \leq j \leq n$. It is a fundamental theorem [81] that, in this case, one can find elements $\{x_\ell\}_{\ell \in \Lambda_k}$ (with $x_\ell \in H_k(F_{b_\ell})$, where $b_\ell \in \mathbb{N}$ is the birth time of x_ℓ and $d_\ell \in \mathbb{N} \cup \{\infty\}$ is the death time of x_ℓ) such that, for each $j \in \{0, \ldots, n\}$, we have that

$$\left\{ f_k^{b_\ell, j}(x_\ell) \mid \ell \in \Lambda_k \ , \ b_\ell \leq j < d_\ell \right\}$$

forms a basis for $H_k(F_j)$.

If the number of births (respectively, deaths) of homology classes in $\{x_\ell\}_{\ell \in \Lambda_k}$ exceeds the number of deaths (respectively, births) at F_j, then β_k increases (respectively, decreases). Tracking the change of Betti numbers during a filtration is thus useful for monitoring the topological evolution of the growing complex. A topological feature that emerges with the birth of a homology class ξ and disappears with the death of ξ is said to 'correspond to' ξ and has persistence $\text{pers}(\xi)$. The features that persist for a long time interval are usually considered to be the defining features of the complex, although this is not always the case.

The collection of persistence intervals $[b_\ell, d_\ell)$ corresponding to elements in $\{x_\ell\}_{\ell \in \Lambda_k}$ can be viewed as the filtered analogue of the Betti number β_k. Such persistence

intervals can be represented as barcodes, or as diagrams (see Figure 2.5).

Definition 7. The *persistence diagram* of the persistent homology (at dimension $k \in \mathbb{Z}$) of a filtered simplicial complex is the multiset

$$\left\{(b_\ell, d_\ell) \in \overline{\mathbb{R}}^2 \mid \ell \in \Lambda_k\right\} \cup \left\{(i,j) \in \mathbb{R}^2 \mid i = j\right\} \subseteq \overline{\mathbb{R}}^2,$$

where $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$.

Note that all persistence diagrams have equal cardinality. For dimension $k \in \mathbb{Z}$, the collection of persistence intervals $[b_\ell, d_\ell)$ corresponding to elements in $\{x_\ell\}_{\ell \in \Lambda_k}$ is called a *barcode* (see Figure 2.5)

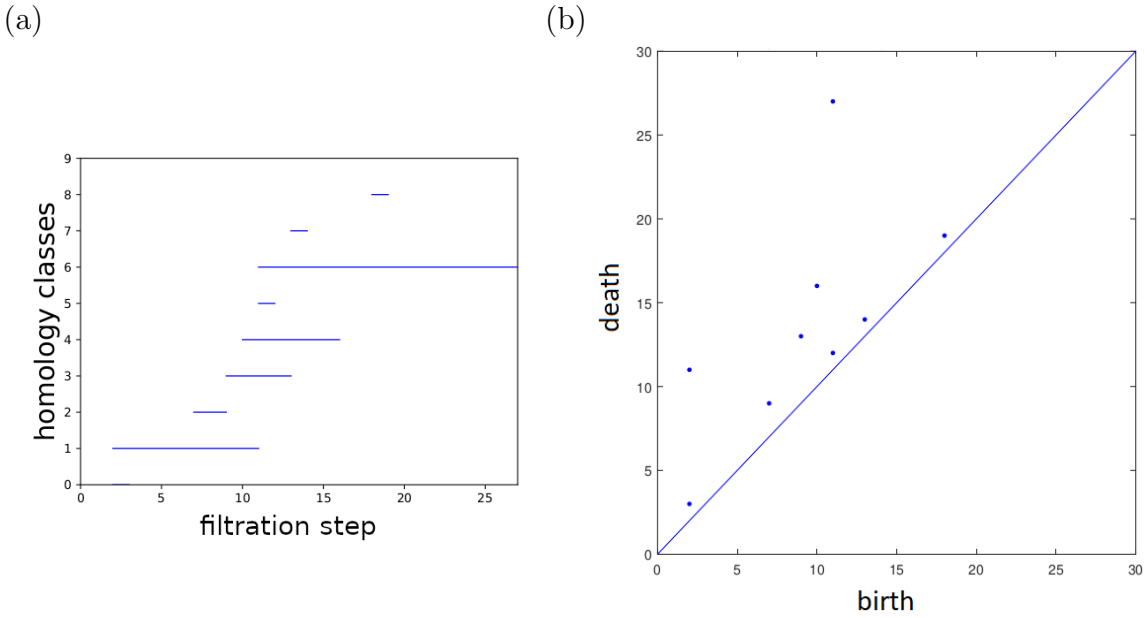


Figure 2.5: (a) An example of a barcode. The horizontal axis represents the time during the filtration. Each horizontal bar indicates the lifespan of a homology class, starting at its time of birth and ending at its time of death. (b) Equivalent persistence diagram. Persistence intervals are marked as points in the extended plane $\overline{\mathbb{R}}^2$, where $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$.

One can turn the space of persistence diagrams (equivalently, the space of barcodes) into an extended metric space by defining the following notion of distance between two persistence diagrams.

Definition 8. Given two persistence diagrams D_1 and D_2, an extended metric d on $\overline{\mathbb{R}}^2$, and a number $p \in [1, \infty]$, the *pth Wasserstein distance* between D_1 and D_2 is defined as

$$W_p[d](D_1, D_2) := \inf_{\phi: D_1 \to D_2} \left[\sum_{x \in D_1} d\left[x, \phi(x)\right]^p \right]^{1/p},$$

where ϕ ranges over all bijections from D_1 to D_2.

If $p = \infty$ and $d = L_\infty$, where $L_\infty\left[(x_1, y_1), (x_2, y_2)\right] = \sup\{|x_1 - x_2|, |y_1 - y_2|\}$, the Wasserstein distance $W_\infty[L_\infty]$ is called the *bottleneck distance*.

One property of persistent homology that is central to its utility in applications is that it is stable and therefore robust to noise: A small perturbation of input data induces only a small perturbation of the corresponding persistence diagram with respect to the bottleneck distance. See [19] for details on and a proof of the first stability theorem, and [12, 18], and [11] for generalizations thereof.

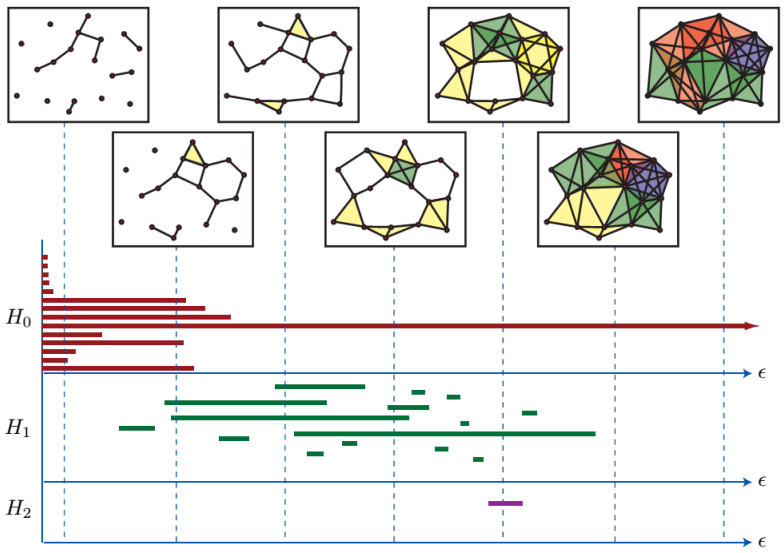


Figure 2.6: A filtration of a simplicial complex and the barcodes illustrating its persistent homology in dimensions 0 (red), 1 (green), and 2 (purple). The horizontal axis represents the time during the filtration. Each horizontal bar represents the lifespan of a homology class. The 0th (1st, 2nd) Betti number of a member of the filtration is the number of red (green, purple) bars intersecting the dashed vertical line that connect this member with the horizontal axis. (Figure taken with permission from [36].)

2.4 Persistent homology on point-cloud data

Persistent homology is a useful tool for analysing point-cloud data. A *point cloud* is a finite set of points,

$$P = \{x_1, x_2, \ldots, x_l\} \subseteq M \ ,$$

in a metric space (M, d), which can be viewed as a sample from some underlying subspace of M. There are various ways to construct a simplicial complex from a point cloud. Two of the most commonly used constructions are the Čech complex and the Vietoris–Rips complex, which we will now describe.

Given a non-negative number ϵ, recall that the closed ball of radius ϵ around a point $x \in M$ is the set of points within distance ϵ from x:

$$B_\epsilon(x) = \{y \in M : d(x, y) \leq \epsilon\} \ .$$

Definition 9. Let (M, d) be a metric space, and let $P = \{x_1, x_2, \ldots, x_l\} \subseteq M$ be a point cloud in M. For $\epsilon \geq 0$, the *Čech complex* $\mathcal{C}_\epsilon$ at ϵ associated with P is the simplicial complex whose simplices are sets of points in P whose closed $\epsilon/2$-balls have non-empty intersection. That is, the set $\{y_1, \ldots, y_k\} \subseteq P$ is a $(k-1)$-simplex of $\mathcal{C}_\epsilon$ if and only if $\bigcap_{i=1}^{k} B_{\epsilon/2}(y_i) \neq \emptyset$.

It is clear that this set of simplices is closed under taking subsets. That is, the conditions for a simplicial complex are indeed satisfied by this definition. The 0-simplices of $\mathcal{C}_\epsilon$ correspond precisely to the points $x_1, x_2, \ldots, x_l$; the 1-simplices are the pairs of points that are within distance ϵ of each other; the 2-simplices are the triples of points whose $\epsilon/2$-balls have non-empty intersection; and so on. The Čech complex $\mathcal{C}_0$ is the set of points in P. For sufficiently large ϵ (to be precise, for ϵ at least as large as the diameter of P), the Čech complex $\mathcal{C}_\epsilon$ is the $(l-1)$-dimensional simplex $\{x_1, x_2, \ldots, x_l\}$ together with all of its faces. If $\epsilon_1 \leq \epsilon_2$, then $\mathcal{C}_{\epsilon_1} \subseteq \mathcal{C}_{\epsilon_2}$, and

increasing ϵ incrementally from 0 to a large enough value gives a filtration of Čech complexes associated to the point cloud P:

$$\emptyset \subseteq \mathcal{C}_0 \subseteq \cdots \subseteq \mathcal{C}_{\epsilon_{\text{large}}}.$$

For $M = \mathbb{R}^n$, the following result, known as the *Nerve Theorem*, states that the Čech complex $\mathcal{C}_\epsilon$ associated to a point cloud P is topologically faithful to the union of the closed $\epsilon/2$-balls around the points in P in the sense that it has the same homotopy type. Intuitively, two spaces have the same homotopy type if they can be transformed into each other by bending, compressing, and expanding them (without having to do any cutting or gluing). Two homotopy equivalent spaces share many topological invariants, including the same homology (see e.g. [39] pages 110–113).

Theorem 2. (Nerve Theorem). *For a point cloud $P = \{x_1, x_2, \ldots, x_l\} \subseteq \mathbb{R}^n$ and $\epsilon \geq 0$, the Čech complex $\mathcal{C}_\epsilon$ is homotopy equivalent to the union of the closed $\epsilon/2$-balls around the points in P.*

See Section III.2 of [28] for a more general discussion and proof of this result.

The Nerve Theorem justifies why, when the point cloud P is a sample of some underlying space, the Čech filtration can reveal topological features of this space. Features that have a long lifespan in the Čech filtration are therefore likely to correspond to features of the underlying space.

Whether the balls of a certain radius around a set of points $P \subseteq M$ have a point of common intersection depends on the entire metric space M and the position of P in it. Checking for a point of common intersection is computationally intensive, so it can be impractical (if not infeasible) to construct the Čech filtration associated to a point cloud [30].

The following construction of a simplicial complex from a point cloud depends

only on the pairwise distances of the points. It is therefore computationally more efficient and hence more useful in practice.

Definition 10. Let (M, d) be a metric space, and let $P = \{x_1, x_2, \ldots, x_l\} \subseteq M$ be a point cloud in M. For $\epsilon \geq 0$, the *Vietoris–Rips complex* $\mathcal{R}_\epsilon$ at ϵ associated to P is the simplicial complex whose simplices are sets of points in P that are pairwise within distance ϵ. The set $\{y_1, \ldots, y_k\} \subseteq P$ is a $(k-1)$-simplex of $\mathcal{R}_\epsilon$ if and only if $d(y_i, y_j) \leq \epsilon$ for all $i, j \in \{1, \ldots, k\}$.

Just as for $\mathcal{C}_\epsilon$, the 0-simplices of $\mathcal{R}_\epsilon$ correspond precisely to the points $x_1, x_2, \ldots, x_l$, and the 1-simplices of $\mathcal{R}_\epsilon$ are the pairs of points that are within distance ϵ of each other. Therefore, the 1-skeleton of $\mathcal{R}_\epsilon$ is the same as that of $\mathcal{C}_\epsilon$. In the definition of higher-dimensional simplices, only the pairwise distances of points play a role. The simplicial complex $\mathcal{R}_\epsilon$ is the maximal simplicial complex that can be built on its 1-skeleton; its k-simplices are the $(k+1)$-cliques of its 1-skeleton. Consequently, the 1-skeleton of $\mathcal{R}_\epsilon$ completely determines the entire complex, which is an attractive quality from a computational point of view because it implies that a Vietoris–Rips complex can be stored as a graph (its 1-skeleton).

If $\epsilon_1 \leq \epsilon_2$, then $\mathcal{R}_{\epsilon_1} \subseteq \mathcal{R}_{\epsilon_2}$, and increasing ϵ incrementally from 0 to a value larger than the maximal distance beween any pair of points in the point cloud gives a filtration of Vietoris–Rips complexes whose 0th member is the collection of 0-simplices and whose final member is the $(l-1)$-dimensional simplex $\{x_1, x_2, \ldots, x_l\}$ together with its faces. That is,

$$\emptyset \subseteq \mathcal{R}_0 \subseteq \cdots \subseteq \mathcal{R}_{\epsilon_{\text{large}}}.$$

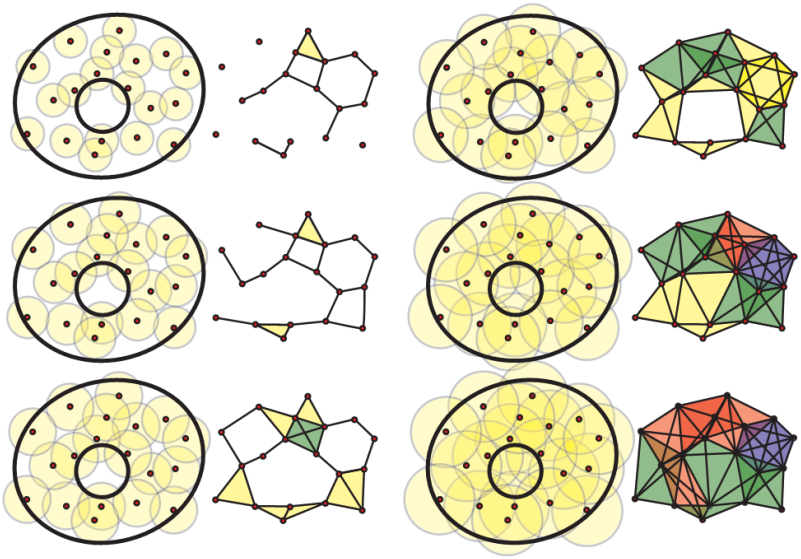


Figure 2.7: A Vietoris–Rips filtration for a point cloud sampled from an annulus. (Figure taken with permission from [36].)

Going back to the special case $M = \mathbb{R}^n$, although a Vietoris–Rips complex $\mathcal{R}_\epsilon$ associated to a point cloud $P \subseteq \mathbb{R}^n$ is not a faithful representation of the union of balls around the points in P and may not even be topologically equivalent to a subspace of $\mathbb{R}^n$, Vietoris–Rips complexes provide a good approximation in the light of persistence, as the following lemma shows.

Lemma 3. *(de Silva, Ghrist [24]). For $M = \mathbb{R}^n$ and any $\epsilon \geq 0$, we have that*

$$\mathcal{R}_\epsilon \subseteq \mathcal{C}_{\epsilon\sqrt{2}} \subseteq \mathcal{R}_{\epsilon\sqrt{2}}.$$

Consequently, any topological feature that persists between $\mathcal{R}_\epsilon$ and $\mathcal{R}_{\epsilon\sqrt{2}}$ in the Vietoris–Rips filtration is also a feature of the Čech complex $\mathcal{C}_{\epsilon\sqrt{2}}$.

The right inclusion in Lemma 3 is obvious: If the balls of radius $\epsilon/2$ around a set of points have a common point of intersection, then the pairwise distances between the points can be at most ϵ (i.e. $\mathcal{C}_\epsilon \subseteq \mathcal{R}_\epsilon$ for any $\epsilon \geq 0$). For a proof of the left inclusion and a tighter bound, see [24].

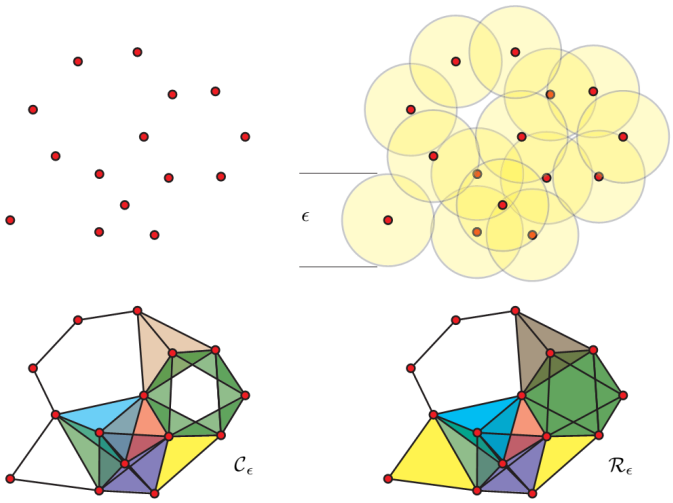


Figure 2.8: Given a point cloud (top left), we can build a Čech complex (bottom left) or a Vietoris–Rips complex (bottom right) based on the balls of diameter ϵ around the points (top right). We can see that the Vietoris–Rips complex contains some simplices that are not in the Čech complex and therefore fails to capture some topological features. (Figure taken with permission from [36].)

Chapter 3

Contagions on Noisy Geometric Networks

3.1 Background

3.1.1 Noisy geometric networks

A *geometric network* is a network whose nodes are embedded in some metric space and whose edges, called *geometric edges*, can occur only between pairs of nodes that are sufficiently close in this space [2, 3]. Geometric networks are useful models for many real-world scenarios. For example, a network describing the interactions between entities that are geographically located on the surface of the earth is a geometric network if interactions are possible only between geographically close entities. Its underlying metric space is the two-dimensional sphere. In addition to geometric networks whose nodes are embedded naturally, nodes of a network can also be embedded in a latent space, e.g. based on the probablity of a connection occurring between two nodes — and this construction has proven useful in a variety of applications [7, 42, 66].

A *noisy geometric network* is a network that is built from a geometric network by adding so-called *non-geometric* (i.e. 'noisy') edges between pairs of nodes that can be remote from each other in the underlying metric space. For example, in synthetic noisy geometric networks, the nodes may be located on a manifold that is embedded in an ambient Euclidean space. Geometric edges can be placed between

all or some of the node pairs that are at distances from each other that are below some fixed threshold. Non-geometric edges can then be added uniformly at random (see Figure 3.1(a)) or following some other probabilistic or deterministic rule. For example, one can fix the number of non-geometric edges that leave a node and connect each of these edges to another node with a probability that is inversely proportional to the distance between the two nodes (compare to our model in Section 3.2.1).

As another example (see Figure 3.1(b)), one can add noise to the nodes' locations in the ambient space and place a non-geometric edge between any two nodes that are close in the ambient space, but not close with respect to the geodesic distance along the manifold. This construction is the origin of the term 'noisy edge'. Many nonlinear dimensionality-reduction techniques, such as diffusion maps and Isomap [4, 20, 34, 69, 72], start by inferring a proximity network from point-cloud data, such as by connecting each point to its k nearest neighbours, with the goal of finding underlying low-dimensional structure of the point cloud. Such proximity networks can be viewed as noisy geometric networks, as it is possible for nodes to be adjacent even when they are not close on the underlying manifold, and nonlinear dimensionality-reduction techniques seek to find purely geometric structures in such networks.

Finally, note that many popular small-world network models [55, 60] — such as Watts–Strogatz [77], Newman–Watts [54], and Kleinberg [45] networks — can be viewed as noisy geometric networks.

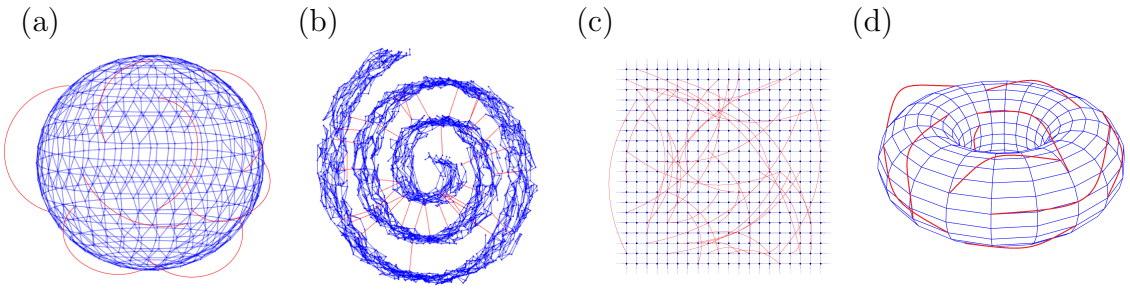


Figure 3.1: Examples of noisy geometric networks. (a) Nodes lie on a two-dimensional sphere embedded in $\mathbb{R}^3$. Geometric edges (blue) occur between nodes that are close on the sphere; non-geometric edges (red) occur uniformly at random. (b) The noisy Swiss-roll. Nodes correspond to a noisy sample from a bounded two-dimensional manifold that is embedded in $\mathbb{R}^3$ as a roll. Geometric edges (blue) occur between nodes that are close on the manifold; non-geometric edges (red) occur between nodes that are close in $\mathbb{R}^3$ but not close on the manifold. (c,d) Kleinberg-like small-world network (see Section 3.2.1) with a geometric degree of $d^{\mathrm{G}} = 4$. (c) A regular lattice of geometric edges (blue) is (d) 'wrapped up' into a torus. Non-geometric edges (red) are placed according to some probability distribution.

3.1.2 Contagions on networks

A *contagion* on a network is a dynamical process in which nodes become successively 'activated', starting from some initial condition. Typically, the initial condition consists of a set of 'seed' nodes, which are active at time $t = 0$ [59, 61].

One of the simplest and best-studied models of a contagion on a network is the discrete-time version of the Watts threshold model (WTM) [76] (see also [37, 74]), which is related to bootstrap percolation [17]. In the WTM, each node i is assigned a *threshold* $T_i \geq 0$. At any given time, each node is in one of two states: It is either *active* (or, depending on the context, 'infected', 'adopted', etc.) or *inactive*. At time $t = 0$, only the seed nodes are active. At each subsequent time step, a node i becomes active if and only if the fraction of neighbours of i that are active is at least T_i. Once a node is activated, it remains active; that is, the WTM is a monotonic process.

A general contagion model can be significantly more complicated than the WTM. For example, biological contagion models are often not monotonic, as an active node can become deactivated following some deactivation rule. Additionally, in many

contagion models, there are more than two possible node states [52]. For example, the nodes in a model describing the spreading of a contagious virus in a population can be susceptible, infective, or recovered [44]. Some models use thresholds in their update rule, but place them on the number of active neighbours of a given node, rather than on the fraction of such nodes in a neighbourhood. Models of the former kind model scenarios where contact with individual agents can induce a change of state of the entity under consideration, regardless of other neighbouring entities. Models of the latter kind, on the contrary, are concerned with scenarios in which the collective of neighbouring entities drives the state of a particular entity. That is, the 'common trend' around a given entity, rather than the influence of collections of individual neighbours, is what affects it (see [37, 74]).

3.2 Our network and contagion models

We consider a family of noisy geometric networks and study a threshold contagion model on it. Our networks can be thought of as being geometrically embedded in a two-dimensional manifold, with their nodes evenly spread on the surface of a flat torus. The torus has Betti numbers $\beta_0 = 1$, $\beta_1 = 2$, and $\beta_2 = 1$ (see Definition 2 in Chapter 2). So there are multiple nontrivial topological features to be taken into account when comparing the structure of the point cloud produced by a contagion map to that of the underlying space.

The geometric edges of our noisy geometric network are distributed evenly between node pairs at a density that depends on a parameter. Non-geometric edges are added at random according to a spatial probability distribution: The larger the distance between two nodes (defined as (3.1)), the less likely it is that they are adjacent via a non-geometric edge. The number of non-geometric edges and the strength of their preference for connecting close nodes are determined by one parameter each.

The regular structure of geometric edges across the torus and the fact that non-geometric edges respect the underlying geometry to some extent leads us to expect that recovery of the topological features of the torus in the point cloud produced by the contagion map will be possible for a wide range of parameters.

The contagion model that we consider on these noisy geometric networks is the WTM with homogeneous thresholds, such that a node becomes active at a given time step if and only if the fraction of its already active neighbours is strictly larger than a fixed threshold T. Threshold contagion models are well-studied and model effects such as social reinforcement on the spreading of a contagion [15, 49]. Taylor *et al.* [71] used the same contagion model, so our investigation serves both as a test of their methodology and as a generalization of their work.

3.2.1 Network model

Our noisy geometric network is a variant of the Kleinberg small-world model [46]. We start with a periodic square lattice of $N = n \times n$ nodes:

$$V = \frac{\mathbb{Z} \times \mathbb{Z}}{n\mathbb{Z} \times n\mathbb{Z}} ,$$

so

$$\overline{(i_x, i_y)} = \overline{(kn + i_x, ln + i_y)} \in V \text{ for all } k, l \in \mathbb{Z} .$$

We define the *periodic lattice distance* μ_{per} between nodes $\overline{(i_x, i_y)}, \overline{(j_x, j_y)} \in V$ to be

$$\mu_{\text{per}}(\overline{(i_x, i_y)}, \overline{(j_x, j_y)}) = |i_x - j_x|_{\text{per}} + |i_y - j_y|_{\text{per}} , \tag{3.1}$$

where

$$|a|_{\text{per}} = \min \{b \in \{0, 1, \ldots, n-2, n-1\} \mid b = kn + a \text{ or } b = kn - a \text{ for some } k \in \mathbb{Z}\} ,$$

and the sum of the two residues in (3.1) is taken in $\mathbb{Z}$. The periodic lattice distance is the regular lattice distance (also known as the 'Manhattan distance') with opposite

sides of the lattice considered to be close. We can thus think of the lattice as being 'wrapped up' into a two-dimensional torus, which has no boundary. In other words, we are using periodic boundary conditions.

We fix $p \in \mathbb{R}_{>0}$ and place a geometric edge between any two nodes whose (periodic) Euclidean distance from each other is within p. A geometric edge is placed between nodes $i = \overline{(i_x, i_y)}$ and $j = \overline{(j_x, j_y)} \in V$ if and only if

$$|i_x - j_x|^2_{\text{per}} + |i_y - j_y|^2_{\text{per}} \leq p^2 \,.$$

We call the number of geometric edges that are incident to a node i its *geometric degree*, and denote it by $d^{\text{G}}(i)$.

Each node also generates $q \in \mathbb{N}$ 'non-geometric stubs', and we connect pairs of stubs to build non-geometric edges. A non-geometric stub emanating from node i is connected to a stub emanating from node j with a probability proportional to $\mu_{\text{per}}(i,j)^{-\gamma}$, where $\gamma \in \mathbb{R}_{\geq 0}$ is a fixed parameter. We call the number of non-geometric edges incident to a node i its *non-geometric degree* and denote it by $d^{NG}(i)$, noting that $d^{NG}(i) = q$. Note that the total degree of a node i is $d_i = d^{G}(i) + d^{NG}(j)$ and that the class of networks defined above consists of regular networks; that is, $d(i) = d(j)$ for all $i, j \in V$. When $\gamma = 0$, the non-geometric stubs are matched uniformly at random; for $\gamma > 0$, non-geometric edges tend to connect nodes that are close with respect to the periodic lattice distance, and this tendency becomes more pronounced for progressively larger γ. The speed of a contagion on this network depends significantly on the parameter γ. If γ is small, non-geometric edges tend to occur in a variety of lengths (including very long ones). That is, the set of non-geometric edges forms many long 'bridges' that enable fast spreading in situations when a single active neighbour is necessary for a node to activate. However, these bridges may not be wide enough to support spreading of a threshold contagion with threshold $T \geq 2/d$. As γ becomes larger, very long non-geometric edges become less

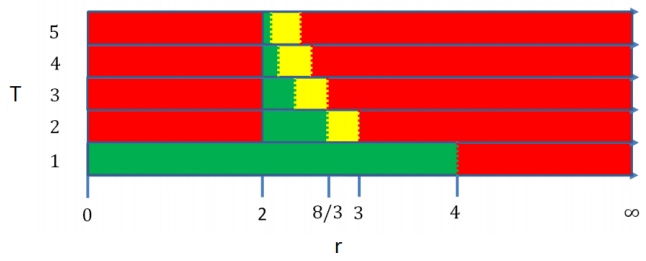


Figure 3.2: The speed of a T-complex contagion on two versions of Kleinberg's small-world model with parameter r (see [26] for details). Green indicates polylogarithmic spreading in both models. Red indicates polynomial spreading in both models. Yellow indicates polylogarithmic spreading in one of the versions and polynomial spreading in the other. (Figure taken with permission from [26].)

likely, and the existing non-geometric edges tend to have lengths in a smaller range, forming several wide bridges that enable fast spreading of complex contagions. If γ is very large, the non-geometric edges tend to be short and form bridges that are wide enough to support spreading of complex contagions, but too short to significantly accelerate the spreading process. These observations can be quantified in a precise form: Ghasemiesfeh et al. [35] showed that the speed of a *T-complex contagion* (a threshold contagion whose update rule is defined by a homogeneous threshold T) on a two-dimensional Kleinberg small-world model with $\gamma = 2$ differs exponentially from the case when $\gamma = 0$. Precisely, the time that it takes for a T-complex contagion to saturate a network, starting from a cluster of T initial seed nodes that are near each other, is a polylogarithmic function of the size (i.e. number of nodes) of the network when $\gamma = 2$ and a polynomial function of the size of the network when $\gamma = 0$. Ebrahimi *et al.* [26] extended these results by providing a characterization of the speed of a T-complex contagion on a Kleinberg small-world network for all values of T and the entire parameter space of γ except at one point (see Figure 3.2).

3.2.2 Contagion model

Our contagion model is the Watts threshold model (WTM) [76]. Let V denote the set of nodes of a network, and let $A = (A_{ij})_{i,j\in V}$ be the adjacency matrix of the network. In our contagion, each node can be either *active* or *inactive*, and we denote the state of node $i \in V$ at time t by $\eta_i(t)$, which takes the value 1 if it is active and the value 0 if it is inactive. We call the set of nodes that are active at time $t = 0$ a contagion *seed*, and we denote the seed set by $S \subseteq V$. That is, $\eta_s(0) = 1$ for all $s \in S$ and $\eta_i(0) = 0$ for all $i \in V\backslash S$. If S consists of a single node, the initial condition is called 'node seeding'; if S consists of a node together with its neighbours, the initial condition is called 'cluster seeding'. For a given homogeneous threshold T, node states are updated synchronously in discrete time steps according to the following rule. If $\eta_i(t) = 1$, then $\eta_i(t+1) = 1$. If $\eta_i(t) = 0$, then

$$\eta_i(t+1) = 1 \text{ if and only if } f_i > T\,, \qquad \text{where} \quad f_i = \frac{1}{d}\sum_{j\in V} A_{ij}\eta_j(t)\,.$$

In other words, a node becomes activated at a time step if the fraction of its neighbours that are active is larger than T at the previous time step. Once a node is active, it stays active forever. For a fixed homogeneous threshold T and a given seed set S, this contagion is a deterministic and monotonic process, which eventually reaches a stable state in which either all of the nodes are active or some nodes are inactive and will never activate. We refer to this deterministic process as a *realization* of the contagion. Formally, a realization R is the nested sequence of subsets of V that are active at successive time steps: $R = \{S = S_0(S), S_1(S), \ldots, S_m(S), \ldots\}$ such that $S_t = \{i \in V \mid \eta_i(t) = 1\}$. Note that, due to the deterministic nature of the process, $S_0 = S$ determines S_m for all $m > 0$.

Chapter 4

Contagion Maps for Spreading Dynamics

4.1 Methods

We construct a point cloud by mapping the network nodes to points in $\mathbb{R}^N$ based on their activation times in different realizations of the contagion dynamics. This so-called *contagion map* is inspired by approaches from nonlinear dimensionality reduction such as diffusion map and Isomap [4, 20, 34, 69, 72], and was first studied in [71]. A point cloud that is the image of a contagion map can be interpreted as a distortion of the underlying network structure that reflects the contagion dynamics. We analyse the structure of this point cloud from three different perspectives — topologically, geometrically, and with respect to dimensionality — and compare it to the structure of the underlying network. We expect the structure of the point cloud in $\mathbb{R}^N$ to resemble the structure of the underlying network when the contagion spreads predominantly via WFP. We perform a bifurcation analysis[1] to identify regions in the parameter space for which WFP is the predominant mode of propagation, and we use the results of this analysis to validate that the structure of the underlying network is recovered in the point cloud whenever the contagion spreads predominantly via

[1]Note that the term 'bifurcation' is usually used for a number of more classical phenomena in dynamical systems. The type of bifurcation that we study in this work is different in flavour from the more classical kind, but is, nonetheless, technically a bifurcation.

WFP.

In order to compare the topology, geometry, and dimensionality of a point cloud to the structure of the network on which it is based, we need to specify this structure precisely. Specifically, we need to choose a metric space associated with the network and we need to specify the locations of the network's nodes in this metric space. We explore different choices of metric on the torus in the Appendix A.

We consider the torus as the Cartesian product of two circles:

$$\mathbb{T} = \frac{1}{2\pi}\mathbb{S}^1 \times \frac{1}{2\pi}\mathbb{S}^1 \subset \mathbb{C} \times \mathbb{C} \cong \mathbb{R}^4 \,. \tag{4.1}$$

We assume that the nodes of our network are 'evenly distributed' on this torus $\mathbb{T}$, forming a regular lattice. The $N = n^2$ nodes of our network are the points on $\mathbb{T}$ with coordinates

$$\frac{1}{2\pi}\left(\cos\frac{2\pi x}{n}, \sin\frac{2\pi x}{n}, \cos\frac{2\pi y}{n}, \sin\frac{2\pi y}{n}\right), \qquad x, y \in \{0, 1, \ldots, n-1\}\,. \tag{4.2}$$

4.1.1 Contagion maps

Consider our contagion, with homogeneous threshold T, on one instantiation of our Kleinberg-like network for some fixed parameter values n (yielding $N = n^2$ nodes), p (determining the number of geometric edges), q (determining the number of non-geometric edges), and γ (determining the 'reach' of the non-geometric edges), as defined in Section 3.2.1. For a given seed set, the contagion dynamics is a deterministic process, and we can record the activation times of the nodes. We consider several realizations of the contagion dynamics initialized with different seeds. We denote the set of realizations by $J = \{R_1, R_2, \ldots, R_{|J|}\}$, and we denote the activation time of node $i \in V$ in realization $R_j \in J$ by $x_j^{(i)}$. If node i is never activated in realization R_j, we set $x_j^{(i)} = 2N$ (i.e. larger than any actual activation time).

The *regular contagion map* associated with the set J of realizations is a function

from the set V of nodes to $\mathbb{R}^{|J|}$. It is defined by

$$i \mapsto x^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_{|J|}^{(i)}] .$$

The regular contagion map associated with J maps each node in V to a vector in $\mathbb{R}^J$ that records its activation times in each of the different realizations.

We take J to be the same size as V and choose the seed sets to be the clusters around the different nodes, such that the seed that initializes realization $R_j \in J$ is $S^{(j)} = \{j\} \cup \{k \mid A_{jk} \neq 0\}$. In this case, the activation time of node i in realization R_j is a proxy for a distance between nodes i and j. To see this, consider the realization of a contagion with homogeneous threshold $T = 0$ initialized with a single seed node $\{j\}$, and observe that the activation time of node i is exactly the length of a shortest path between i and j. For cluster seeding of our contagion, the activation time of node i in realization j may not be precisely the shortest-path distance between i and j; it depends on how the contagion spreads. Moreover, $x_j^{(i)} \neq x_i^{(j)}$ in general[2]. With this in mind, we define the *reflected contagion map*, which maps $i \mapsto y^{(i)} = [x_i^{(1)}, x_i^{(2)} \ldots, x_i^{(|J|)}]$, and the *symmetric contagion map*, which maps $i \mapsto [x_1^{(i)} + x_i^{(1)}, \ldots, x_{|J|}^{(i)} + x_i^{(|J|)}]$.

The image of a contagion map is a point cloud in $\mathbb{R}^{|J|}$ that can be interpreted as a distortion of a network that takes into account contagion dynamics for different realizations of a contagion.

4.1.2 Geometry

To quantify the similarity of the geometric structure of a contagion map to that of the network on which it is based, we calculate the Pearson correlation coefficient of

[2]Consider, for example, the network with adjacency matrix $\begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$ and our contagion with threshold $T = 0.4$. Then $2 = x_1^{(2)} \neq x_2^{(1)} = 1$.

pairwise distances between points in the contagion point cloud and pairwise distances between corresponding nodes of the network. We use Euclidean distance in $\mathbb{R}^4$ for the nodes and Euclidean distance in $\mathbb{R}^N$ for points in the point cloud.

Recall that the network nodes lie on the torus (4.1) at points with coordinates (4.2). Let

$$w^{(i)} = \frac{1}{2\pi}\left(\cos\frac{2\pi i_x}{n}, \sin\frac{2\pi i_x}{n}, \cos\frac{2\pi i_y}{n}, \sin\frac{2\pi i_y}{n}\right)$$

denote the point in $\mathbb{T}$ that is associated with node $i = \overline{(i_x, i_y)}$. The distance between two such points (according to the $d_{2,2}$ metric as defined in Section A.2 in the Appendix) is

$$d\left(w^{(i)}, w^{(j)}\right) = \sqrt{\sum_{k=1}^{4}\left(w_k^{(i)} - w_k^{(j)}\right)^2} = \frac{1}{\pi}\left(\sin^2\frac{(i_x - j_x)\pi}{n} + \sin^2\frac{(i_y - j_y)\pi}{n}\right)^{\frac{1}{2}},$$

and the distance between the corresponding points, $x^{(i)}$ and $x^{(j)}$, in the point cloud is

$$d(x^{(i)}, x^{(j)}) = \sqrt{\sum_{k=1}^{N}\left(x_k^{(i)} - x_k^{(j)}\right)^2}.$$

Given ordered sets, D^{net} and D^{map}, of pairwise distances between nodes of the network and points in the point cloud, respectively, we compute the Pearson correlation coefficient between these sets:

$$\rho = \frac{\sum_{i=1}^{N}\sum_{j=i+1}^{N}\left[d'(w^{(i)}, w^{(j)})d'(x^{(i)}, x^{(j)})\right]}{\sqrt{\sum_{i=1}^{N}\sum_{j=i+1}^{N} d'(w^{(i)}, w^{(j)})^2}\sqrt{\sum_{i=1}^{N}\sum_{j=i+1}^{N} d'(x^{(i)}, x^{(j)})^2}}, \tag{4.3}$$

where

$$d'(w^{(i)}, w^{(j)}) = d(w^{(i)}, w^{(j)}) - \overline{d(w^{(i)}, w^{(j)})},$$

and

$$d'(x^{(i)}, x^{(j)}) = d(x^{(i)}, x^{(j)}) - \overline{d(x^{(i)}, x^{(j)})},$$

with

$$\overline{d(w^{(i)}, w^{(j)})} = \frac{\sum_{i=1}^{N} \sum_{j=i+1}^{N} d(w^{(i)}, w^{(j)})}{(N^2 - N)/2}$$

denoting the mean pairwise distance between nodes and $\overline{d(x^{(i)}, x^{(j)})}$ denoting the mean pairwise distance between points in the contagion map. The scaling by the variances of both sets of pairwise distances in (4.3) compensates for any artefacts arising from the difference in dimension between the ambient space of the network nodes and the ambient space of the contagion point cloud. Progressively larger Pearson correlation coefficients ρ indicate progressively more similar geometric structures between a contagion map and its associated network.

4.1.3 Topology

We examine the topology of a contagion map by studying the persistent homology of the Vietoris–Rips filtration (see Definition 10 in Chapter 2) on its associated point cloud. We calculate persistent homology using the software package Ripser[3]. We seek to quantify the extent to which topological features of a torus manifest in the barcode that represents persistent homology of a contagion point cloud at a given dimension. To do this, we calculate the Wasserstein distance $W_2[d_2]$ (see Definition 8 in Chapter 2) between this barcode and a 'model barcode' that represents topological features of a torus in the given dimension.

It is not automatically clear what exactly it means for a barcode to 'represent the topological features of a torus'. Intuitively, such a barcode has two dominant bars in dimension 1 and one dominant bar in dimension 2, capturing the topological features of a torus in these dimensions. One way to obtain such a 'model torus-barcode' is to compute the persistent homology of a Vietoris–Rips filtration on a regular point cloud on a torus according to some underlying metric. There are many different metrics that one can choose for this purpose, including those that are intrinsic to the torus

[3]Ripser is publicly available at `https://github.com/Ripser/ripser`.

and those that are defined on its ambient space. We explore some of these options in the Appendix A. The 'model barcode' representing the topological features of a torus that we choose is the one corresponding to the persistent homology of the Vietoris–Rips filtration on the regular point cloud on the torus described in formula (4.2) with the Euclidean metric (see Appendix A.2). Smaller Wasserstein distances correspond to more 'torus-like' point clouds, recovering the topology of the manifold in which the network nodes are embedded. We will work with networks of $N = 50 \times 50$ nodes (see Section 3.2.1). Due to the computational complexity of computing the two-dimensional persistent homology of a Vietoris–Rips filtration on 2500 points (it involves up to 1.6263×10^{12} simplices to be built), we compute persistent homology only up to dimension 1, which requires building only up to 2.6042×10^{9} simplices for a given point cloud.

The Wasserstein distance between barcodes is sensitive to scaling. Consider, for instance, two barcodes that have the same number of bars, such that the relative lengths of the bars within each barcode are the same, but the overall bar lengths differ between them. Although these two barcodes represent the exact same topological features — albeit of different sizes — the Wasserstein distance between them is nonzero (see Figure 4.1). Similarly, two barcodes that represent very similar topological features, but are at very different scales, may be at a larger Wasserstein distance from each other than two barcodes that represent different features but are close in 'scale'[4]. See Figure 4.2 (a)–(c) for an illustration of this phenomenon.

In the present application, this sensitivity to scaling can manifest as follows. The 'model torus' barcode corresponding to regularly spaced points on a torus constructed as the Cartesian product of two circles of circumference 1 has relatively short bars. For our contagion, larger values of T entail slower spreading. Therefore, if we have

[4]Our use of the term 'scale' differs from existing uses in topological data analysis. An example use of 'scale' in topological data analysis is for a point in a filtration.

two contagion maps that both arise from spreading via WFP without ANC, but one with large T (entailing slow propagation) and the other with small T (entailing fast propagation), then both contagion maps have the same (torus-like) shape, but the former yields a point cloud that is much 'larger' than that of the latter. This implies, in turn, that the former's corresponding barcode is farther away from the model torus barcode than is the case for the latter. Similarly, when there are many non-geometric edges, there is fast spreading via ANC. Therefore, although the shape of the contagion map should not look torus-like in this case, but rather like a cluster of points, the Wasserstein distance from the corresponding barcode to the model torus barcode may still be small, simply by virtue of the size of the point cloud, rather than because of its shape.

To counteract this effect, we 'calibrate' all barcodes before taking the Wasserstein distance. We find the longest bar in each barcode and divide the birth and death times of all bars by that length. This yields barcodes whose longest bar has length 1, so they can be considered to be at the same 'scale'. Consequently, the Wasserstein distance between these calibrated barcodes can serve as a measure for comparing the topological features of the corresponding point clouds without taking absolute distances (i.e. geometry) into account (see Figure 4.2 (d)–(f)).

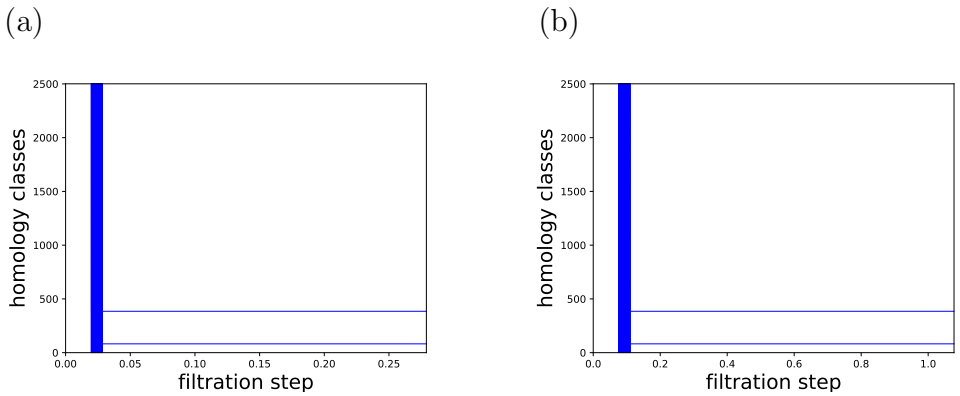


Figure 4.1: (a) Uncalibrated and (b) calibrated barcodes for the persistent homology of the Vietoris–Rips filtration on the regularly spaced point cloud on the torus from formula (4.2). The Wasserstein distance between the two barcodes is ~ 51.55, although they are identical aside from scale.

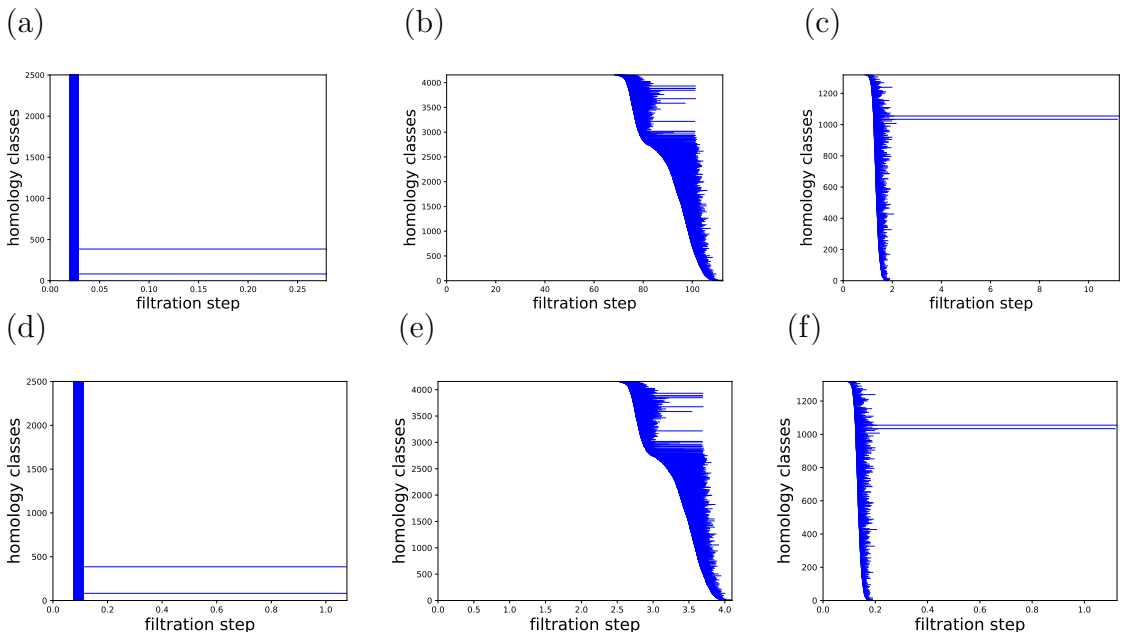


Figure 4.2: Illustrative examples to demonstrate the sensitivity of the Wasserstein distance to barcode 'scales'. Barcodes (d), (e), and (f) are the calibrated versions of barcodes (a), (b), and (c), respectively. Barcodes (a) and (c) (and, consequently, (d) and (f)) represent 1-dimensional topological features of a torus, whereas barcode (b) (and, consequently, (e)) does not. However, in the uncalibrated versions of the barcodes, the Wasserstein distance $W_2[d_2]$ between barcodes (a) and (b) is $\sim$ 12.01, and the Wasserstein distance $W_2[d_2]$ between barcodes (a) and (c) is $\sim$ 138.86, suggesting that (a) and (b) — rather than (a) and (c) — have similar topological features. By contrast, the Wasserstein distances between the calibrated versions of the barcodes illustrate the true topological proximities: ~ 350.41 between barcodes (d) and (e) and ~ 51.85 between barcodes (d) and (f).

To compute the Wasserstein distance, we use the software package HERA[5]. HERA currently provides the fastest algorithm for computing Wasserstein distances, outperforming the one implemented in the DIONYSUS library [43]. See Figure 4.3 for examples of calibrated barcodes representing various spreading regimes on an instantiation of our network model.

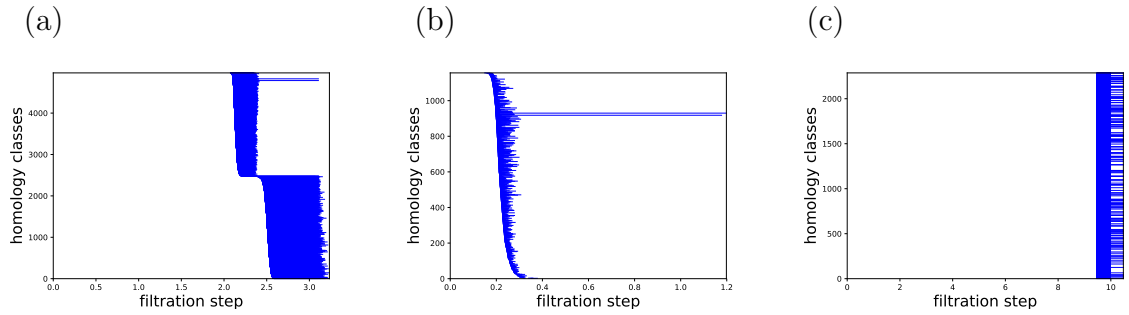


Figure 4.3: Calibrated barcodes for the persistent homology of the Vietoris–Rips filtrations on contagion maps arising from a Kleinberg-like small-world network with parameter values $N = 2500$, $d^{\mathrm{G}} = 8$, $d^{\mathrm{NG}} = 2$ and $\gamma = 0$ for different values for the contagion threshold T. In (a), $T = 0.05$, i.e. a small threshold for which fast spreading via both WFP and ANC is expected. In (b), $T = 0.25$, a threshold for which ANC is unlikely and spreading is expected to follow WFP. In (c), $T = 0.4$, i.e. a large threshold for which no spreading is expected to occur. (See Figure 4.25 to locate these parameter values in the (d^{NG}, T)-parameter space and identify their respective associated spreading regimes.) Note that in panel (b), the barcode corresponding to the parameter combination for which we expect predominantly spreading via WFP, we observe two relatively long bars, which represent the one-dimensional topological features of the torus. In panel (a), the torus features show up to some extent (in the form of two slightly dominant bars), despite the fast spreading via ANC due to the small value of T. This shows that WFP can still affect contagion maps noticeably, even in the presence of ANC. In panel (c), there are many bars of equal birth and length. This stems from the fact that for this parameter combination no spreading is expected, and so most nodes do not become infected in most realizations of the contagion. Hence, for most points in the contagion map between 9 and 11 coordinates are 0 (for the realizations that include them in their seed) and the other coordinates are $2N$ (for the realizations in which they do not become activated). In other words, most points lie on corners of a 2500-dimensional hypercube in the first orthant of $\mathbb{R}^{2500}$. The Wasserstein distances from the calibrated barcode representing the persistent homology of the Vietoris–Rips filtration on the regularly spaced point cloud on the torus (see Figure 4.1) are ~ 943.25 in panel (a), ~ 52.01 in panel (b), and ~ 669.68 in panel (c).

[5]HERA is publicly available at `https://bitbucket.org/grey_narn/hera`.

4.1.4 Dimensionality

We determine the approximate embedding dimension P of a point cloud by finding the smallest dimension such that we lose less than 5% of the variance when projecting to that dimension using principal component analysis (PCA) [69]. That is, for each $p \in \{1, 2, \dots\}$, we project the point cloud $\{x^{(i)} \in \mathbb{R}^N\}_{i \in V}$ to $\mathbb{R}^p$ using PCA, resulting in a point cloud $\{\hat{x}_p^{(i)} \in \mathbb{R}^p\}_{i \in V}$.

We estimate the extent to which this projection preserves the original point cloud by calculating the residual variance [23]

$$R_p = 1 - \left(\rho^{(p)}\right)^2 ,$$

where $\rho^{(p)}$ is the Pearson correlation coefficient between the pairwise Euclidean distances of points in $\{x^{(i)} \in \mathbb{R}^N\}_{i \in V}$ and corresponding pairwise Euclidean distances between points in $\{\hat{x}_p^{(i)} \in \mathbb{R}^p\}_{i \in V}$ (see Section 4.1.2). The approximate embedding dimension P is the smallest dimension for which the residual variance is less than 5%; that is, $P = \min\{p \mid R_p < 0.05\}$.

In practice, we put a cap of 100 on P, so if the approximate embedding dimension is 100 or larger, we record it as 100. As we consider the torus to be embedded in $\mathbb{R}^4$, an approximate embedding dimension of $P = 4$ indicates that the contagion map recovers the dimensionality of the torus.

4.2 Results

In Section 4.2.1, we explore contagion maps on a variety of small networks that are based on a periodic lattice as defined in Section 3.2.1. Initially, we consider the non-noisy network, and we subsequently add non-geometric edges, first by adding a single non-geometric edge, then by adding non-geometric edges in the orientation of one side of the lattice, then by adding non-geometric edges in the orientation of the

diagonal of the lattice, and finally by adding non-geometric edges randomly according to the procedure described in Section 3.2.1. In each case, we compute the contagion map, build a Vietoris–Rips filtration on it, and compute its persistent homology. By examining and comparing the corresponding barcodes, we obtain information on the spreading behaviour of a contagion on the different versions of this lattice-based network. In the case of the non-noisy network, we compute contagion maps by hand.

In Sections 4.2.2 and 4.2.3, we perform numerical experiments on large 50-by-50 networks of the class defined in Section 3.2.1. We study the shape of contagion maps arising from networks with varying non-geometric degree d^{NG} and varying contagion threshold T (Section 4.2.2), as well as of those arising from networks constructed using varying values of the parameter γ, which regulates the range of non-geometric edges (Section 4.2.3).

Finally, in Section 4.2.4, we conduct a bifurcation analysis for the spreading behaviour of our threshold contagion and compare it to our numerical findings from Section 4.2.2.

4.2.1 Explorations on various small networks

This section contains preliminary calculations on some small-scale networks. The main purpose of these calculations is to get a feel for the behaviour of contagion maps on lattice-based noisy geometric networks. The reader may want to skip this section and proceed to Section 4.2.2, which contains numerical results for larger Kleinberg-like small-world networks.

4.2.1.1 Non–noisy periodic lattice

Consider our network model (see Section 3.2.1) with $N = n^2$ nodes (for some integer n), geometric degree $d^{\mathrm{G}} = 4$ (i.e. $p = 1$), and non-geometric degree $d^{\mathrm{NG}} = 0$. That is, consider an $n \times n$ square lattice that is 'wrapped up' into a torus. Having no noisy

edges, this is a purely geometric network, and it is regular of degree 4.

We run our threshold contagion model with cluster seeding (see Section 3.2.2) on this network for various values of the threshold T. For $T \geq 0.5$, no nodes other than the seed nodes become active. That is, the contagion does not spread at all. For $0.25 \leq T < 0.5$ another four additional nodes are activated at time step 1, but the dynamics stop after that. For all thresholds $0 \leq T < 0.25$, the contagion exhibits identical behaviour. In this case, an inactive node becomes active at a given time step, if at least one of its neighbours was active at the previous time step. At each time step, additional nodes are activated until all nodes are active at time step $n-2$ (when n is odd), or time step $n-1$ (when n is even). The activation time of a node is one less than the periodic lattice distance from this node to the node around which the seed is clustered.

Now consider the case when $n = 5$, i.e. the network has $N = 25$ nodes, and $0 \leq T < 0.25$ (see Figure 4.4).

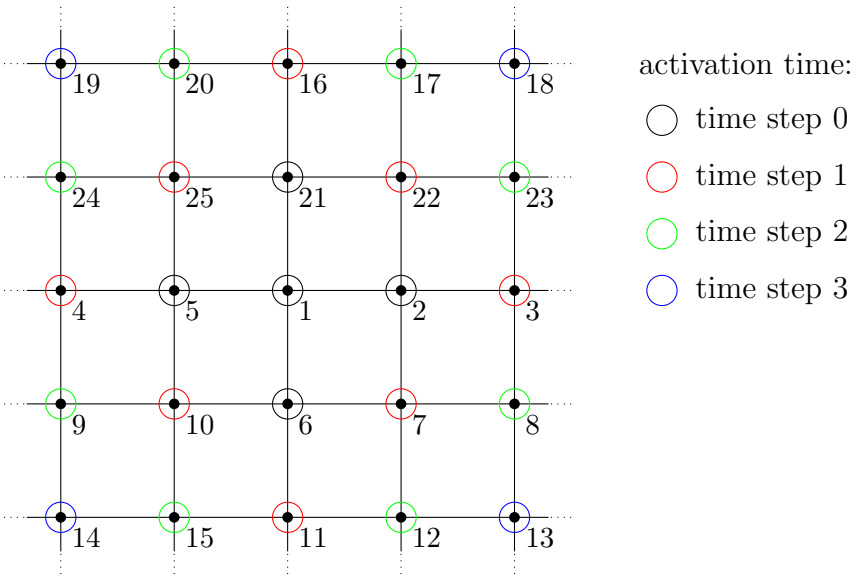


Figure 4.4: Cartoon depicting the spreading of a cluster-seeded threshold contagion with threshold $0 \leq T < 0.25$ on a 5×5 'wrapped up' lattice. Nodes are successively activated starting from the five seed nodes $1, 2, 5, 6$ and 21 until all nodes are active at time step 3.

For each of the 25 nodes, we run the contagion starting from a cluster seed around this node and record the activation times of all nodes. Running the contagion starting from a cluster seed around node 1 gives the following activation times:

node	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
$\mathbf{x_j^{(i)}}$	0	0	1	1	0	0	1	2	2	1	1	2	3	3	2	1	2	3	3	2	0	1	2	2	1

By symmetry, the list of activation times that arises from running the contagion starting with a cluster seed around another node is just a permutation of this list of activation times. The permutation matrix for shifting the seed cluster one node to the right is the following:

$$S_{\text{right}} = \begin{pmatrix}
0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0
\end{pmatrix}$$

The permutation matrix for shifting the seed cluster one node down is the following:

$$S_{\text{down}} = \begin{pmatrix}
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1\\
1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0
\end{pmatrix}$$

We can record the activation times of all nodes in each realization of the contagion in a matrix D_{reg} as follows. Each row of D_{reg} holds the activation times of all nodes when the contagion is seeded around the node that is labeled by the number of the row. That is, entry $(D_{\text{reg}})_{i,j}$ is the activation time of node j when the contagion is seeded around node i:

$$D_{\text{reg}} = \begin{pmatrix}
0 & 0 & 1 & 1 & 0 & 0 & 1 & 2 & 2 & 1 & 1 & 2 & 3 & 3 & 2 & 1 & 2 & 3 & 3 & 2 & 0 & 1 & 2 & 2 & 1 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 2 & 2 & 2 & 1 & 2 & 3 & 3 & 2 & 1 & 2 & 3 & 3 & 1 & 0 & 1 & 2 & 2 \\
1 & 0 & 0 & 0 & 1 & 2 & 1 & 0 & 1 & 2 & 3 & 2 & 1 & 2 & 3 & 3 & 2 & 1 & 2 & 3 & 2 & 1 & 0 & 1 & 2 \\
1 & 1 & 0 & 0 & 0 & 2 & 2 & 1 & 0 & 1 & 3 & 3 & 2 & 1 & 2 & 3 & 3 & 2 & 1 & 2 & 2 & 2 & 1 & 0 & 1 \\
0 & 1 & 1 & 0 & 0 & 1 & 2 & 2 & 1 & 0 & 2 & 3 & 3 & 2 & 1 & 2 & 3 & 3 & 2 & 1 & 1 & 2 & 2 & 1 & 0 \\
0 & 1 & 2 & 2 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 2 & 2 & 1 & 1 & 2 & 3 & 3 & 2 & 1 & 2 & 3 & 3 & 2 \\
1 & 0 & 1 & 2 & 2 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 2 & 2 & 2 & 1 & 2 & 3 & 3 & 2 & 1 & 2 & 3 & 3 \\
2 & 1 & 0 & 1 & 2 & 1 & 0 & 0 & 0 & 1 & 2 & 1 & 0 & 1 & 2 & 3 & 2 & 1 & 2 & 3 & 3 & 2 & 1 & 2 & 3 \\
2 & 2 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 2 & 2 & 1 & 0 & 1 & 3 & 3 & 2 & 1 & 2 & 3 & 3 & 2 & 1 & 2 \\
1 & 2 & 2 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 2 & 2 & 1 & 0 & 2 & 3 & 3 & 2 & 1 & 2 & 3 & 3 & 2 & 1 \\
1 & 2 & 3 & 3 & 2 & 0 & 1 & 2 & 2 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 2 & 2 & 1 & 1 & 2 & 3 & 3 & 2 \\
2 & 1 & 2 & 3 & 3 & 1 & 0 & 1 & 2 & 2 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 2 & 2 & 2 & 1 & 2 & 3 & 3 \\
3 & 2 & 1 & 2 & 3 & 2 & 1 & 0 & 1 & 2 & 1 & 0 & 0 & 0 & 1 & 2 & 1 & 0 & 1 & 2 & 3 & 2 & 1 & 2 & 3 \\
3 & 3 & 2 & 1 & 2 & 2 & 2 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 2 & 2 & 1 & 0 & 1 & 3 & 3 & 2 & 1 & 2 \\
2 & 3 & 3 & 2 & 1 & 1 & 2 & 2 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 2 & 2 & 1 & 0 & 2 & 3 & 3 & 2 & 1 \\
1 & 2 & 3 & 3 & 2 & 1 & 2 & 3 & 3 & 2 & 0 & 1 & 2 & 2 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 2 & 2 & 1 \\
2 & 1 & 2 & 3 & 3 & 2 & 1 & 2 & 3 & 3 & 1 & 0 & 1 & 2 & 2 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 2 & 2 \\
3 & 2 & 1 & 2 & 3 & 3 & 2 & 1 & 2 & 3 & 2 & 1 & 0 & 1 & 2 & 1 & 0 & 0 & 0 & 1 & 2 & 1 & 0 & 1 & 2 \\
3 & 3 & 2 & 1 & 2 & 3 & 3 & 2 & 1 & 2 & 2 & 2 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 2 & 2 & 1 & 0 & 1 \\
2 & 3 & 3 & 2 & 1 & 2 & 3 & 3 & 2 & 1 & 1 & 2 & 2 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 2 & 2 & 1 & 0 \\
0 & 1 & 2 & 2 & 1 & 1 & 2 & 3 & 3 & 2 & 1 & 2 & 3 & 3 & 2 & 0 & 1 & 2 & 2 & 1 & 0 & 0 & 1 & 1 & 0 \\
1 & 0 & 1 & 2 & 2 & 2 & 1 & 2 & 3 & 3 & 2 & 1 & 2 & 3 & 3 & 1 & 0 & 1 & 2 & 2 & 0 & 0 & 0 & 1 & 1 \\
2 & 1 & 0 & 1 & 2 & 3 & 2 & 1 & 2 & 3 & 3 & 2 & 1 & 2 & 3 & 2 & 1 & 0 & 1 & 2 & 1 & 0 & 0 & 0 & 1 \\
2 & 2 & 1 & 0 & 1 & 3 & 3 & 2 & 1 & 2 & 3 & 3 & 2 & 1 & 2 & 2 & 2 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
1 & 2 & 2 & 1 & 0 & 2 & 3 & 3 & 2 & 1 & 2 & 3 & 3 & 2 & 1 & 1 & 2 & 2 & 1 & 0 & 0 & 1 & 1 & 0 & 0
\end{pmatrix}$$

Taking the rows (or columns, by symmetry) of this matrix to be the coordinate vectors of points in $\mathbb{R}^{25}$ gives the image of the regular contagion map (see Section 4.1.1). We compute the persistent homology of the Vietoris–Rips filtration on this point cloud in dimensions $0, 1$ and 2 with PERSEUS[6], using a step size of $\Delta = \frac{1}{3}\min_{i,j\in V}\left(d(x^{(i)}, x^{(j)})\right)$, and $\left\lceil \frac{\max_{i,j\in V}\left(d(x^{(i)},x^{(j)})\right)}{\Delta} \right\rceil$ steps (see barcodes in Figure 4.5).

[6]PERSEUS is publicly available at `http://people.maths.ox.ac.uk/nanda/perseus/index.html`.

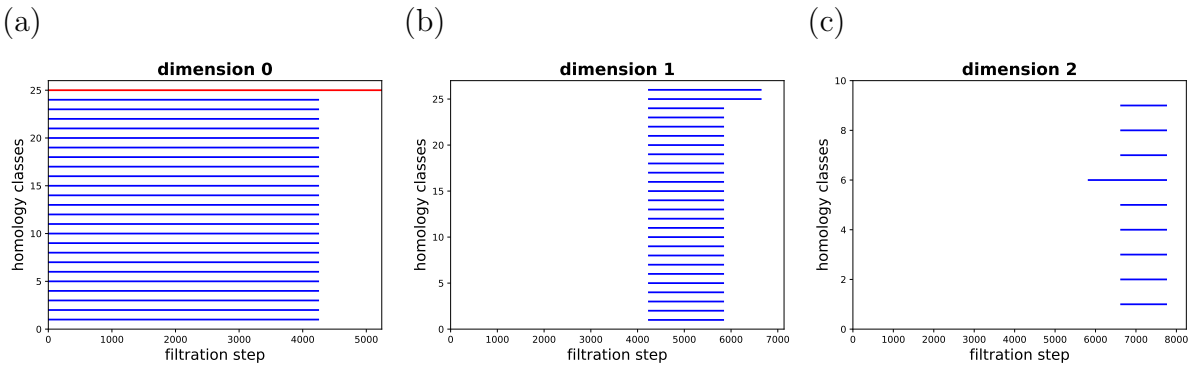


Figure 4.5: Barcodes for the persistent homology in dimension (a) 0, (b) 1, and (c) 2 of the Vietoris–Rips filtration on a regular contagion map for a periodic 5-by-5 lattice. (The red bar in (a) never dies.)

The barcode for dimension 0 (Figure 4.5 (a)) shows 25 homology classes. All of them are present at time 0, and all but one of them die simultaneously at one filtration step, leaving the final one to live forever. At the same filtration step, 26 homology classes in dimension 1 are born (see barcode for dimension 1 in Figure 4.5 (b)). Two of these classes live slightly longer than the others, which all die at the same filtration step. The barcode for dimension 2 shows 9 homology classes, one of which lives slightly longer than the others. Precisely, the dominant class is born with the death of 24 of the 1-dimensional classes, while the other 8 classes are born with the death of the two dominant 1-dimensional classes.

These barcodes are consistent with our interpretation of the image of a contagion map as a distortion of the underlying network according to the spreading dynamics. As our network is purely geometric, completely symmetric, and can be thought of as being embedded in a torus, our contagion spreads uniformly by WFP along this underlying torus in a way that is essentially the same for all realizations, regardless of the node around which the contagion seed is clustered. (As explained above, activation times are just permutations of each other in the different realizations.) The contagion map is therefore expected to produce a completely symmetric point cloud in $\mathbb{R}^{25}$ that has toroidal structure. The barcodes corresponding to the persistent

homology groups of the Vietoris–Rips filtration on this point cloud can be explained as follows. The 25 homology classes in dimension 0 correspond to the 25 points, (the complexes around) which merge eventually to produce one component that is shaped like a thickening of a torus with 25 holes in its 'surface'. These 25 holes together with the two loops that are intrinsic to the torus are captured by 26 homology classes in dimension 1[7]. Once the 'superficial' holes close, only the two classes corresponding to the toroidal loops remain, and a void, captured by the slightly dominant homology class in dimension 2, is created. Eventually, the two toroidal loops close, forming 8 additional voids, which are finally filled simulataneously with the dominant void that is intrinsic to the toroidal structure of the point cloud.

The matrix corresponding to the symmetric contagion map (see Section 4.1.1 is the sum of the matrix corresponding to the regular contagion map and its transpose. In this case, by symmetry, it is just the same matrix with doubled entries:

$$D_{\text{sym}} = \begin{pmatrix}
0 & 0 & 2 & 2 & 0 & 0 & 2 & 4 & 4 & 2 & 2 & 4 & 6 & 6 & 4 & 2 & 4 & 6 & 6 & 4 & 0 & 2 & 4 & 4 & 2 \\
0 & 0 & 0 & 2 & 2 & 2 & 0 & 2 & 4 & 4 & 4 & 2 & 4 & 6 & 6 & 4 & 2 & 4 & 6 & 6 & 2 & 0 & 2 & 4 & 4 \\
2 & 0 & 0 & 0 & 2 & 4 & 2 & 0 & 2 & 4 & 6 & 4 & 2 & 4 & 6 & 6 & 4 & 2 & 4 & 6 & 4 & 2 & 0 & 2 & 4 \\
2 & 2 & 0 & 0 & 0 & 4 & 4 & 2 & 0 & 2 & 6 & 6 & 4 & 2 & 4 & 6 & 6 & 4 & 2 & 4 & 4 & 4 & 2 & 0 & 2 \\
0 & 2 & 2 & 0 & 0 & 2 & 4 & 4 & 2 & 0 & 4 & 6 & 6 & 4 & 2 & 4 & 6 & 6 & 4 & 2 & 2 & 4 & 4 & 2 & 0 \\
0 & 2 & 4 & 4 & 2 & 0 & 0 & 2 & 2 & 0 & 0 & 2 & 4 & 4 & 2 & 2 & 4 & 6 & 6 & 4 & 2 & 4 & 6 & 6 & 4 \\
2 & 0 & 2 & 4 & 4 & 0 & 0 & 0 & 2 & 2 & 2 & 0 & 2 & 4 & 4 & 4 & 2 & 4 & 6 & 6 & 4 & 2 & 4 & 6 & 6 \\
4 & 2 & 0 & 2 & 4 & 2 & 0 & 0 & 0 & 2 & 4 & 2 & 0 & 2 & 4 & 6 & 4 & 2 & 4 & 6 & 6 & 4 & 2 & 4 & 6 \\
4 & 4 & 2 & 0 & 2 & 2 & 2 & 0 & 0 & 0 & 4 & 4 & 2 & 0 & 2 & 6 & 6 & 4 & 2 & 4 & 6 & 6 & 4 & 2 & 4 \\
2 & 4 & 4 & 2 & 0 & 0 & 2 & 2 & 0 & 0 & 2 & 4 & 4 & 2 & 0 & 4 & 6 & 6 & 4 & 2 & 4 & 6 & 6 & 4 & 2 \\
2 & 4 & 6 & 6 & 4 & 0 & 2 & 4 & 4 & 2 & 0 & 0 & 2 & 2 & 0 & 0 & 2 & 4 & 4 & 2 & 2 & 4 & 6 & 6 & 4 \\
4 & 2 & 4 & 6 & 6 & 2 & 0 & 2 & 4 & 4 & 0 & 0 & 0 & 2 & 2 & 2 & 0 & 2 & 4 & 4 & 4 & 2 & 4 & 6 & 6 \\
6 & 4 & 2 & 4 & 6 & 4 & 2 & 0 & 2 & 4 & 2 & 0 & 0 & 0 & 2 & 4 & 2 & 0 & 2 & 4 & 6 & 4 & 2 & 4 & 6 \\
6 & 6 & 4 & 2 & 4 & 4 & 4 & 2 & 0 & 2 & 2 & 2 & 0 & 0 & 0 & 4 & 4 & 2 & 0 & 2 & 6 & 6 & 4 & 2 & 4 \\
4 & 6 & 6 & 4 & 2 & 2 & 4 & 4 & 2 & 0 & 0 & 2 & 2 & 0 & 0 & 2 & 4 & 4 & 2 & 0 & 4 & 6 & 6 & 4 & 2 \\
2 & 4 & 6 & 6 & 4 & 2 & 4 & 6 & 6 & 4 & 0 & 2 & 4 & 4 & 2 & 0 & 0 & 2 & 2 & 0 & 0 & 2 & 4 & 4 & 2 \\
4 & 2 & 4 & 6 & 6 & 4 & 2 & 4 & 6 & 6 & 2 & 0 & 2 & 4 & 4 & 0 & 0 & 0 & 2 & 2 & 2 & 0 & 2 & 4 & 4 \\
6 & 4 & 2 & 4 & 6 & 6 & 4 & 2 & 4 & 6 & 4 & 2 & 0 & 2 & 4 & 2 & 0 & 0 & 0 & 2 & 4 & 2 & 0 & 2 & 4 \\
6 & 6 & 4 & 2 & 4 & 6 & 6 & 4 & 2 & 4 & 4 & 4 & 2 & 0 & 2 & 2 & 2 & 0 & 0 & 0 & 4 & 4 & 2 & 0 & 2 \\
4 & 6 & 6 & 4 & 2 & 4 & 6 & 6 & 4 & 2 & 2 & 4 & 4 & 2 & 0 & 0 & 2 & 2 & 0 & 0 & 2 & 4 & 4 & 2 & 0 \\
0 & 2 & 4 & 4 & 2 & 2 & 4 & 6 & 6 & 4 & 2 & 4 & 6 & 6 & 4 & 0 & 2 & 4 & 4 & 2 & 0 & 0 & 2 & 2 & 0 \\
2 & 0 & 2 & 4 & 4 & 4 & 2 & 4 & 6 & 6 & 4 & 2 & 4 & 6 & 6 & 2 & 0 & 2 & 4 & 4 & 0 & 0 & 0 & 2 & 2 \\
4 & 2 & 0 & 2 & 4 & 6 & 4 & 2 & 4 & 6 & 6 & 4 & 2 & 4 & 6 & 4 & 2 & 0 & 2 & 4 & 2 & 0 & 0 & 0 & 2 \\
4 & 4 & 2 & 0 & 2 & 6 & 6 & 4 & 2 & 4 & 6 & 6 & 4 & 2 & 4 & 4 & 4 & 2 & 0 & 2 & 2 & 2 & 0 & 0 & 0 \\
2 & 4 & 4 & 2 & 0 & 4 & 6 & 6 & 4 & 2 & 4 & 6 & 6 & 4 & 2 & 2 & 4 & 4 & 2 & 0 & 0 & 2 & 2 & 0 & 0
\end{pmatrix}$$

As with the regular contagion map, we compute the persistent homology of the Vietoris–Rips filtration on this point cloud in dimensions $0, 1$ and 2 with Perseus, using a step size of $\Delta = \frac{1}{3} \min_{i,j \in V} \left(d(x^{(i)}, x^{(j)}) \right)$, and $\left\lceil \frac{\max_{i,j \in V} \left(d(x^{(i)}, x^{(j)}) \right)}{\Delta} \right\rceil$ steps (see barcodes in Figure 4.6).

[7]Note that $\mathbb{T} \backslash \{k \text{ points}\} \simeq \mathbb{S}^1 \vee \mathbb{S}^1 \vee \bigvee_{k-1} \mathbb{S}^1$, so $H_1 \left(\mathbb{T} \backslash \{k \text{ points}\} \right) \cong \mathbb{Z}^{k+1}$ and $\beta_1 \left(\mathbb{T} \backslash \{k \text{ points}\} \right) = k + 1$.

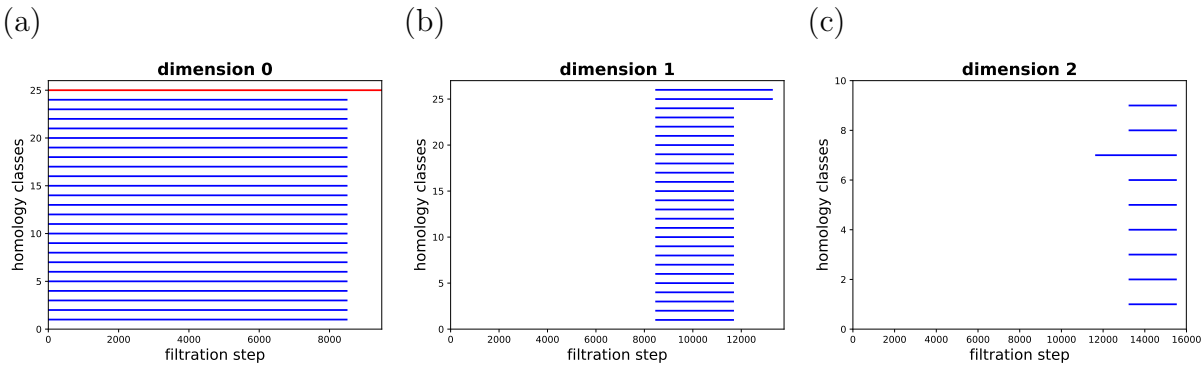


Figure 4.6: Barcodes for the persistent homology in dimension (a) 0, (b) 1, and (c) 2 of the Vietoris–Rips filtration on a symmetric contagion map for a periodic 5-by-5 lattice. (The red bar in (a) never dies.)

The barcodes corresponding to the symmetric contagion map are simply the barcodes corresponding to the regular contagion map stretched by a factor of 2, as expected.

4.2.1.2 Noisy networks based on a periodic lattice

Now consider an $n \times n$ periodic lattice, as in Section 4.2.1.1, that is supplemented with long-range edges to form a noisy geometric network (see Section 3.1.1). The long-range (noisy) edges can be added in different ways and we will consider the following cases: (1) Adding a single noisy edge that connects the top left node in the lattice to the node in the centre of the lattice (i.e. a noisy edge that goes halfway around both loops on the torus); (2) adding noisy edges randomly following a probability distribution that favours edges between nearby nodes (similarly to our network model from Section 3.2.1); (3) adding noisy edges in the orientation of the side of the lattice (i.e. in the orientation of one of the loops on the torus); and (4) adding noisy edges in the orientation of the diagonal of the lattice (i.e. in the orientation that goes around both loops on the torus 'simultaneously').

As in Section 4.2.1.1, we consider a threshold contagion that is seeded in a node and its direct neighbours and whose update rule is that an inactive node becomes

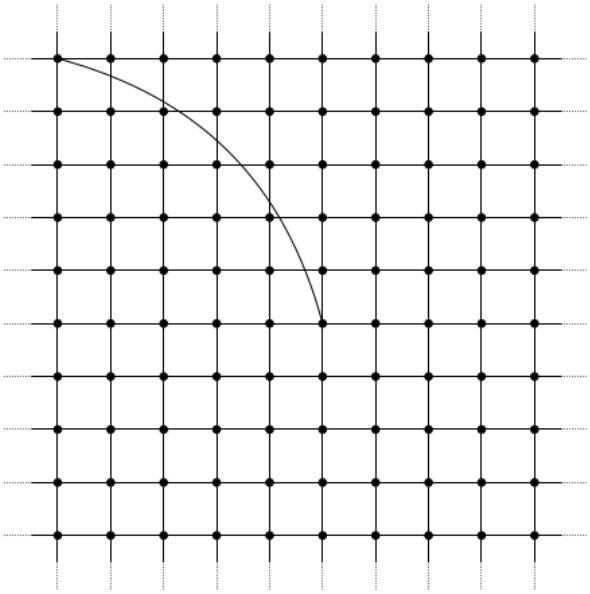


Figure 4.7: Sketch of a 10×10 lattice with a single noisy edge added.

activated at a time step if the fraction of its neighbours that were active in the previous time step is strictly greater than a fixed threshold T (see Section 3.2.2). For each type of noisy geometric network, we record the activation times for all nodes in all realizations of the contagion that are seeded in such a cluster around a single node, and we thus compute the symmetric contagion map. We build Vietoris–Rips filtrations on the resulting point clouds and compare the barcodes corresponding to their persistent homology groups to the barcodes resulting from a non-noisy lattice.

Lattice with a single noisy edge added. We take an $n \times n$ periodic lattice, as in Section 4.2.1.1, and add a single noisy edge going from the node in the top left corner of the lattice (with coordinates $\overline{(1,1)}$) to the node in the centre of the lattice (with coordinates $\overline{(\frac{n}{2}+1, \frac{n}{2}+1)}$ if n is even, and $\overline{(\frac{n}{2}+\frac{1}{2}, \frac{n}{2}+\frac{1}{2})}$ if n is odd). Figure 4.7 shows a sketch of this network when $n = 10$.

We then compute the symmetric contagion map with a threshold of $T = 0$ for

both this noisy network and the non-noisy base network, and compute the persistent homology groups in dimensions 1 and 2 of the Vietoris–Rips filtrations on the resulting point clouds using PERSEUS (with step size and number of steps chosen as in Section 4.2.1.1). For the case $n = 10$ (i.e. 100 nodes), the resulting barcodes are shown in Figure 4.8. We observe an additional bar in both dimensions 1 and 2 for the network with the single noisy edge added, as compared to the non-noisy periodic lattice. This makes sense, because, for our chosen threshold $T = 0$, the contagion is a simple contagion that is able to spread via a single edge. The added noisy edge thus adds a loop to the toroidal structure of the image of the contagion map.

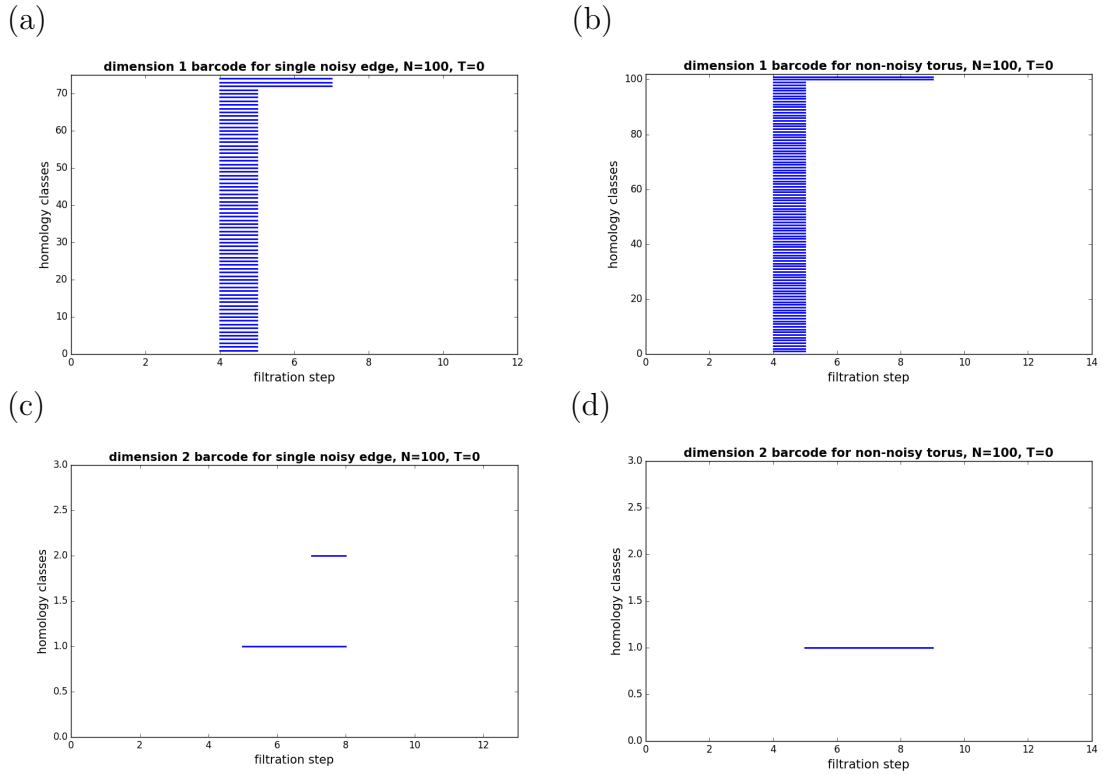


Figure 4.8: Barcodes for Vietoris–Rips filtrations on symmetric contagion maps on 10×10 noisy (panels (a) and (c)) versus non-noisy (panels (b) and (d)) periodic lattice networks. Dimension 1 barcodes for the periodic lattice network with a single noisy edge added (a) and the non-noisy periodic lattice network (b). Dimension 2 barcodes for the periodic lattice network with a single noisy edge added (c) and the non-noisy periodic lattice network (d).

Lattice with noisy edges added randomly. As before, we start with an $n \times n$ lattice, as in Section 4.2.1.1. We then add noisy edges randomly according to a probablity distribution that favours edges between nodes that are close to each other, similarly to our network model in Section 3.2.1. This noisy geometric network is defined by the following parameters:

- $n \in \mathbb{N}$: The 'side length' of the lattice. So the network has $N = n^2$ nodes.
- $p \in \mathbb{R}_{>0}$: The 'length' of the geometric edges. Two nodes have a geometric edge between them if the (periodic) Euclidean distance between them is at most p.
- $q_{\text{prop}} \in [0, 1]$: The 'proportion' of non-geometric edges. The number of non-geometric edges being placed is $q_{\text{prop}} \cdot N$ rounded to the nearest natural number. So the number of non-geometric edges is between 0 and n^2 (i.e. at most half of the number of geometric edges if $p = 1$). However, as multi-edges are only counted once, $\text{round}(q_{\text{prop}} \cdot N) \leq n^2$ is really an upper bound for the actual number of non–geometric edges.
- $\gamma \in \mathbb{R}_{\geq 0}$: The exponent in the probability distribution that determines which nodes are connected by the non-geometric egdes. The probability for two non-geometric stubs emanating from nodes i and j, respectively, is proportional to $\mu_{\text{per}}(i, j)^{-\gamma}$, where μ_{per} is the periodic lattice distance between i and j.

Figure 4.9 shows a sketch of such a network with $n = 10, p = 1, q_{\text{dens}} = 0.2$ and $\gamma = 0$. Note that, although $q_{\text{prop}} \cdot N = 0.2 \cdot 100 = 20$ non-geometric edges are placed, only 18 non-geometric show up effectively due to the presence of multi-edges.

Computing the symmetric contagion map with a threshold of $T = 0$ on this network (with parameters $n = 10, q_{\text{dens}} = 0.2$ and $\gamma = 0$) and computing the persistent homology groups in dimensions 1 and 2 of the Vietoris–Rips filtrations on the resulting point clouds using Perseus (with step size and number of steps chosen as

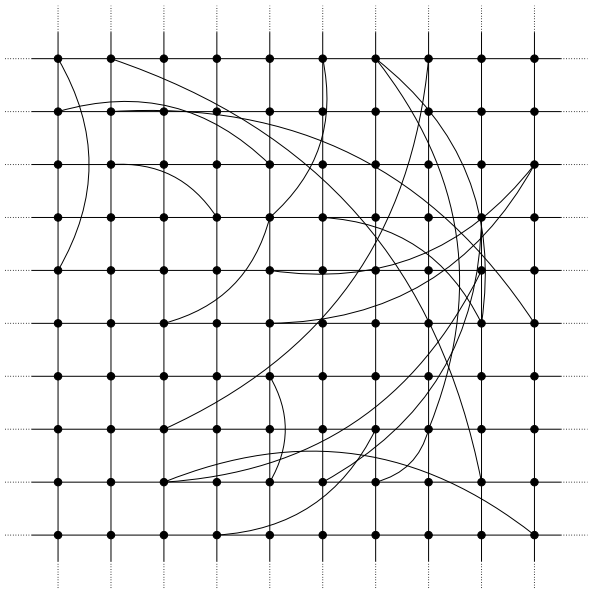


Figure 4.9: Sketch of a 10×10 lattice with noisy edges added randomly.

in Section 4.2.1.1) yields the barcodes in Figure 4.10 (a) and (c). We observe that, due to the addition of the randomly placed non-geometric edges in the network, the toroidal structure is no longer detectable in the image of the contagion map.

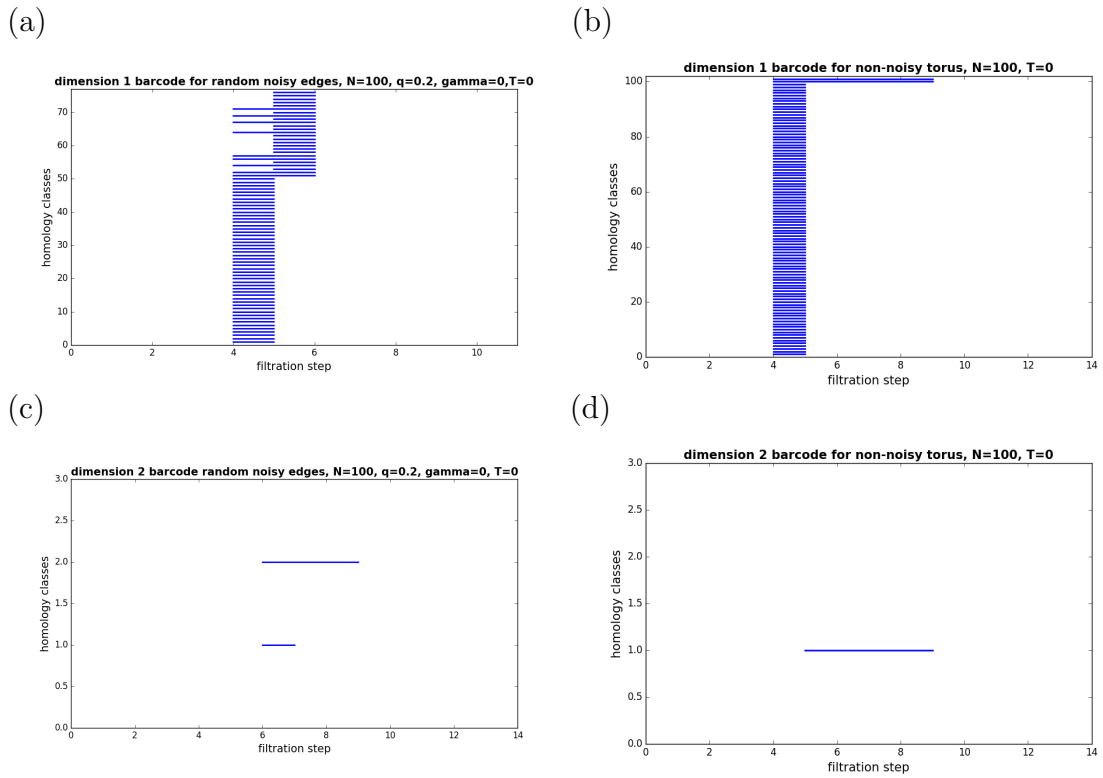


Figure 4.10: Barcodes for Vietoris–Rips filtrations on symmetric contagion maps on 10×10 noisy (panels (a) and (c)) versus non-noisy (panels (b) and (d)) periodic lattice networks. Dimension 1 barcodes for the periodic lattice network with noisy edges added randomly (a) and the non-noisy periodic lattice network (b). Dimension 2 barcodes for the periodic lattice network with noisy edges added randomly (c) and the non-noisy periodic lattice network (d).

Lattice with noisy edges added parallel to one side. As before, we start with an $n \times n$ lattice, as in Section 4.2.1.1. We then add a certain number of non-geometric edges uniformly at random, such that their span is fixed to be a of certain 'length', and their direction is parallel to one of the sides of the lattice. This noisy geometric network is defined by the following parameters:

- $n \in \mathbb{N}$: The 'side length' of the lattice. So the network has $N = n^2$ nodes.
- $p \in \mathbb{R}_{>0}$: The 'length' of the geometric edges. Two nodes have a geometric edge between them if the (periodic) Euclidean distance between them is at most p.

- $m_{\text{par}} \in \mathbb{N}_0$: The total number of non-geometric edges in the network. Multi-edges are counted only once, so the actual number of non-geometric edges may be less than m_{par}.

- $l_{\text{par}} \in \{1, 2, \ldots, n\}$: The 'length' of the non-geometric edges. A non-geometric edge can connect node $\overline{(i_x, i_y)}$ to node $\overline{(i_x + l_{\text{par}}, i_y)}$.

Figure 4.11 shows a sketch of such a network for $n = 10, p = 1, m_{\text{par}} = 10$, and $l_{\text{par}} = 5$.

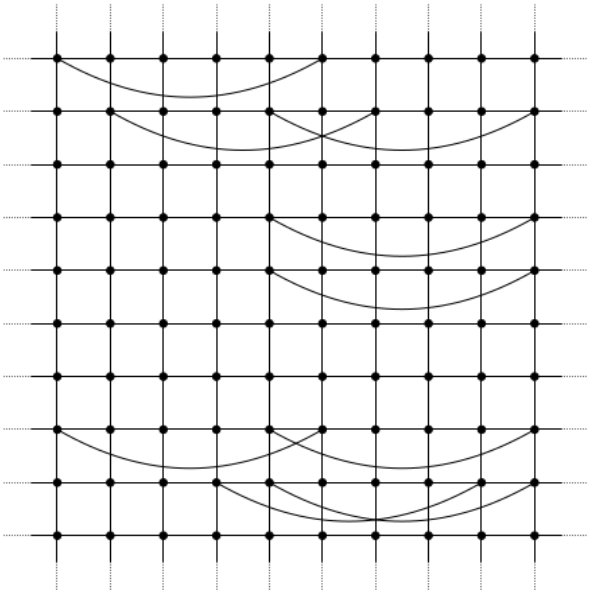


Figure 4.11: Sketch of a 10×10 lattice with noisy edges added parallel to one of the sides.

Figure 4.12 (a) and (c) show barcodes corresponding to the symmetric contagion map with threshold $T = 0$ on such a network with parameters $n = 10, m_{\text{par}} = 10$ and $l_{\text{par}} = 5$ (computed in the same way as in previous sections). We observe that one of the two feature loops of the torus is not detected anymore in the barcode for dimension 1, leaving only one dominant bar in this barcode (see Figure 4.12 (a)). Similarly, there are no dominant bars in the barcode for dimension 2, indicating that there are no feature voids in the image of the contagion map. The non-geometric

edges form shortcuts across one loop of the torus that underlies our network. In fact, at a 'length' of 5, they run exactly halfway across the lattice; in other words, they go through this loop. From the perspective of a simple contagion (with threshold $T = 0$), the non-geometric edges thus 'close' one toroidal loop, essentially transforming the network into a ring. This explains why the barcodes based on this network detect only one loop in dimension 1 and no void in dimension 2.

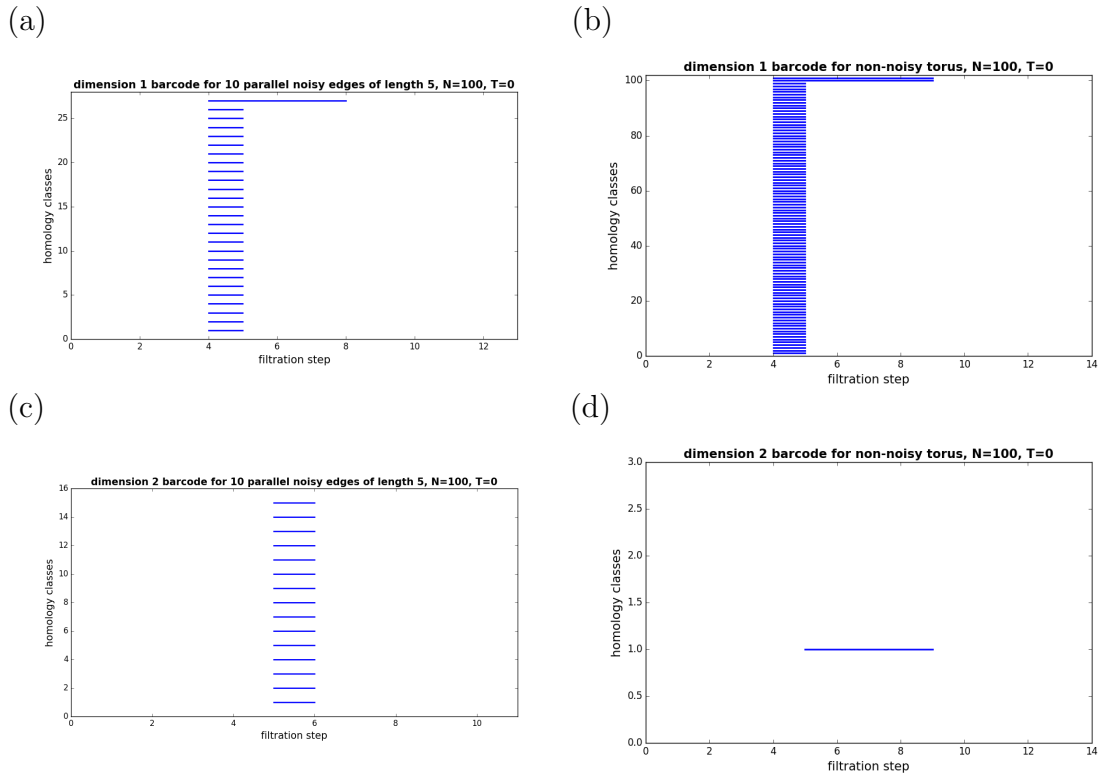


Figure 4.12: Barcodes for Vietoris–Rips filtrations on symmetric contagion maps on 10×10 noisy (panels (a) and (c)) versus non-noisy (panels (b) and (d)) periodic lattice networks. Dimension 1 barcodes for the periodic lattice network with noisy edges added parallelly to a side of the lattice (a) and the non-noisy periodic lattice network (b). Dimension 2 barcodes for the periodic lattice network with noisy edges added parallelly to a side of the lattice (c) and the non-noisy periodic lattice network (d).

Lattice with noisy edges added diagonally. As before, we start with an $n \times n$ lattice, as in Section 4.2.1.1. We then add a certain number of non-geometric edges uniformly at random, such that their span is fixed to be of a certain 'length' and

they go diagonally across the lattice. This noisy geometric network is defined by the following parameters:

- $n \in \mathbb{N}$: The 'side length' of the lattice. So the network has $N = n^2$ nodes.
- $p \in \mathbb{R}_{>0}$: The 'length' of the geometric edges. Two nodes have a geometric edge between them if the (periodic) Euclidean distance between them is at most p.
- $m_{\text{diag}} \in \mathbb{N}_0$: The total number of non-geometric edges placed in the network. Multi-edges are counted only once, so the actual number of non-geometric edges may be less than m_{diag}.
- $l_{\text{diag}} \in \{1, 2, \ldots, n\}$: The 'length' of the non-geometric edges. A non-geometric edge can connect node $\overline{(i_x, i_y)}$ to node $\overline{(i_x + l_{\text{diag}}, i_y + l_{\text{diag}})}$.

Figure 4.13 shows a sketch of such a network for $n = 10, p = 1, m_{\text{diag}} = 10$ and $l_{\text{diag}} = 5$.

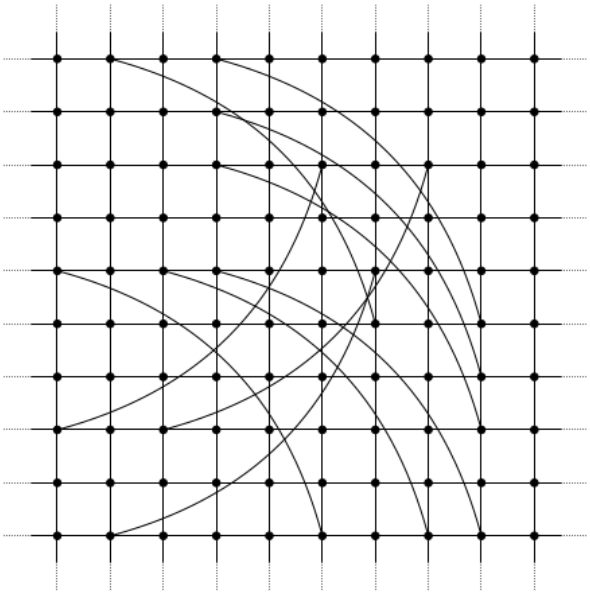


Figure 4.13: Sketch of a 10×10 lattice with noisy edges added diagonally to the sides.

Figure 4.14 (a) and (c) show barcodes corresponding to the symmetric contagion map with threshold $T = 0$ on such a network with parameters $n = 10, m_{\text{diag}} = 10$, and $l_{\text{diag}} = 5$ (computed in the same way as in previous sections). We observe an additional bar in dimension 1.

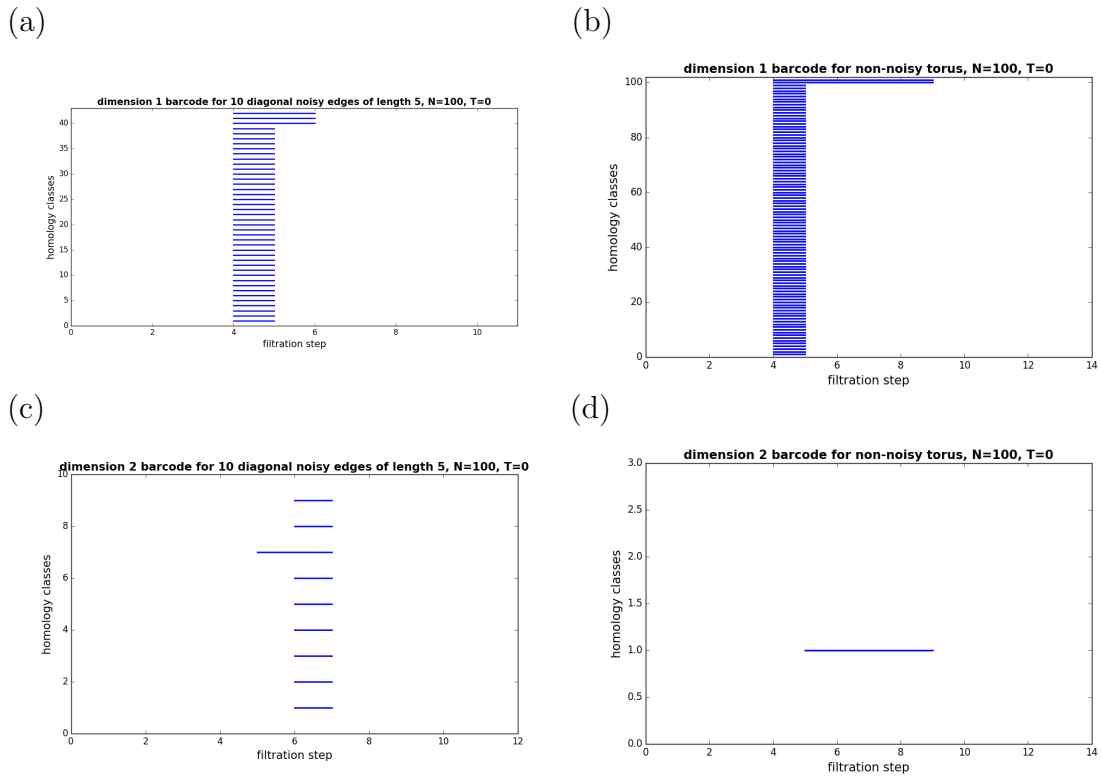


Figure 4.14: Barcodes for Vietoris–Rips filtrations on symmetric contagion maps on 10×10 noisy (panels (a) and (c)) versus non-noisy (panels (b) and (d)) periodic lattice networks. Dimension 1 barcodes for the periodic lattice network with noisy edges added diagonally to a side of the lattice (a) and the non-noisy periodic lattice network (b). Dimension 2 barcodes for the periodic lattice network with noisy edges added diagonally to a side of the lattice (c) and the non-noisy periodic lattice network (d).

4.2.2 Numerical results for 50-by-50 networks

We construct Kleinberg-like small-world networks (as detailed in Section 3.2.1) for the following parameter values: $N = 2500$ nodes (i.e. $n = 50$), geometric degrees $d^{\text{G}} = 4, 8, 12$ (corresponding to $p = 1, \sqrt{2}, 2$), non-geometric degrees $d^{\text{NG}} = 0, 1, 2, \ldots, 25$, and distance decay parameter $\gamma = 0, 0.1, 0.2, 0.3, \ldots, 3$. For each of these $3 \times 26 \times 31 =$

2418 networks and for each threshold value $T = 0, 0.01, 0.02, 0.03, \ldots, 1$, we run our contagion model (see Section 3.2.2) with cluster seeding around each of its 2500 nodes, recording the activation times of each node. Using the activation times of each node in each of these realizations as coordinates, we map the nodes of each network to a point cloud in $\mathbb{R}^{2500}$ via the symmetric contagion map (see Section 4.1.1).

We compute quantitative measures of the similarity of these point clouds to the underlying torus in terms of geometry (see Section 4.1.2), topology (see Section 4.1.3), and dimensionality (see Section 4.1.4) when we place non-geometric edges uniformly at random (i.e. when $\gamma = 0$). We illustrate our results by separately displaying the values of the Pearson correlation coefficient ρ, the Wasserstein distance $W_2[d_2]$, and the embedding dimension P in the (d^{NG}, T)-parameter space for each value of d^{G} (see Figure 4.15). When examining topological similarity, we cover only the case $d^{\mathrm{G}} = 8$, as computing persistent homology of the Vietoris–Rips filtration on a point cloud of 2500 points is extremely time-consuming because of the large number of simplices involved. Brighter regions in our plots signify larger Pearson correlation coefficients ρ (in the geometry computation), smaller Wasserstein distances $W_2[d_2]$ (in the topology computation), and lower approximate embedding dimensions P. In each plot, we can identify a region in parameter space for which ρ is large, and $W_2[d_2]$ and P are small, indicating that WFP dominates for these parameter values.

The first column of each plot in Figure 4.15 shows our results for $d^{\mathrm{NG}} = 0$, which corresponds to a network being purely geometric. In this case, the network formation is fully deterministic and contagion dynamics (in particular the presence versus absence of WFP) are easy to determine analytically. (See Section 4.2.4 for details.) We see in these first columns of the plots that ρ, $W_2[d_2]$, and P take only extreme values for $d^{\mathrm{NG}} = 0$, and that the transition between extreme values occurs at the same threshold T for all three measures. Below this threshold, the Pearson correla-

tion coefficients are large and the Wasserstein distances and approximate embedding dimensions are small ($P = 4$, to be precise). Above this threshold, the Pearson correlation coefficients are small and P is large (at the cap of 100), and these values of T are marked to yield 'infinite activation times' of nodes in the plot for the Wasserstein distance. The observations described in this paragraph are consistent with our analytical considerations, which demonstrate that spreading (by WFP) can occur only below this threshold.

There is a band along the transition between the region in which we expect WFP (see Figures 4.19–4.21) — and observe associated structure of the contagion map (see Figures 4.15) — and the region in which we do not. This band is dark in the plots of the geometric and the topological structure, and it is bright in the plot depicting dimensionality. This implies that for the corresponding parameter combinations the point cloud is low-dimensional, but that it does not exhibit torus-like structure in terms of geometry or topology. Although this feature was not discussed in [71], one also observes such a band (see their figures and Figure 4.23 of the present paper) for a threshold contagion on the noisy ring lattice that they studied.

Some irregularities and outliers in our figures are likely due to the probabilistic nature of non-geometric edges in the network construction. One example is that of the non-white spots within the white region in Figure 4.15(e). These correspond to parameter combinations for which according to our bifurcation analysis (see Section 4.2.4) we expect infinite activation times, but (by chance) all nodes end up having finite activation times in all realizations.

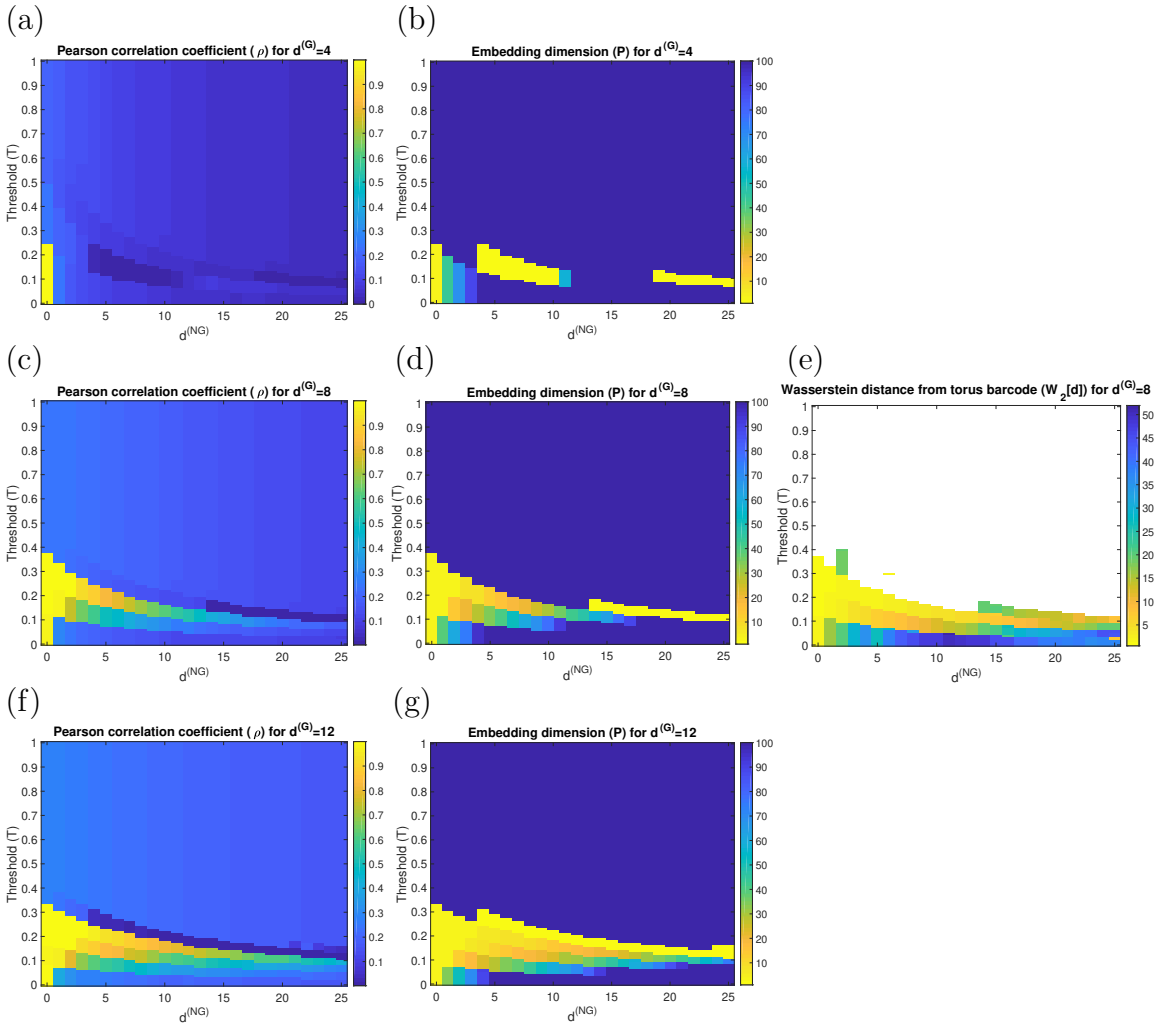


Figure 4.15: (a) Geometry (as quantified by the Pearson correlation coefficient) and (b) dimensionality (as determined by the approximate embedding dimension, capped at 100) of contagion maps from Kleinberg-like small-world networks with parameter values $N = 2500$, $d^{\mathrm{G}} = 4$, $\gamma = 0$, and $d^{\mathrm{NG}} = 0, 1, 2, \ldots, 25$ and contagion thresholds $T = 0, 0.01, 0.02, 0.03, \ldots, 1$. (c) Geometry (as quantified by the Pearson correlation coefficient) and (d) dimensionality (as determined by the approximate embedding dimension, capped at 100) of contagion maps from Kleinberg-like small-world networks with parameter values $N = 2500$, $d^{\mathrm{G}} = 8$, $\gamma = 0$, and $d^{\mathrm{NG}} = 0, 1, 2, \ldots, 25$ and contagion thresholds $T = 0, 0.01, 0.02, 0.03, \ldots, 1$. (e) Wasserstein distances between the scaled barcode of the persistent homology of the Vietoris–Rips filtration on the regularly spaced point cloud on the torus (see formula (4.2)) and the scaled barcodes of the persistent homology of the Vietoris–Rips filtrations on the contagion maps from Kleinberg-like small-world networks with parameter values $N = 2500$, $d^{\mathrm{G}} = 12$, $\gamma = 0$, and $d^{\mathrm{NG}} = 0, 1, 2, \ldots, 25$ and contagion thresholds $T = 0, 0.01, 0.02, 0.03, \ldots, 1$. The white regions correspond to parameter combinations for which there are nodes that do not activate (i.e. they have 'infinite' activation times) in some realizations of the contagion. (f) Pearson correlation coefficient and (g) approximate embedding dimension (capped at 100) of contagion maps from Kleinberg-like small-world networks with parameter values $N = 2500$, $d^{\mathrm{G}} = 8$, $\gamma = 0$, and $d^{\mathrm{NG}} = 0, 1, 2, \ldots, 25$ and contagion thresholds $T = 0, 0.01, 0.02, 0.03, \ldots, 1$.

4.2.3 Effect of the distance-decay parameter γ on contagion maps

In our Kleinberg-like small-world networks (see Section 3.2.1), recall that the range of non-geometric edges is regulated by the decay parameter $\gamma \in \mathbb{R}_{\geq 0}$. Each node has a fixed number of non-geometric stubs, and two stubs emanating from nodes i and j, respectively, are joined to form a non-geometric edge with a probability proportional to $\mu_{\text{per}}(i,j)^{-\gamma}$. That is, for $\gamma = 0$, the non-geometric stubs are matched uniformly at random, regardless of the distance between the corresponding nodes, so the length of the non-geometric edges can take any value with equal probability. For $\gamma > 0$, non-geometric edges have a bias to connect nodes that are close with respect to the periodic lattice distance. This bias becomes more pronounced as γ increases, so the larger the value of γ, the shorter the non-geometric edges tend to be. See Section 3.2.1 for a more detailed discussion on the role of γ on the spreading of a contagion.

We examine the effect of γ on the shape of a contagion map. For fixed values of the geometric degree d^{G} and non-geometric degree d^{NG} of our network and threshold T of our contagion, we vary the value of γ. Specifically, we choose a geometric degree of $d^{\text{G}} = 8$, a non-geometric degree of $d^{\text{NG}} = 2$, and one value for the contagion threshold T for each predicted spreading regime when $\gamma = 0$. We use the value $T = 0.05$ to represent the regime in which we expect both WFP and ANC when $\gamma = 0$, the value $T = 0.25$ to represent the regime in which we expect WFP but no ANC when $\gamma = 0$, and the value $T = 0.4$ to represent the regime in which we expect neither WFP nor ANC when $\gamma = 0$. See Figure 4.25 to locate these parameter values in the (d^{NG}, T)-parameter plane and identify their respective associated spreading regimes. For each of these three values for T, we let γ vary from 0 to 3 in increments of 0.1. For each value of γ, we map the nodes of the associated network via the contagion map using the given value of T and analyse the resulting point cloud as described in Section 4.1.

For $T = 0.05$, the Pearson correlation coefficient increases significantly in an almost linear fashion as γ increases, while the Wasserstein distance and the approximate embedding dimension both decrease. This can be attributed to the fact that non-geometric edges change in function from drivers of ANC to contributors to WFP. For $\gamma = 0$, we expect fast spreading of the contagion that is dominated by ANC. This leads to a contagion map whose image is a cluster of tightly bunched points that are distributed fairly evenly in the region that they occupy. In particular, the pairwise distances between the points are not influenced much by the pairwise distances between their corresponding nodes. Such a cluster of points has a high approximate embedding dimension P, because its points are distributed with roughly constant density across the region that they occupy, so the point cloud does not have a lower intrinsic dimension than its ambient space.

As γ increases, the non-geometric edges tend to become shorter and contribute increasingly to WFP, instead of facilitating spreading across large distances on the network, thus producing a point cloud that is still contained in a small volume (as the spreading is still fast with such a low threshold), but such that pairwise distances between points become increasingly faithful to the pairwise distances between their corresponding nodes.

For $T = 0.25$, the Pearson correlation coefficient starts out large and increases further for progressively larger γ. By contrast, the Wasserstein distance is small throughout the range of γ, with a slight decrease at the lower end of the range. The approximate embedding dimension is $P = 4$ throughout the range of γ. This relative stability of all three measures stems from the fact that, for this value of T, WFP dominates over ANC even when $\gamma = 0$, as the non-geometric edges are not sufficiently numerous to drive ANC. For progressively larger γ, the gradually shortening non-geometric edges only contribute increasingly to WFP.

For $T = 0.4$, the Pearson correlation coefficient and the approximate embedding dimension remain, respectively, fairly small and large throughout the range of γ. As γ increases, the Wasserstein distance decreases steadily throughout the range of γ.

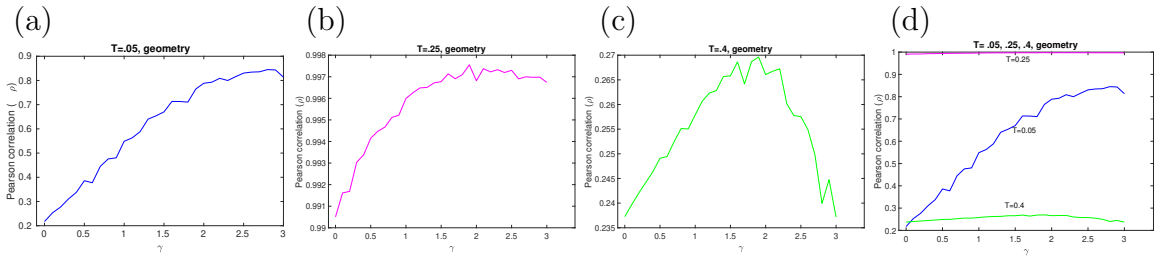


Figure 4.16: Pearson correlation coefficient between point-cloud distances and node–node distances as we increase γ from 0 to 3 in increments of 0.1 for Kleinberg-like small-world networks with parameters $N = 2500$, $d^{\mathrm{G}} = 8$, and $d^{\mathrm{NG}} = 2$ and contagion thresholds of (a) $T = 0.05$, (b) $T = 0.25$, and (c) $T = 0.4$. In panel (d), we show the plots for all three values of T.

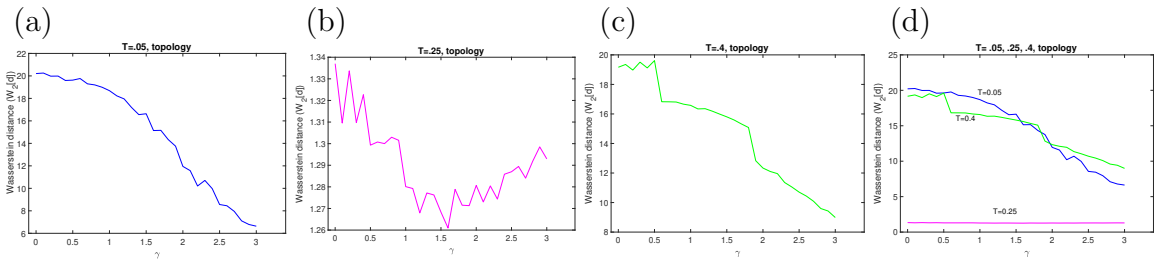


Figure 4.17: Wasserstein distance between scaled barcodes as we increase γ from 0 to 3 in increments of 0.1 for Kleinberg-like small-world networks with parameters $N = 2500$, $d^{\mathrm{G}} = 8$, and $d^{\mathrm{NG}} = 2$ and contagion thresholds of (a) $T = 0.05$, (b) $T = 0.25$, and (c) $T = 0.4$. In panel (d), we show the plots for all three values of T.

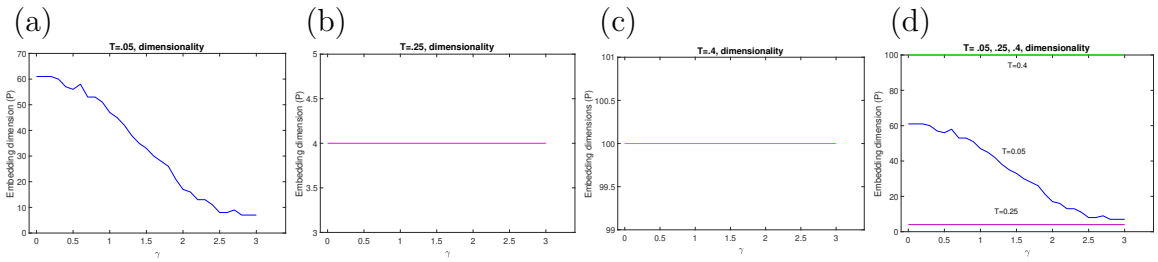


Figure 4.18: Approximate embedding dimension as we increase γ from 0 to 3 in increments of 0.1 for Kleinberg-like small-world networks with parameters $N = 2500$, $d^{\mathrm{G}} = 8$, and $d^{\mathrm{NG}} = 2$ and contagion thresholds of (a) $T = 0.05$, (b) $T = 0.25$, and (c) $T = 0.4$. In panel (d), we show the plots for all three values of T.

4.2.4 Bifurcation analysis

We conduct a bifurcation analysis of the spreading behaviour of the threshold contagion (see Section 3.2.2 for its definition) initialized with cluster seeding on the family of Kleinberg-like small-world networks that we described in Section 3.2.1. The results of this bifurcation analysis give a guideline for interpreting our prior numerical computations. We want to determine analytically which combinations of network parameter values, d^{G} and d^{NG} (see Section 3.2.1), and threshold parameter value T (see Section 3.2.2) allow the contagion to spread by WFP and which allow it to spread by ANC. That is, we want to identify regions in parameter space for which the spreading dynamics follow specific regimes that are characterized by the presence and absence of WFP and ANC. We are especially interested in the region of parameter space for which there is WFP but no ANC, as this region should comprise the parameter combinations for which the contagion map exhibits structural features of a torus.

We consider $N = 2500$ exclusively, as — at least locally and sufficiently early in the contagion process — the total size of a network should not affect contagion behaviour. At later stages, a contagion process that saturates a network will speed up earlier for smaller networks, as the active region of the network is now proportionately larger with respect to the total network size. Additionally, we restrict our analysis to the case where $\gamma = 0$ (i.e. we place non-geometric edges uniformly at random).

We fix the geometric degree d^{G} and examine the spreading behaviour as we vary the non-geometric degree d^{NG} and the threshold T. The possible values for d^{G} are constrained by the number of nodes that are within a distance $p \in \mathbb{R}_{>0}$ of a given node (see Section 3.2.1). In other words, for a given $p \in \mathbb{R}_{>0}$, the corresponding d^{G} is 1 less than the number of integer lattice points that lie inside a circle of radius p that is centered at the origin. This number is approximately equal to the area of the circle, and the problem of determining it is known as "Gauss' circle problem" [38].

The three smallest values of non-geometric degree d^{G} are 4, 8, and 12, corresponding to $1 \leq p < \sqrt{2}$, $\sqrt{2} \leq p < 2$, and $2 \leq p < \sqrt{5}$, respectively, in the definition of our Kleinberg-like small-world network.

4.2.4.1 Wavefront Propagation (WFP)

We consider the networks for $d^{\mathrm{G}} = 4, 8, 12$ individually, and we work out the maximal threshold for which the network can support sustained spreading via only geometric edges.

If $d^{\mathrm{G}} = 4$, then for WFP to occur, the threshold T needs to be small enough to allow spreading via a single edge. Therefore, for variable d^{NG}, for WFP to occur, the threshold T needs to be smaller than

$$T^{\mathrm{WFP}} = \frac{1}{4 + d^{\mathrm{NG}}} \,.$$

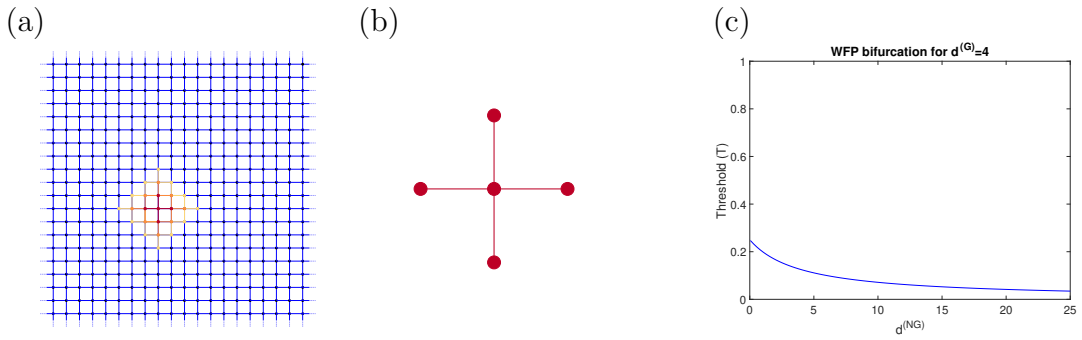


Figure 4.19: (a) Purely geometric network with geometric neighbours up to radius $r = 1$ around a node. (This corresponds to $p = 1$ and $q = 0$ in our Kleinberg-like small-world networks.) The geometric degree is $d^{\mathrm{G}} = 4$, and the non-geometric degree is $d^{\mathrm{NG}} = 0$. The contagion seed nodes are coloured dark, the nodes activated at the first time step are coloured medium, and nodes activated at the second time step are coloured light. (b) A node with its four direct neighbours, which are the nodes that are within Euclidean distance $p = 1$ from it. (c) Bifurcation diagram for the occurrence of WFP in a network with $d^{\mathrm{G}} = 4$. We vary the non-geometric degree d^{NG} from 0 to 25, and we vary the contagion threshold T from 0 to 1. WFP occurs only in the region below the curve.

If $d^{\mathrm{G}} = 8$, then for WFP to occur, the threshold T needs to be small enough to allow spreading via three edges. Therefore, for variable d^{NG}, for WFP to occur, the

threshold T needs to be smaller than

$$T^{\mathrm{WFP}} = \frac{3}{8 + d^{\mathrm{NG}}} \,.$$

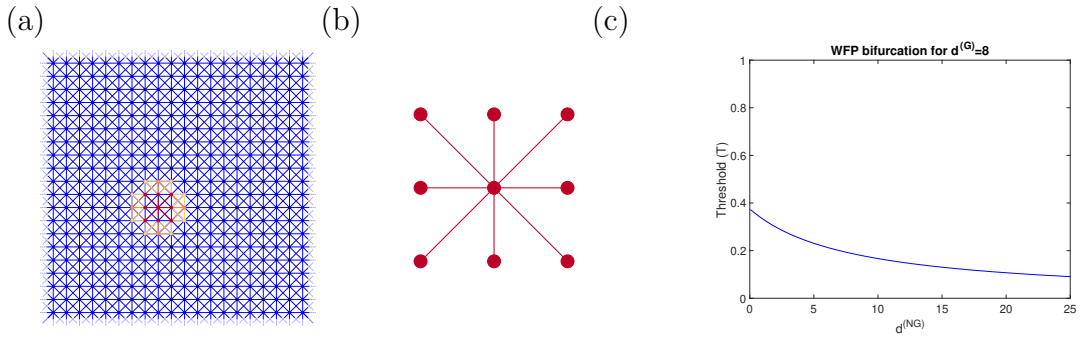


Figure 4.20: (a) Purely geometric network with geometric neighbours up to radius $r = \sqrt{2}$ around a node. (This corresponds to $p = \sqrt{2}$ and $q = 0$ in our Kleinberg-like small-world networks.) The geometric degree is $d^{\mathrm{G}} = 8$, and the non-geometric degree is $d^{\mathrm{NG}} = 0$. The contagion seed nodes are coloured dark, the nodes activated at the first time step are coloured medium, and nodes activated at the second time step are coloured light. (b) A node with its eight direct neighbours, which are the nodes that are within Euclidean distance $p = \sqrt{2}$ from it. (c) Bifurcation diagram for the occurrence of WFP in a network with $d^{\mathrm{G}} = 8$. We vary the non-geometric degree d^{NG} from 0 to 25, and we vary the contagion threshold T from 0 to 1. WFP occurs only in the region below the curve.

If $d^{\mathrm{G}} = 12$, then for WFP to occur, the threshold T needs to be small enough to allow spreading via four edges. Therefore, for variable d^{NG}, for WFP to occur, the threshold T needs to be smaller than

$$T^{\mathrm{WFP}} = \frac{4}{12 + d^{\mathrm{NG}}} \,.$$

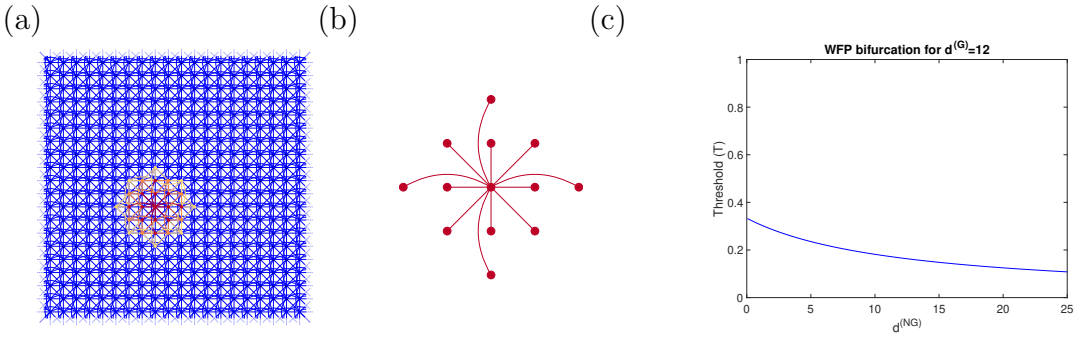


Figure 4.21: (a) Purely geometric network with geometric neighbours up to radius $r = 2$ around a node. (This corresponds to $p = 2$ and $q = 0$ in our Kleinberg-like small-world networks.) The geometric degree is $d^{\mathrm{G}} = 12$, and the non-geometric degree is $d^{\mathrm{NG}} = 0$. The contagion seed nodes are coloured dark, the nodes activated at the first time step are coloured medium, and nodes activated at the second time step are coloured light. (b) A node with its twelve direct neighbours, which are the nodes that are within Euclidean distance $p = 2$ from it. (c) Bifurcation diagram for the occurrence of WFP in a network with $d^{\mathrm{G}} = 12$. We vary the non-geometric degree d^{NG} from 0 to 25, and we vary the contagion threshold T from 0 to 1. WFP occurs only in the region below the curve.

There does not seem to be a closed form for T^{WFP} that holds for general values of d^{G}. One needs to find the maximum threshold that allows spreading by WFP for each value of d^{G} individually by finding the edges that can support spreading from the contagion seed.

4.2.4.2 Appearance of New Clusters (ANC)

The activation of an inactive node by ANC occurs, by definition, exclusively via non-geometric edges. That is, a node is activated via ANC if it is adjacent to at least $T \times (d^{\mathrm{G}} + d^{\mathrm{NG}})$ active nodes by non-geometric edges and all of its geometric neighbours are inactive. Consequently, if the threshold T is larger than or equal to the ratio of the non-geometric degree to the total degree (i.e. $T \geq \frac{d^{\mathrm{NG}}}{d^{\mathrm{G}}+d^{\mathrm{NG}}}$), then ANC is impossible. If $T < \frac{d^{\mathrm{NG}}}{d^{\mathrm{G}}+d^{\mathrm{NG}}}$, then ANC is possible. When $\frac{d^{\mathrm{NG}}-1}{d^{\mathrm{G}}+d^{\mathrm{NG}}} \leq T < \frac{d^{\mathrm{NG}}}{d^{\mathrm{G}}+d^{\mathrm{NG}}}$, ANC can occur in principle, but only if all of the non-geometric edges of an inactive node that has no active geometric neighbours 'reach into' contagion clusters. This is very unlikely to occur in practice, so a threshold T for which ANC is possible in

principle is not a good indicator in practice for the presence of ANC. We will explore this issue.

We define the *horizon*

$$H^{\mathrm{ANC}} = \frac{d^{\mathrm{NG}}}{d^{\mathrm{G}} + d^{\mathrm{NG}}}$$

of ANC to be the boundary between thresholds for which ANC is possible in theory and thresholds for which ANC is impossible. Using the horizon as a boundary curve for ANC generates an 'idealized' bifurcation diagram that tends to overestimate the size of the region of parameter space for which ANC occurs.

In practice, one needs to think about the probability of a number k among all non-geometric edges of a given node reaching into clusters of active nodes. If non-geometric edges are placed uniformly at random (which occurs when $\gamma = 0$ in the construction of our networks), the expected probability for a non-geometric edge of an inactive node to be incident to an active node at time t is $\frac{q(t)}{N-1}$, where $q(t)$ is the number of active nodes at time t. Consequently, if the non-geometric degree is d^{NG}, the expected number of non-geometric edges of an inactive node that are incident to active nodes is $\frac{q(t)}{N-1} d^{\mathrm{NG}}$. It follows that the maximal threshold for which one can expect every node that is inactive before time t to become activated via ANC at time t is

$$T = \frac{\frac{q(t)}{N-1} d^{\mathrm{NG}}}{d^{\mathrm{G}} + d^{\mathrm{NG}}} \,. \tag{4.4}$$

However, for ANC to occur at a certain time step, it is not necessary for every inactive node to become activated via ANC at that time step. It suffices for any inactive node that is sufficiently far away from the contagion to be activated via ANC, and the conditions on T for that to occur are generally less strict than it being smaller than (4.4). Consequently, we expect (4.4) to be a lower bound for T^{ANC}, the critical threshold for ANC to occur.

The numerator in (4.4) depends linearly on the number of active nodes $q(t)$, which

is time-dependent. This raises the question of what may be a sensible choice for t (and $q(t)$). Intuitively, if ANC occurs towards the end of a spreading process, when large parts of the network are already active, then its contribution to the spreading of the contagion is a minor one and it has only a negligible distortive effect on the contagion map. The activation times of nodes that are infected via ANC late in a contagion process are only mildly shorter than what occurs for spreading purely via WFP, so the points in the image of the contagion map are perturbed only slightly. To make (4.4) a meaningful bound for T^{ANC}, we thus need to work out the point in the spreading process up to which the occurrence of ANC plays a significant role in the overall spreading behaviour and accordingly has a noticeable effect on the contagion map. The later this point occurs, the larger $\frac{q(t)}{N-1}d^{\mathrm{NG}}$ will be and the larger we expect the critical threshold (bifurcation point) T^{ANC} to be. We have

$$\delta\, H^{\mathrm{ANC}} = \frac{\delta\, d^{\mathrm{NG}}}{d^{\mathrm{G}} + d^{\mathrm{NG}}} < T^{\mathrm{ANC}} < \frac{d^{\mathrm{NG}}}{d^{\mathrm{G}} + d^{\mathrm{NG}}} = H^{\mathrm{ANC}} \tag{4.5}$$

for some $\delta \in (0, 1)$, where the parameter δ is determined by how late in the spreading process the occurrence of ANC plays a significant role. If, for instance, the occurrence of ANC plays a significant role in overall spreading behaviour and thus has a noticeable distortive effect on a contagion map only if it takes place by the time that three fifths of the nodes of the network are active, then the bifurcation curve for ANC is bounded below as follows:

$$T^{\mathrm{ANC}} > \frac{\frac{1500}{2499}d^{\mathrm{NG}}}{d^{\mathrm{G}} + d^{\mathrm{NG}}}\,.$$

We compare the idealized bifurcation diagram (using the horizon of ANC as its bifurcation curve) and the diagram that we obtain from (4.4) with the number of active nodes equaling $(3/5)N$, where N is the total number of nodes, to our numerical results for the geometry and dimensionality (see Figure 4.22).

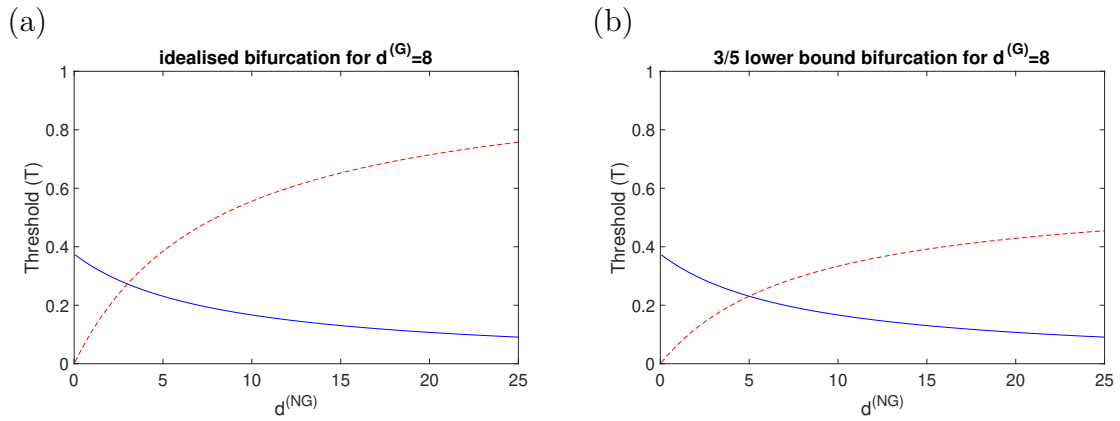


Figure 4.22: (a) Idealized bifurcation curve and (b) approximate bifurcation curve for WFP and ANC in a Kleinberg-like small-world network with geometric neighbours up to $r = \sqrt{2}$ around a node (i.e. with geometric degree $d^{\mathrm{G}} = 8$). The blue curve shows T^{WFP}; and the red curve shows the idealized and approximate T^{ANC} in (a) and (b), respectively.

The above argument is independent of the particular geometry that underlies a network, as long as the non-geometric edges are placed uniformly at random. In particular, the inequalities (4.5) should also hold for the ring lattice in the computations of Taylor *et al.* [71]. Indeed, looking at their results (see Figure 4.23), their T_0^{ANC} curve (the dotted curve) does seem to sit a bit high with respect to their numerical results, suggesting that T_0^{ANC} is indeed bounded above by the idealized bifurcation curve.

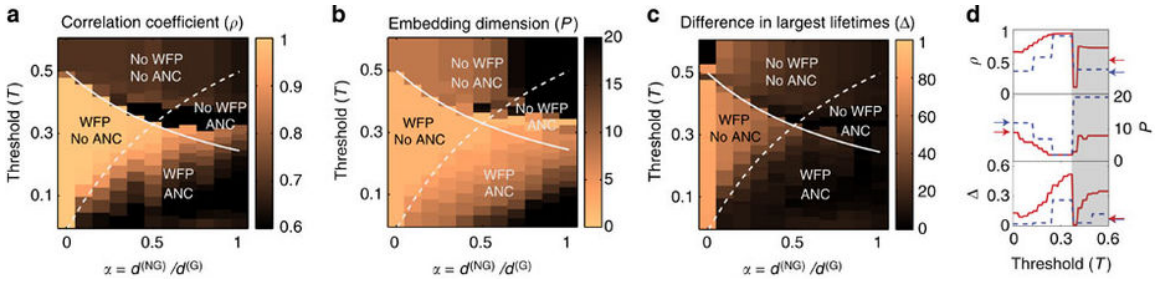


Figure 4.23: Results for the noisy ring lattices with $N = 200$ nodes and various ratios $\alpha = d^{NG}/d^{G}$ (with $d^{G} = 20$ and $d^{NG} = 0, \ldots, 20$). (Figure taken with permission from [71].)

If three fifths of the total number of nodes is indeed the correct choice for the maximal number of nodes that have been activated before a certain time step t for the occurrence of ANC to be significant at time t, then one should expect the actual

bifurcation curve T^{ANC} to lie somewhere between the red curves in Figure 4.22.

To find the actual bifurcation curve T^{ANC}, we need to find (for a given value of d^{NG}) the largest threshold T that realistically allows ANC to arise before the active region of the network is so large that the occurrence of ANC no longer has a significant impact on the spreading dynamics. This amounts to finding a threshold T that is as large as possible, but still small enough such that we obtain a value of at least 1 for the expected number of inactive nodes with sufficiently many non-geometric edges reaching into the active region.

To avoid taking into account the activation of nodes via ANC in the boundary of the active region or very close to it, one can count only inactive nodes that lie 'sufficiently far away' from active nodes. We use the term *neighbourhood*[8] and the notation $\mathcal{N}$ for the set of nodes outside which we count nodes being activated via ANC. The neighbourhood can consist either of the active nodes only, in which case $|\mathcal{N}| = q(t)$; or it can include some additional nodes around the active region, such that $\mathcal{N} > q(t)$.

At a given time t, suppose that the $q(t)$ nodes that are active at time t were activated at previous time steps by WFP from the cluster seed and that they form an active cluster of roughly square shape. We denote this set of nodes by

$$I = \{i \in V \mid \eta_i(t) = 1\},$$

so $|I| = q(t)$. We define the neighbourhood $\mathcal{N}(I)$ of this active cluster to be the cluster itself together with the nodes in its periphery of a certain 'width'. That is, $\mathcal{N}(I)$ is the active cluster I itself, its boundary, and — depending on the width — some more nodes around it. Given a width w, we approximate the number of nodes in $\mathcal{N}(I)$ as

$$|\mathcal{N}(I)| \approx \left(\sqrt{q(t)} + 2w\right)^2 . \tag{4.6}$$

[8]Our use of the term 'neighbourhood' is different from its usual use in graph theory.

We can choose any natural number for the width w, and one plausible choice is $d^{\mathrm{G}}/2$. If $w = 0$, the neighbourhood $\mathcal{N}(I)$ is just I itself (and then $|\mathcal{N}| = q(t)$). To make a sensible choice for the size of the active region of a network at the latest point in the spreading process at which we consider ANC to be significant, we estimate the largest number $q_{\max}$ of active nodes such that the active region together with a periphery of inactive nodes of width d^{G} takes up at most 90% of the network. We make this estimate by choosing $q_{\max}$ to be the largest integer such that

$$|\mathcal{N}_{\max}| = \left(\sqrt{q_{\max}} + 2d^{\mathrm{G}}\right)^2 \leq 0.9\,N\,. \tag{4.7}$$

Consequently,

$$q_{\max} \leq \left(\sqrt{0.9\,N} - 2d^{\mathrm{G}}\right)^2\,. \tag{4.8}$$

For $N = 2500$ and $d^{\mathrm{G}} = 8$, this gives $q_{\max} = 988$.

If we consider ANC to be significant up to time t, the corrected bifurcation curve for ANC is

$$T^{\mathrm{ANC}} = \frac{k}{d^{\mathrm{G}} + d^{\mathrm{NG}}}\,, \tag{4.9}$$

where k is the largest integer such that, at time t, the expected number of nodes outside the neighbourhood $\mathcal{N}(I)$ of I that have more than k active neighbours is at least 1.

Let X_k be the number of nodes outside $\mathcal{N}(I)$ with more than k active neighbours. It satisfies the binomial distribution

$$X_k \sim \mathrm{Bin}\left(N - |\mathcal{N}(I)|, P(d_{\mathrm{in}} > k)\right)\,.$$

so

$$P(X_k = x) = \binom{N - |\mathcal{N}(I)|}{x} P(d_{\mathrm{in}} > k)^x P(d_{\mathrm{in}} \leq k)^{N - |\mathcal{N}(I)| - x}\,,$$

where d_{in} is the number of edges of a node outside $\mathcal{N}(I)$ that are incident to an active node. Therefore, the expected number of nodes outside $\mathcal{N}(I)$ that have more than k

active neighbours is

$$\begin{aligned} E[X_k] &= \sum_{x=0}^{N-|\mathcal{N}(I)|} P(X_k = x)x \\ &= (N - |\mathcal{N}(I)|)\, P(d_{\text{in}} > k) \\ &= (N - |\mathcal{N}(I)|)\, (1 - P(d_{\text{in}} \leq k))\ . \end{aligned} \tag{4.10}$$

We have (approximately)

$$d_{\text{in}} \sim \text{Bin}\left(d^{\text{NG}}, \frac{q(t)}{N}\right)\ . \tag{4.11}$$

That is, d_{in} approximately follows a binomial distribution. Its associated (approximate) cumulative distribution function is

$$\begin{aligned} P(d_{\text{in}} \leq k) &\approx \sum_{d=0}^{k} \binom{d^{\text{NG}}}{d} \left(\frac{q(t)}{N}\right)^d \left(1 - \frac{q(t)}{N}\right)^{d^{\text{NG}}-d} \\ &= \left(d^{\text{NG}} - k\right) \binom{d^{\text{NG}}}{k} \int_0^{1-\frac{q(t)}{N}} s^{d^{NG}-k-1}(1-s)^k ds\,. \end{aligned} \tag{4.12}$$

Actually, the random variable d_{in} follows a binomial distribution only approximately, because — with our network model avoiding double edges — the event that a non-geometric edge of a node is incident to an active node is not entirely independent of another one of this node's non-geometric edges being incident to an active node. Consequently, if we want to determine the probability that a node outside $\mathcal{N}(I)$ has k non-geometric edges that reach into the contagion cluster (i.e. $P(d_{\text{in}}) = k$), we have to pick k of the node's d_{NG} non-geometric edges $\{e_1, e_2, \ldots e_k, e_{k+1}, \ldots, e_{d^{\text{NG}}-1}, e_{d^{\text{NG}}}\}$ and calculate consecutively, for each $\ell \in \{1, 2, \ldots, d^{\text{NG}}\}$, the probability that e_ℓ is incident to an active (if $\ell \in \{1, \ldots, k\}$) or inactive (if $\ell \in \{k+1, \ldots, d^{\text{NG}}\}$) node, given the states of the incident nodes of the e_j with $j < \ell$.

The precise probability of a node having k non-geometric edges that are incident to an active node is thus

$$P\left(d_{\text{in}} = k\right) = \binom{d^{\text{NG}}}{k} \frac{\left(\prod_{k'=0}^{k-1} q(t) - k'\right)\left(\prod_{k'=0}^{d^{\text{NG}}-1-k} N - 1 - q(t) - k'\right)}{\prod_{k'=0}^{d^{\text{NG}}-1}(N - 1 - k')}\,.$$

However, for $N \gg k$, it is reasonable to approximate the probability that a given non-geometric edge of an inactive node is incident to an active node as $\frac{q(t)}{N}$ (i.e. the number of active nodes divided by the total number of nodes). Consequently, d_{in} asymptotically follows the binomial distribution and equation (4.12) is correct asymptotically.

To determine the numerator of T^{ANC} (see formula (4.9)), we seek the largest integer k such that $E[X_k] \geq 1$. That is, using equation 4.12, we seek the largest integer k such that

$$\frac{N - |\mathcal{N}(I)| - 1}{N - |\mathcal{N}(I)|} \geq \sum_{d=0}^{k} \binom{d^{\text{NG}}}{d} \left(\frac{q(t)}{N}\right)^{d} \left(1 - \frac{q(t)}{N}\right)^{d^{\text{NG}}-d}. \tag{4.13}$$

We can find k for each value of d^{NG} and deduce $T^{\text{ANC}}(d^{\text{NG}})$ from that value for a given d^{G}. For $d^{\text{G}} = 8$, this yields the plots in Figure 4.24 for T^{ANC} for various choices of the width of the neighbourhood $\mathcal{N}(I)$ and of the maximum value of $q(t)$ at which we consider ANC to be significant.

Observe that the curves are essentially increasing, but in an oscillatory manner, resulting in a staircase-like shape. This shape arises from the fact that increasing d^{NG} from 0 to 25 in integer increments affects the largest integer k such that inequality (4.13) is satisfied (i.e. the numerator of formula (4.9)) as well as the denominator of formula (4.9). The larger the value of d^{NG}, the larger the value of $\binom{d^{\text{NG}}}{d}$ and the smaller the value of $\left(1 - \frac{q(t)}{N}\right)^{d^{\text{NG}}-d}$ in the sum on the right-hand side of inequality (4.13), with the latter factor being dominant. This explains the overall increasing tendency of the curve. For a given value of k, an increase of d^{NG} leads to a steady decrease of T^{ANC}, explaining the small intervals of decrease of the curve.

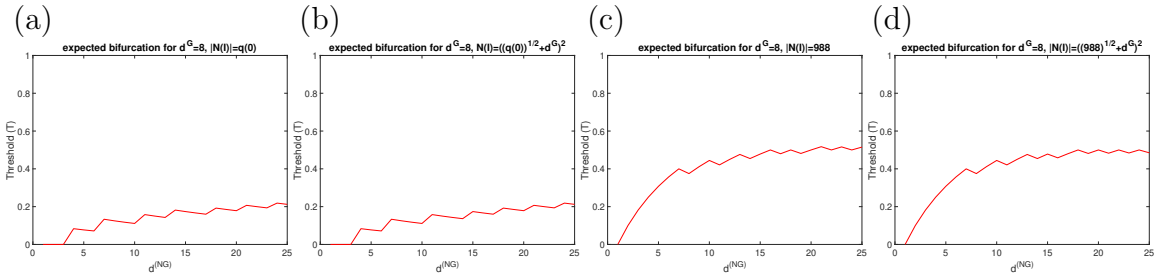


Figure 4.24: Behaviour of T^{ANC} in a Kleinberg-like small-world network geometric neighbours up to $r = \sqrt{2}$ around a node (i.e. with geometric degree $d^{\mathrm{G}} = 8$) for various choices of what constitutes the neighbourhood of an active cluster: (a) $|\mathcal{N}(I)| = q(0)$ (pathological case), (b) $|\mathcal{N}(I)| = \left(\sqrt{q(0)} + d^{\mathrm{G}}\right)^2$ (pathological case), (c) $|\mathcal{N}(I)| = 988$, and (d) $|\mathcal{N}(I)| = \left(\sqrt{988} + d^{\mathrm{G}}\right)^2$.

Note (again) that our central reasoning in the above argument is independent of the geometry that underlies our noisy geometric network, provided we place non-geometric edges uniformly at random. The only point at which the particular geometry of the two-dimensional torus comes into play is in our estimation of the approximation (4.6), which we calculate by assuming that the contagion cluster I is roughly square-shaped and that its neighborhood $\mathcal{N}(I)$ forms a larger square-shaped area of some width around I. If we take this width to be 0 (i.e., if $\mathcal{N}(I) = q(t)$), the formula for the expectation (4.10) is the same for any noisy geometric network.

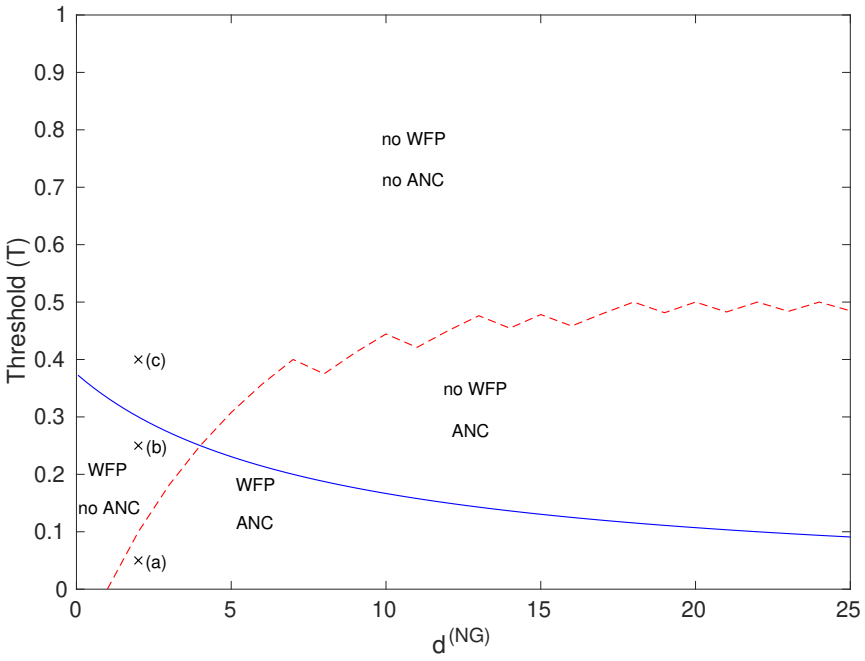


Figure 4.25: Bifurcation diagram for WFP and ANC on Kleinberg-like small-world network with geometric neighbours up to $r = \sqrt{2}$ around a node (i.e. geometric degree $d^{\mathrm{G}} = 8$). The blue curve shows T^{WFP}, and the red curve shows T^{ANC}. The points (a), (b), and (c) mark parameter combinations that represent three different spreading regimes. Point (a) indicates the parameter combination $(d^{\mathrm{NG}}, T) = (2, 0.05)$, for which (due to the low threshold T) we expect fast spreading via WFP and ANC; point (b) indicates the parameter combination $(d^{\mathrm{NG}}, T) = (2, 0.25)$, representing spreading predominantly via WFP; and point (c) indicates the parameter combination $(d^{\mathrm{NG}}, T) = (2, 0.4)$, for which we do not expect spreading. If $d^{\mathrm{G}} = 8$ and $d^{\mathrm{NG}} = 2$, the total degree is 10. This implies that a contagion requires 1, 3, and 5 adjacent activated nodes, respectively, to activate an inactive node for (a) $T = 0.05$, (b) $T = 0.25$, and (c) $T = 0.4$.

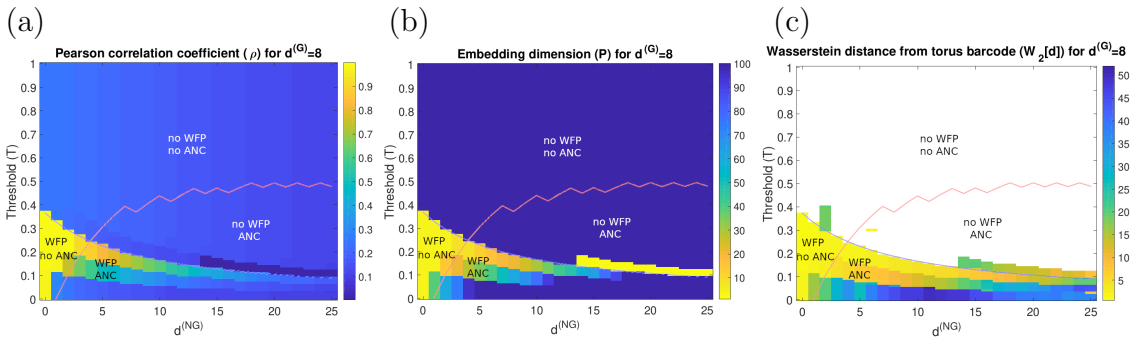


Figure 4.26: Our numerical results for (a) geometry, (b) topology, and (c) dimensionality superimposed on the bifurcation diagram for WFP and ANC for a Kleinberg-like small-world network with geometric neighbours up to $r = \sqrt{2}$ around a node (i.e. geometric degree $d^{\mathrm{G}} = 8$). The transitions between qualitatively different structures according to the numerical results align well with the curves for T^{WFP} and T^{ANC}. In the absence of WFP, the numerical results cannot distinguish between the presence and absence of ANC. Additionally, in the presence of WFP, the values indicating toroidal structure of the point cloud are somewhat pronounced even in the region in which no ANC is expected, and become weaker with progressively larger values of d^{NG}. This is a sensible observation, as toroidal structure can only be expected when WFP is present and is disturbed by the presence of ANC to varying degrees, depending on the rate of ANC.

4.3 Conclusions and future work

Networks that have some underlying geometry and include both geometric edges (which are short according to that geometry) and non-geometric edges (which can occur between nodes regardless of their distance) arise in many applications, including modelling of human communication and transportation. The spreading of a contagion on such a network can be influenced heavily by the underlying geometry, and it is useful to investigate the strength of such influence.

To study this problem, we considered a family of networks whose nodes can be considered to lie on a two-dimensional torus and whose edges include both geometric edges (which are deterministic and close to each other in the underlying geometry) and non-geometric edges (which are formed randomly and can be between nodes that are far apart in the underlying geometry). We investigated the spreading behaviour of threshold contagions on this family of networks. We did so by mapping network

nodes to a high-dimensional point cloud via a contagion map (following [71]) and analysing the structure of this point cloud from three perspectives: geometrically, topologically, and in terms of dimensionality. To examine the point cloud's geometry and dimensionality, we calculated the Pearson correlation coefficient and the embedding dimension, both of which are well established measures and easy to compute. To study the topology of the point cloud, we computed persistent homology of the Vietoris–Rips filtration and then calculated the Wasserstein distance between the corresponding barcode and a reference barcode. This was the most challenging and time-consuming part of our work, as algorithms for the computation of persistent homology are computationally expensive and software development in the field is still relatively young and evolving. Despite using Ripser, the currently best-performing documented software package for the computation of persistent homology of Vietoris–Rips filtrations [57], a single run on a contagion map took several hours to finish for some parameter combinations. We therefore restricted ourselves to computing persistent homology in dimension 1, although persistent homology in dimension 2 may also be insightful. In our analysis of the topological structure of the point clouds, we also illustrated the sensitivity of the Wasserstein distance to the overall scale of barcodes and the ensuing need to correct for geometric factors when using the Wasserstein distance as a measure of purely topological similarity. We proposed a way of calibrating barcodes so as to eliminate geometric factors and found that our method was effective at quantifying the 'torus-ness' of a point cloud.

Our main finding is that the computational analysis of the point clouds that are the images of contagion maps aligns with the bifurcation analysis of spreading behaviours in the network and contagion parameter space. We also found that the nature of the non-geometric edges affects contagion maps in the expected way: the shorter these edges are likely to be, the more they appear to contribute to the wave-

like spreading of a contagion along the underlying torus and the more torus-like the structure of the point cloud is. This provides empirical evidence to support the effectiveness of contagion maps as a tool for estimating spreading behaviour. It also suggests the potential of contagion maps as a manifold-learning tool that has some robustness to noise, a direction that we explore in Chapter 5. Complementarily, it should be valuable to analyse contagion maps mathematically to make precise statements about their performance under various conditions.

With the topological analysis of contagion maps having proved the most challenging part of our work, this aspect leaves room for further investigations in both computational experiments and mathematical theory. First, it may be insightful to examine the persistent homology in dimension 2 of the Vietoris–Rips filtration based on our focal network family, as there are nontrivial features to consider. Second, one might challenge the methodology for providing a quantitative measure for the topology of the contagion map itself: Both the Vietoris–Rips complex and the Wasserstein distance — despite being used as topological tools — are intrinsically geometric in nature. For the Wasserstein distance, we have addressed this issue by preprocessing barcodes by calibrating them. For the Vietoris–Rips filtration itself, we note that its construction is based purely on pairwise distances, just like the Pearson correlation coefficient that we used as a measure of geometry. While the persistent homology of the Vietoris–Rips filtration provides a richer, more nuanced summary of the set of pairwise distances than the Pearson correlation coefficient, this information is subsequently integrated when we calculate the Wasserstein distance from the corresponding barcode to a reference barcode. One may therefore question how much additional information one gains from our topological measure, especially in light of its computational complexity compared to that of the easily computed Pearson correlation coefficient. To understand this issue, it should be useful to explore how our

geometric and topological measures relate to each other and if such a relation can be exploited to speed up computations of our topological measure.

Finally, we note that it makes sense to use the approach of the present paper to study various other monotonic spreading processes, including those with stochastic update rules (such as compartmental models for diseases), on networks that have some underlying geometric structure (including ones that are more complicated than a torus).

Chapter 5

Contagion Maps for Manifold Learning

We test contagion maps and variants thereof as a manifold-learning tool on several different data sets, and we compare their performance to that of Isomap, one of the most well-known manifold-learning algorithms [72].

5.1 Algorithms

Both Isomap and contagion maps follow the same basic steps. First, if a data set is given in the form of a point cloud, we start by constructing a *neighbourhood graph* based on this point cloud whose nodes are the points and whose edges connect nearby pairs of nodes. We construct such a graph in two different ways: (1) by connecting each point to its k closest neighbours, producing the *k-nearest-neighbour graph*; and (2) by connecting any two points that are within ϵ from one another, producing the *ϵ-neighbourhood graph*. In both types of neighbourhood graph, one can either weight the edges by the corresponding pairwise distances, or treat the edges as unweighted[1]. Second, each algorithm calculates a notion of distance between points that is aimed to estimate the actual geodesic distance on the underlying manifold. Isomap takes the shortest-path length between two nodes in the neighbourhood graph as an estimate

[1]Note that, as the 'weights' in this graph are really distances, a small weight of an edge indicates a strong connection between its incident nodes.

for the actual geodesic distance between the two corresponding points; contagion maps take the activation times of nodes in a contagion process.

The idea is that only the pairwise distances between neighbouring pairs of points approximate the geodesic distances sufficiently, and that estimates for geodesic distances between non-neighbouring pairs of points can be inferred from the distances between neighbouring points by 'tracing' through the neighbourhood graph. The precise pairwise distances between adjacent points in a neighbourhood graph provide information that is relevant to estimating the geodesic distances between non-neighbouring points. Unweighted neighbourhood graphs forget this information and thus tend to lead to less accurate approximations to the geodesic distances. The loss of information from taking an unweighted graph can be greater for k-nearest-neighbour graphs than for ϵ-neighbourhood graphs, because, in the latter case, the pairwise distances are within a range that is capped by ϵ.

5.1.1 Isomap

The Isomap algorithm [72] is essentially a combination of a shortest-path algorithm with classical multidimensional scaling (MDS) [73].

MDS aims to embed a point cloud in a low-dimensional vector space in a way that preserves given distances between pairs of points as well as possible. It does so by minimizing a cost function called 'strain': Given a matrix $D = (d_{ij})_{i,j\in I}$ of pairwise distances, or 'dissimilarities' (not necessarily satisfying the defining properties of a metric), and some Euclidean target space Y, MDS finds coordinate vectors $\{y_i \in Y\}_{i\in I}$ that minimize the cost function

$$E = \|\tau(D) - \tau(D_Y)\|_{L^2} \ ,$$

where $D_Y = (d_2(y_i, y_j))_{i,j\in I}$, $\|M\|_{L^2} = \sqrt{\sum_{i,j} M_{ij}^2}$, and $\tau(M) = -HSH/2$ with $S_{ij} = M_{ij}^2$ and $H_{ij} = \delta_{ij} - 1/|I|$ [50].

MDS is a linear dimensionality-reduction technique. In a sense, Isomap works as a 'nonlinear version' of MDS which accommodates for potential curvature of data by incorporating a shortest-paths algorithm to estimate geodesic distances between data points.

The original Isomap algorithm proceeds as follows. First, given data in point-cloud form, Isomap starts by building a neighbourhood graph (V, E) on the point cloud. It does so by associating a set V of nodes to the data points (where i is the node associated to point p_i) and defining the edge set E to build either (1) a k-nearest-neighbour graph, or (2) an ϵ-neighbourhood graph:

(1) Given some $k \in \mathbb{N}$, $(i, j) \in E$ if and only if p_i is in the set of the k nearest neighbours of p_j, or vice versa.

(2) Given some $\epsilon \in \mathbb{R}_{>0}$, $(i, j) \in E$ if and only if $d_2(p_i, p_j) \leq \epsilon$.

If we are given data in the form of a network, we use this given (weighted or unweighted) network instead. Next, Isomap calculates the shortest-path lengths between pairs of nodes in this network using some shortest-path algorithm. We use the Floyd–Warshall algorithm [32,75] in the following work. The set of shortest-path lengths can be recorded in a *dissimilarity matrix* $D_{\text{iso}} = (d_{\text{G}}(i, j))_{i,j \in V}$ [2], to which Isomap finally applies MDS to map the data points to a low-dimensional space.

In addition to the original Isomap algorithm, which projects points via MDS based on the set of shortest-path lengths (i.e. the entries in D_{iso}), we also examine this set topologically (via a Vietoris–Rips filtration based on these shortest-path lengths) and geometrically (via a Pearson correlation with some given base-geometry), when possible. Furthermore, we analyse the point cloud given by the rows (or, equivalently, columns) of D_{iso}. Namely, we analyse this point cloud with the methods measuring

[2] The length of a shortest path between nodes i and j is denoted by $d_{\text{G}}(i, j)$. Note that the function $d_{\text{G}} : V \times V \to [0, \infty)$ is a metric on the set of nodes, or — equivalently — on the set of data points.

geometry (see Section 4.1.2), topology (see Section 4.1.3), and dimensionality (see Section 4.1.4) that we used on contagion maps (see Section 4.1), when possible. This allows us to compare Isomap to contagion maps under these three measures. When analysing the point cloud, we are essentially applying our methods to a dissimilarity matrix that encodes the pairwise distances between the column vectors of D_{iso}. We denote the operator that maps D_{iso} to that matrix as

$$p_{\text{dist}} \ : \ \mathcal{M}_{m,n}(\mathbb{R}) \longrightarrow \mathcal{M}_{n,n}(\mathbb{R}) \ ,$$

with $(p_{\text{dist}}(D))_{ij} = d_2(D_{*i}, D_{*j})$.

Note that our measure of geometry can only be calculated when there is a known base-geometry to compare our processed data to, a scenario which is not usually given in manifold-learning applications. The measure is, however, useful when testing the algorithm on benchmark data.

5.1.2 Contagion maps

Given point-cloud data, our contagion-map algorithm starts by creating a neighbourhood graph on the point cloud, as described for the Isomap algorithm in the previous section. If data is given in the form of a network, our contagion-map algorithm works with this given network instead. We denote the graph by (V, E), the number of nodes (i.e. the number of points in the case of point-cloud data) by $|V| = N$, and the graph's binary adjacency matrix by A. For a given seed set, the threshold contagion (defined in Section 3.2.2) on a given network yields a deterministic process, and one can record the activation times of the nodes. Recall from Section 4.1.1 that contagion maps consider realizations of the contagion dynamics initialized with seeds consisting of a node j together with its direct neighbours; that is the jth seed set is $S^{(j)} = \{j\} \cup \{k \mid A_{jk} \neq 0\}$, for each $j \in V$. This produces N realizations of the contagion. We denote the activation time of node $i \in V$ during the realization that

is seeded at the cluster around node $j \in V$ by $x_j^{(i)}$. If node i is never activated in the realization that is seeded around node $j \in V$, we set $x_j^{(i)} = 2N$ (i.e. larger than any actual activation time).

One can now work directly with this set of activation times, that is, treat the activation times as estimates for the geodesic distances between points on the underlying manifold and apply our methods to examine geometry, topology, and dimensionality to them. Alternatively, one can work with points whose coordinate vectors are given by the rows of the dissimilarity matrix $D_{\text{cont}} = \left(x_j^{(i)}\right)_{i,j \in V}$ that holds the activation times (or of a symmetrization of this dissimilarity matrix given by $D_{\text{cont}} + (D_{\text{cont}})^{\mathrm{T}}$). Using terminology from Chapter 4, producing such a point cloud means mapping the nodes via a *contagion map*.

Recall that the *regular contagion map* is the function from the set V of nodes to $\mathbb{R}^N$ that is defined by

$$i \mapsto x^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_N^{(i)}] .$$

Similarly, the *symmetric contagion map* is the function from the set V of nodes to $\mathbb{R}^N$ that is defined by

$$i \mapsto [x_1^{(i)} + x_i^{(1)}, \ldots, x_N^{(i)} + x_i^{(N)}] .$$

We work with symmetric contagion maps exclusively.

Note that a contagion map with threshold $T = 0$ is approximately equivalent to a version of Isomap that uses an unweighted neighbourhood graph.

5.1.3 Workflow

Our workflow is composed of multiple stages, at each of which one can choose from a number of different options. This leads to exponentially many possible procedures that one can use to analyse a given data set. First, given point-cloud data, one needs to choose the type of neighbourhood graph to build on this data set, as well as the

defining parameter k or ϵ. Next, one needs to pick a way of estimating geodesic distances based on this graph. We choose either shortest-path distance or activation times in a threshold contagion; that is, we follow either the Isomap algorithm or that of contagion maps. In the case of contagion maps, one also needs to choose a threshold parameter T. Given the set of estimates for the geodesic distances, i.e. the dissimilarity matrix that encodes the shortest-path (D_{iso}) distances or activation times (D_{cont}), one can apply further methods either to these estimates directly, or to the pairwise distances between the points whose coordinate vectors are the columns (or rows, by symmetry) of this dissimilarity matrix. For either of these choices one can finally apply methods to measure dimensionality, topology, and geometry. Figure 5.1 shows a schematic representation of our workflow in tree form. We apply different subsets of this full analysis to the different data sets that we consider.

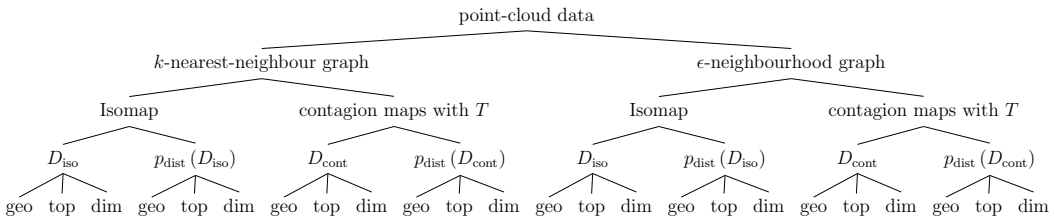


Figure 5.1: Schematic representation of the workflow.

5.2 Data sets

We apply Isomap, as well as contagion maps with several different thresholds, to four different data sets. First, we consider point **samples from a Swiss roll** that are obtained by taking regularly spaced points on the Swiss roll surface and then adding various levels of noise to the points. This way of generating data allows us to have direct control over the data, so we can explore how the different algorithms react to slight variations of the data (in terms of e.g. density, noise level, or uniformity). We use this data set primarily to explore the effects of the parameter k or ϵ when

building a neighbourhood graph. We also examine the **Swiss roll data set from Tenenbaum *et al.*** [72], a classical benchmark data set for dimensionality reduction and one of the first data sets to which Isomap was applied. We then analyse a couple of representatives of the **torus-based networks** on which our contagion-map-based analysis in the previous chapter was focused. This serves as a complement to the previous chapters, which revolved around the identification of spreading behavior on known networks (as opposed to manifold identification using spreading dynamics, which is what we aim to do in this chapter). Finally, we consider a data set that represents the **conformation space of the cyclo-octane molecule** [51]. This data set is known to have non-trivial topological features and is an example of a naturally occurring data set.

5.3 Results

5.3.1 Noisy regularly spaced points from Swiss roll

We take points on a Swiss roll at a density of approximately 50 per unit area. The points are regularly spaced with respect to the intrinsic geodesic distance on the Swiss roll (see Figure 5.2(a)).

We parameterize our Swiss roll as follows:

$$\text{BSR} = \left\{ \phi(s,t) = \left(\frac{t}{8\pi}\cos(t), \frac{t}{8\pi}\sin(t), s \right) : t \in \mathbb{R}_{>0}, s \in \mathbb{R} \right\} .$$

The factor of $\frac{1}{8\pi}$ ensures that, as t increases, the radius increases by $\frac{1}{4}$ every time the angle increases by 2π (i.e. every time a circle is completed). So adjacent sheets of this Swiss roll are at a distance of $\frac{1}{4}$ from each other.

We take regularly spaced points on BSR by varying s in equal increments of d to get a sequence $\{s_1, s_2, \dots s_l\}$ of parameter values for s with $s_{i+1} - s_i = d$, and calculating a sequence $\{t_1, t_2, \dots t_k\}$ of values for t such that the points $\phi(s_i, t_j)$ and $\phi(s_i, t_{j+1})$ are apart by a geodesic distance of d on the Swiss roll surface. To find the

sequence $\{t_1, t_2, \ldots t_k\}$ we calculate the length $|P_{s_{\text{fix}}}|$ of a path $P_{s_{\text{fix}}}$ on BSR obtained by fixing a value s_{fix} for s and varying t:

$$P_{s_{\text{fix}}}(t_{\text{start}}, t_{\text{end}}) = \{\phi\left(s_{\text{fix}}, t\right) \, : t_{\text{start}} \leq t \leq t_{\text{end}}\} \ .$$

$$\begin{aligned}
|P_{s_{\text{fix}}}(t_{\text{start}}, t_{\text{end}})| &= \int_{t_{\text{start}}}^{t_{\text{end}}} \left| \frac{d\phi\left(s_{\text{fix}}, t\right)}{dt} \right| \mathrm{d}t \\
&= \int_{t_{\text{start}}}^{t_{\text{end}}} \left| \left(\frac{\cos(t)}{8\pi} - \frac{t}{8\pi}\sin(t), \frac{\sin(t)}{8\pi} + \frac{t}{8\pi}\cos(t), 0 \right) \right| \mathrm{d}t \\
&= \frac{1}{8\pi} \int_{t_{\text{start}}}^{t_{\text{end}}} \sqrt{\left(\cos(t) - t\sin(t)\right)^2 + \left(\sin(t) + t\cos(t)\right)^2} \mathrm{d}t \\
&= \frac{1}{8\pi} \int_{t_{\text{start}}}^{t_{\text{end}}} \sqrt{1 + t^2} \mathrm{d}t \\
&= \frac{1}{8\pi} \int_{\theta(t_{\text{start}})}^{\theta(t_{\text{end}})} \sqrt{1 + \sinh^2(\theta)} \cosh(\theta) d\theta \qquad (t = \sinh(\theta)) \\
&= \frac{1}{8\pi} \int_{\theta(t_{\text{start}})}^{\theta(t_{\text{end}})} \cosh^2(\theta) d\theta \\
&= \frac{1}{8\pi} \int_{\theta(t_{\text{start}})}^{\theta(t_{\text{end}})} \frac{\cosh(2\theta) + 1}{2} d\theta \\
&= \frac{1}{16\pi} \left[\frac{1}{2}\sinh(2\theta) + \theta \right]_{\theta(t_{\text{start}})}^{\theta(t_{\text{end}})} \\
&= \frac{1}{16\pi} \left[t\sqrt{1 + t^2} + \sinh^{-1}(t) \right]_{t_{\text{start}}}^{t_{\text{end}}} \\
&= \frac{1}{16\pi} \left[t\sqrt{1 + t^2} + \ln(t + \sqrt{1 + t^2}) \right]_{t_{\text{start}}}^{t_{\text{end}}} \ .
\end{aligned}$$

We start by setting $t_0 = 0$ and then find t_1 such that

$$|P_{s_{\text{fix}}}(t_0, t_1)| = \frac{1}{16\pi} \left[t\sqrt{1 + t^2} + \ln(t + \sqrt{1 + t^2}) \right]_{t_0}^{t_1} = d \ .$$

Likewise, given t_j, we find t_{j+1} such that

$$|P_{s_{\text{fix}}}(t_j, t_{j+1})| = \frac{1}{16\pi} \left[t\sqrt{1 + t^2} + \ln(t + \sqrt{1 + t^2}) \right]_{t_j}^{t_{j+1}} = d \ .$$

We then add *Gaussian noise* with a specified signal-to-noise ratio (S/N) to these regularly spaced points (see Figure 5.2(b)). That is, for each point $p = [p_1, p_2, p_3] \in$

BSR, we add independent, identically distributed noise drawn from a zero-mean normal distribution to each of its coordinates to obtain a perturbation p_{noisy} of the point:

$$p_{\text{noisy}} = [p_1 + n_1, p_2 + n_2, p_3 + n_3] \ ,$$

where $n_i \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 10^{\frac{-S/N}{10}}$.

We test Isomap and contagion maps on this noisy point cloud to see how well each of them identifies the underlying 2-dimensional space. All algorithms start by building a neighbourhood graph on the points (see Figure 5.2(c,d)).

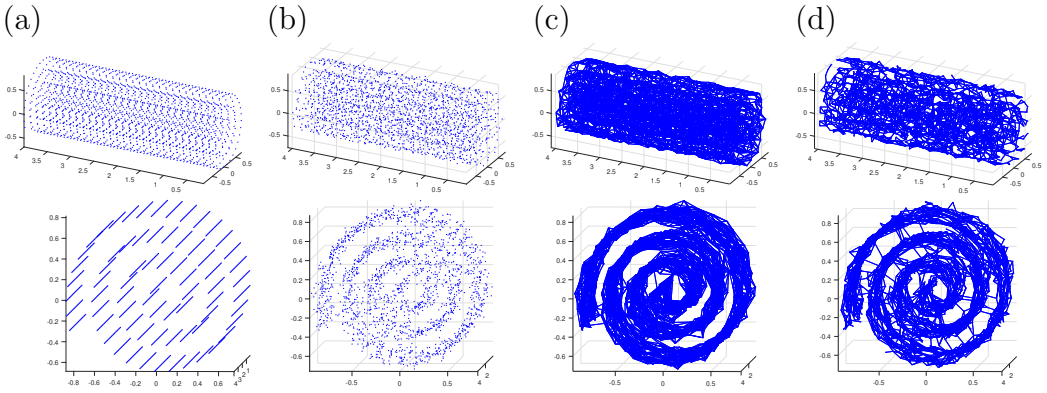


Figure 5.2: (a) Regularly spaced points on a Swiss roll at density 50 per unit area. (b) Gaussian noise is added at a signal-to-noise ratio of 30. (c) The 5-nearest-neighbour graph on the noisy point cloud. (d) The 0.18-neighbourhood graph on the noisy point cloud. (We show two views of each plot.)

We have already touched on Isomap's sensitivity to 'short-circuit errors' in Section 1.2 of the introduction. However, a careful choice of ϵ (or k) when constructing the neighbourhood graph can mitigate such errors to some extent. The goal is to find an ϵ (or k) that is small enough to avoid short-circuit edges, but not so small that the resulting graph 'corrupts' the underlying space. One may choose to simply vary ϵ over a range and implement Isomap on all of the neighbourhood graphs that thus arise. This approach is in the same vein as considering the full range of thresholds for the contagion map algorithm, and it works whenever there exists a range of ϵ (or k) for which the resulting neighbourhood graphs correctly represent the underlying

topology. However, for some data sets — particularly those that are sparse and incorporate a high level of noise — it is impossible to find a value for ϵ (or k) that strikes a balance between covering the underlying topology and not making 'short-circuit errors'. In other words, the range of ϵ (or k) that 'corrupt' the underlying topology and the range of ϵ (or k) that make 'short-circuit errors' overlap, leaving no values of ϵ (or k) that yield neighbourhood graphs that correctly represent the underlying topology. For such data sets, Isomap is inadequate as a manifold-learning tool, but contagion maps may be effective.

See Figures 5.3 and 5.4 for examples on the Swiss roll that illustrate the concepts in this paragraph. In particular, Figure 5.3 shows an example of a data set for which a careful choice of ϵ generates a neighbourhood graph that both captures the underlying manifold and does not include 'short-circuit' edges. By contrast, Figure 5.4 shows an example of a data set, for which no choice of ϵ produces a neighbourhood graph that accurately represents the underlying manifold. Figure 5.5 shows the residual variances of projecting this data set via MDS to dimensions 1 to 10 when based on Isompap, and when based on the contagion map with $T = 0.2$ (both starting with a 0.18-neighbourhood graph). For the contagion map, the residual variance plunges at dimension 2, thereby correctly identifying the intrinsic dimension of the data. The residual variances for Isomap, on the other hand, only decrease slighty and continuosly across the increasing target dimensions. That is, Isomap fails to see the correct intrinsic structure of the data when starting with an ϵ-neighbourhood graph with $\epsilon = 0.18$ (or any other value of ϵ), while contagion map — with a suitable choice of ϵ and T — correctly identifies the underlying structure. For contagion maps to work in this case, the value of ϵ had to be chosen large enough for the neighbourhood graph to cover the underlying manifold; and the value of T had to be picked small enough to carry the contagion and large enough to be resistant to the unavoidable

noisy inter-sheet edges in the ϵ-neighbourhood graph.

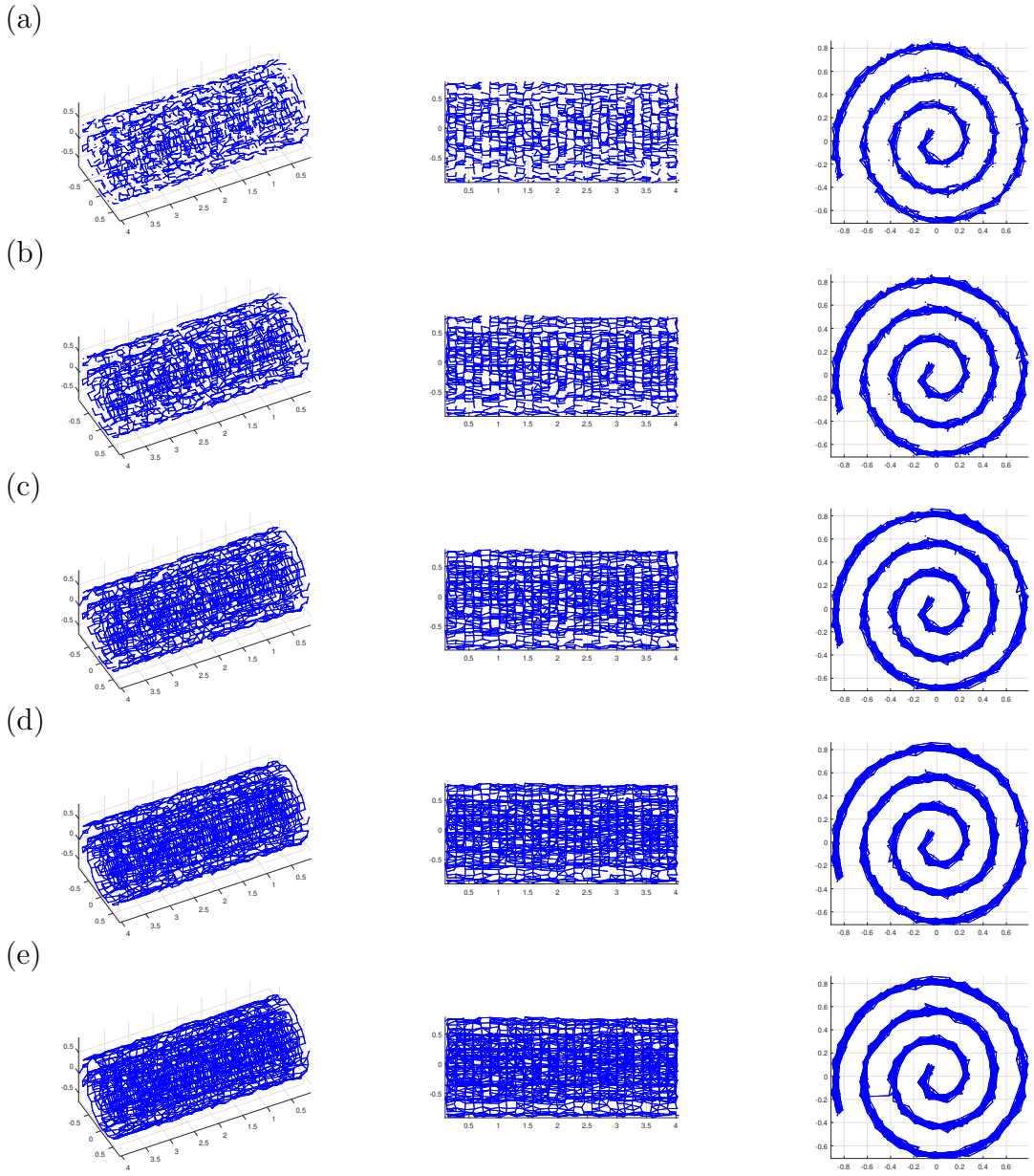


Figure 5.3: The ϵ-neighbourhood graphs on a noisy point sample from a Swiss roll with $S/N = 35$ for (a) $\epsilon = 0.14$, (b) $\epsilon = 0.15$, (c) $\epsilon = 0.16$, (d) $\epsilon = 0.17$, and (e) $\epsilon = 0.18$. (We show three views of each plot.) If ϵ is too small (e.g. $\epsilon = 0.14$), the ϵ-neighbourhood graph does not adequately 'cover' the underlying Swiss roll. If ϵ is too large (e.g. $\epsilon = 0.18$), the ϵ-neighbourhood graph includes inter-sheet edges, and thus does not represent the underlying Swiss roll. However, there exists a range of ϵ for which the ϵ-neighbourhood graph covers the underlying surface and does not include inter-sheet edges, thus providing an authentic representation of the underlying Swiss roll. In other words, the pairwise distances between points whose corresponding nodes are adjacent in the ϵ-neighbourhood graph for such ϵ approximate the actual geodesic sufficiently, and approximate pairwise geodesic distances between other point pairs can be inferred through the Isomap algorithm. This is an example of a data set for which Isomap is suitable as a manifold-learning technique with a careful choice of ϵ.

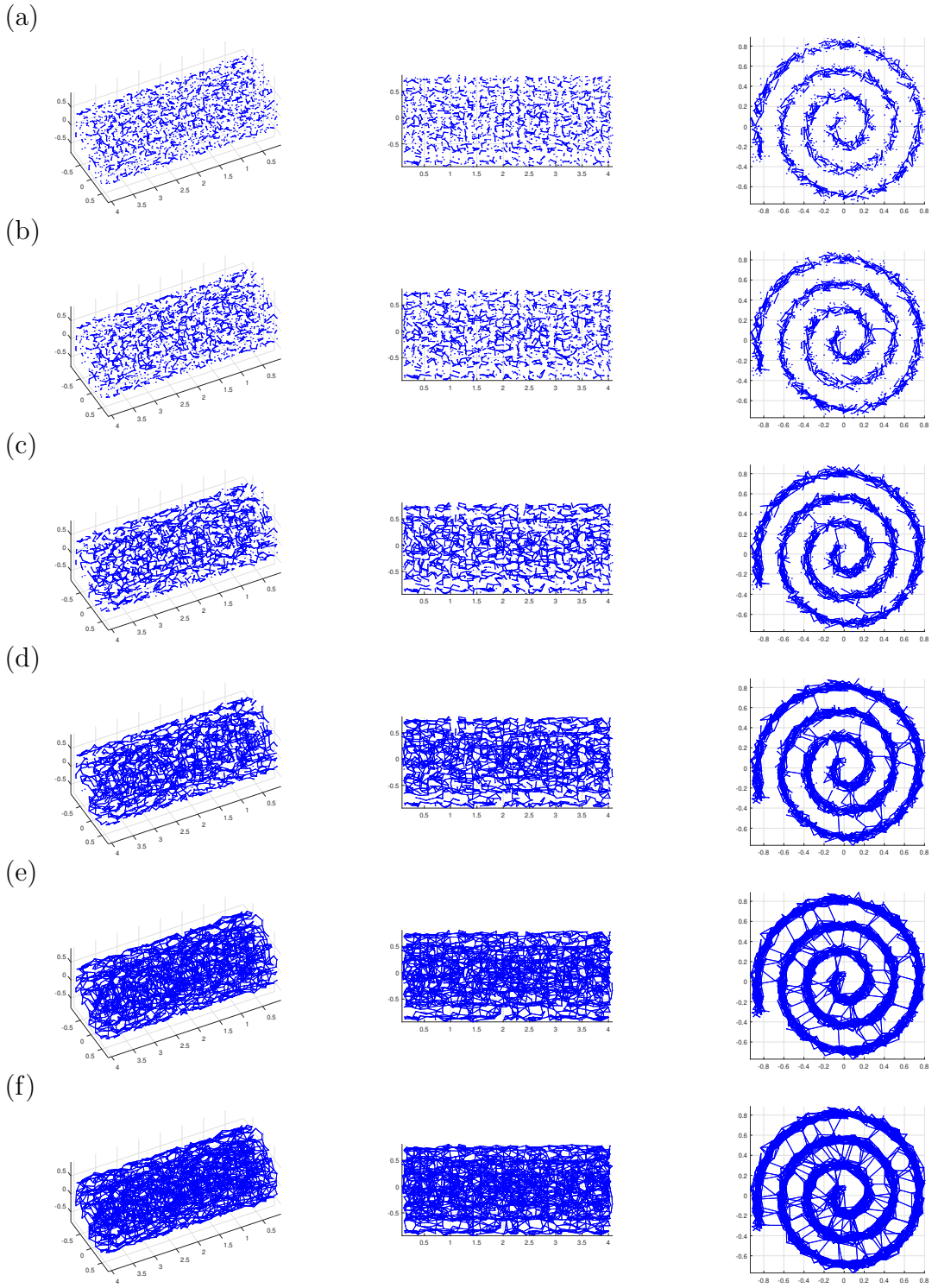


Figure 5.4: The ϵ-neighbourhood graphs on a noisy point sample from a Swiss roll with $S/N = 30$ for (a) $\epsilon = 0.11$, (b) $\epsilon = 0.12$, (c) $\epsilon = 0.14$, (d) $\epsilon = 0.16$, (e) $\epsilon = 0.18$, and (f) $\epsilon = 0.19$. (We show three views of each plot.) For small ϵ, the ϵ-neighbourhood graph does not adequately 'cover' the underlying Swiss roll. As ϵ increases, noisy inter-sheet edges appear (for e.g. $\epsilon = 0.12$) before ϵ is large enough for the neighbourhood graph to adequately 'cover' the underlying Swiss roll. This is an example of a data set for which Isomap cannot be used successfully as a manifold-learning technique with any choice of ϵ.

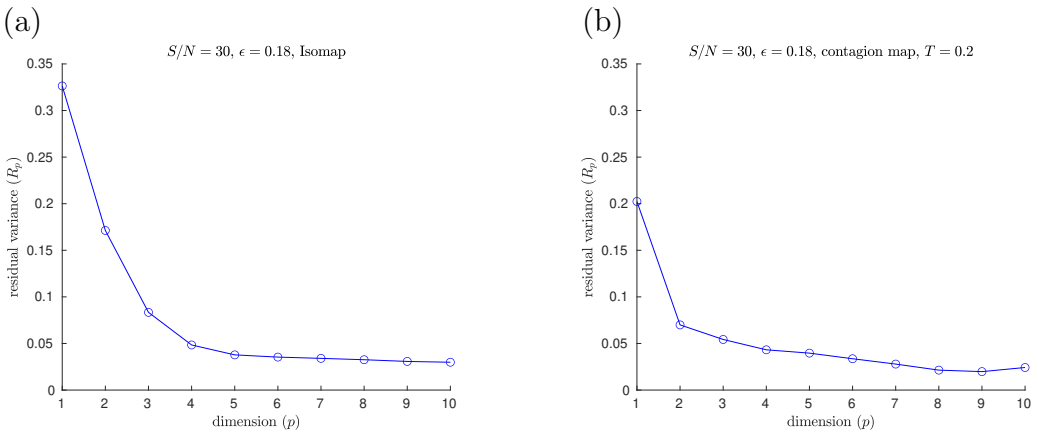


Figure 5.5: (a) Isomap and (b) contagion map with $T = 0.2$ (both starting with a 0.18-neighbourhood graph) on BSR with $S/N = 30$.

5.3.2 Swiss roll data set from Tenenbaum *et al.*

We consider 2000 points of the Swiss roll data set from [72]. This data set consists of 20000 points in total.

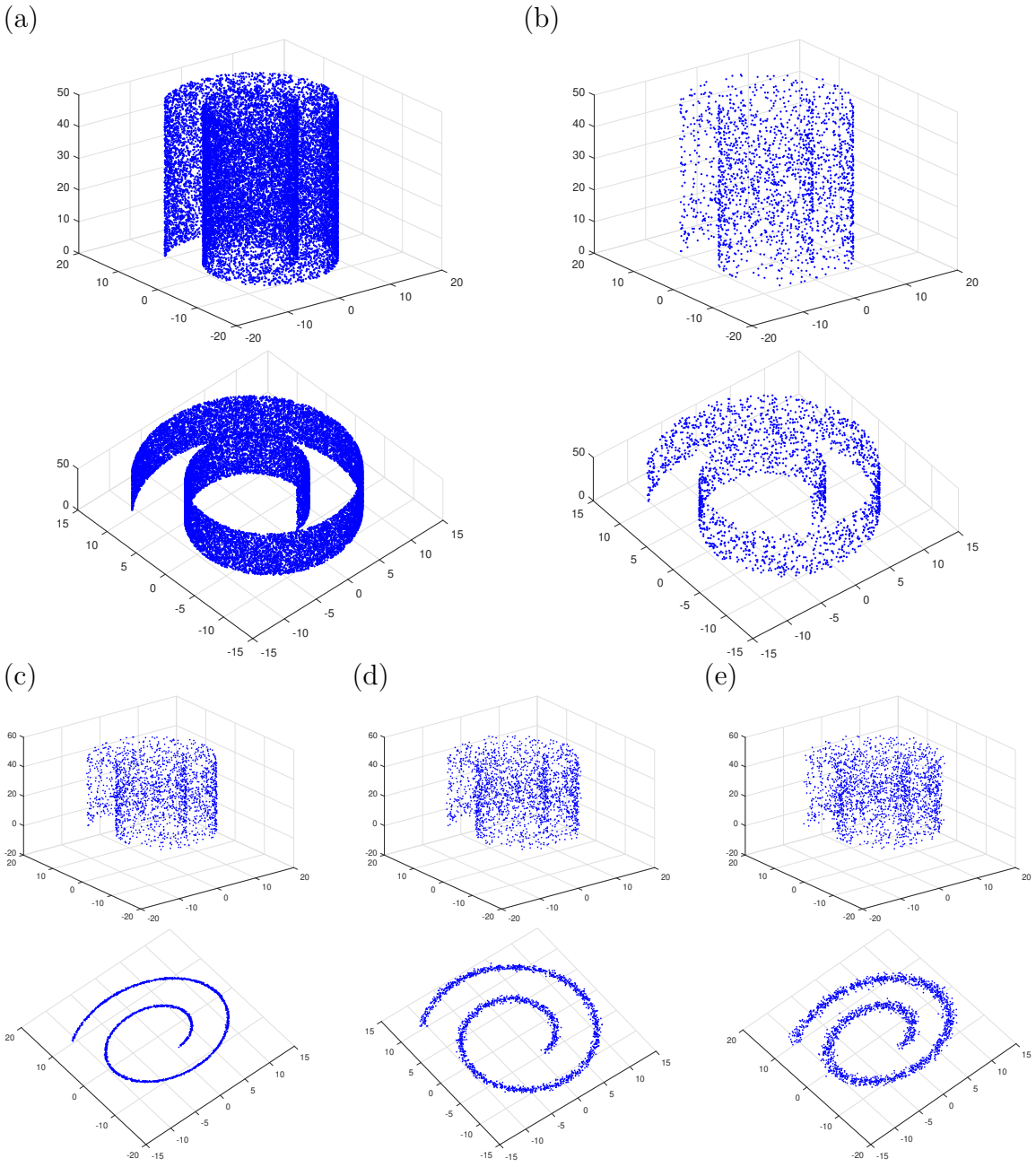


Figure 5.6: (a) All 20000 points, and (b) the first 2000 points in the Swiss roll data set from [72]. We add white Gaussian noise to the 2000 points with several different signal-to-noise ratios (S/N): (c) $S/N = 20$, (d) $S/N = 10$, (e) $S/N = 5$. (We show two views of each plot.)

Both Isomap and contagion maps perform adequately for the high signal-to-noise ration of $S/N = 20$ (i.e. low noise level) with all four examined versions of neighbourhood graphs (see Figure 5.7).

For the lowest signal-to-noise ratio (i.e. greatest noise level) that we consider — namely $S/N = 5$ — the 8-nearest-neighbour graph includes noisy inter-sheet edges. As a result, Isomap does not detect the intrinsic dimension 2 of the Swiss roll when using the 8-nearest neighbour graph, and neither does the contagion map with $T = 0$ or $T = 0.1$, thresholds for which the activation times of our threshold contagion are close to the shortest paths in the unweighted neighbourhood graph (see Figure 5.9). With $T = 0.2$, however, the contagion map does correctly recover the intrinsic dimension 2, as this threshold is just large enough to be robust to the occurring noisy edges. For thresholds larger than $T = 0.2$, many of the realizations of our threshold contagion leave nodes in the neighbourhood graph inactive (recorded as 'infinite' activation times), and, as a result, the residual variances of the embeddings based on these activation times are large for all considered dimensions (1 to 10). This illustrates that, while contagion maps can be a powerful tool when dealing with such noisy edges, a suitable choice of threshold can be a delicate matter.

Similarly, for signal-to-noise ratio $S/N = 5$, Isomap fails when based on the 4.5-neighbourhood graph or the 5-neighbourhood graph, but the contagion map for a threshold of $T = 0.2$ successfully identifies the intrinsic dimension 2 for both examined values of the neighbourhood parameter ϵ. Furthermore, when based on a 5-neighbourhood graph, Isomap fails even for the higher $S/N = 10$, as the 5-neighbourhood graph includes noisy inter-sheet edges even for this lower noise level (see Figure 5.8).

Note that our measures for topology and geometry are not useful for this data set, as the underlying manifold has no non-trivial topological features, and there is

no base-geometry provided.

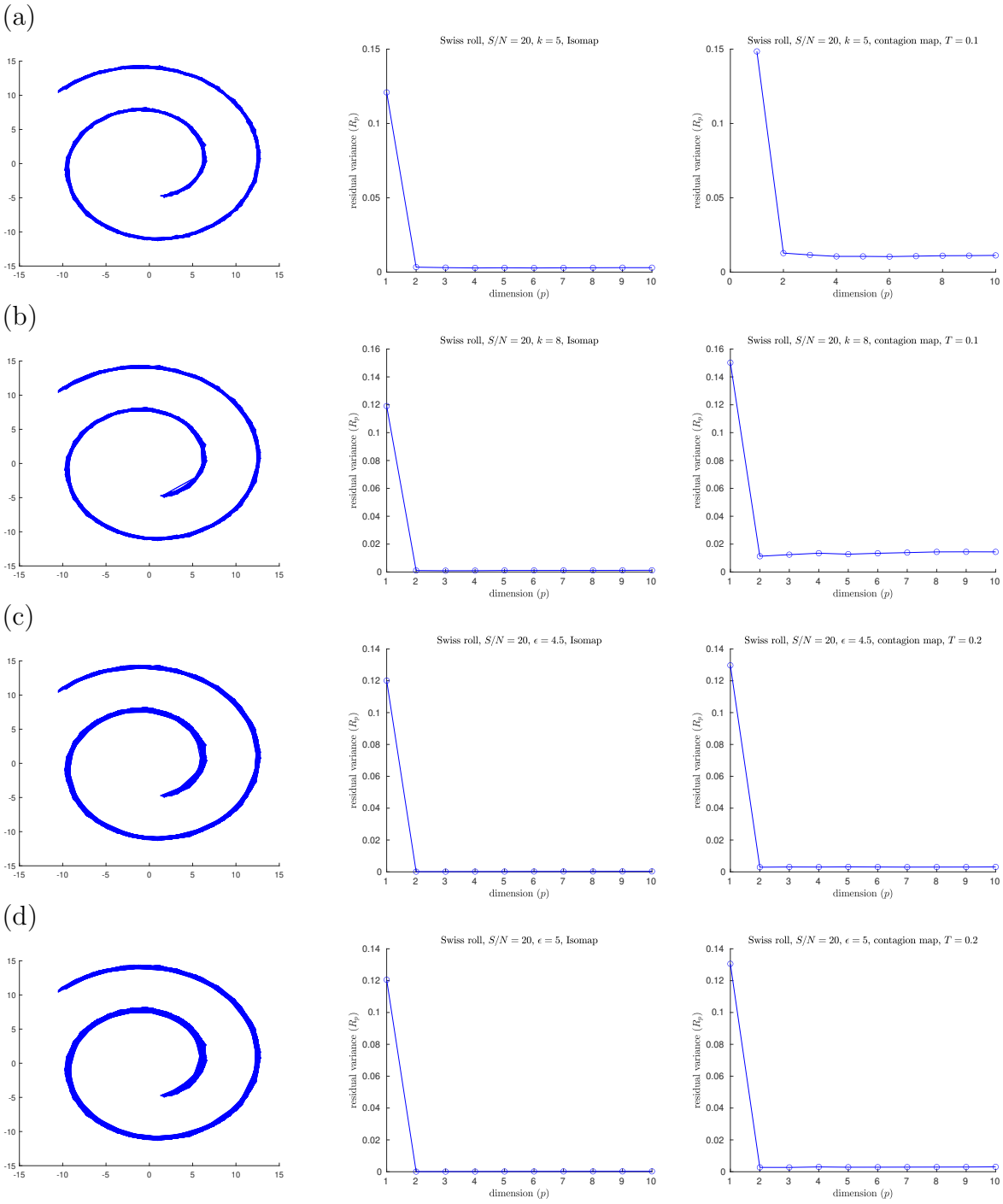


Figure 5.7: Neighbourhood graphs on 2000 points of the Swiss roll data set from [72] with white Gaussian noise added with $S/N = 20$ (left column). The residual variances of MDS projections to dimensions 1 to 10 resulting from (middle column) the approximate pairwise geodesic distances between nodes based on shortest paths on the weighted graph (i.e. Isomap) and (right column) the activation times in a threshold contagion on the unweighted graph (i.e. contagion maps). The neighbourhood graphs are (a) the 5-nearest-neighbour graph, (b) the 8-nearest-neighbour graph, (c) the 4.5-neighbourhood graph, and (d) the 5-neighbourhood graph.

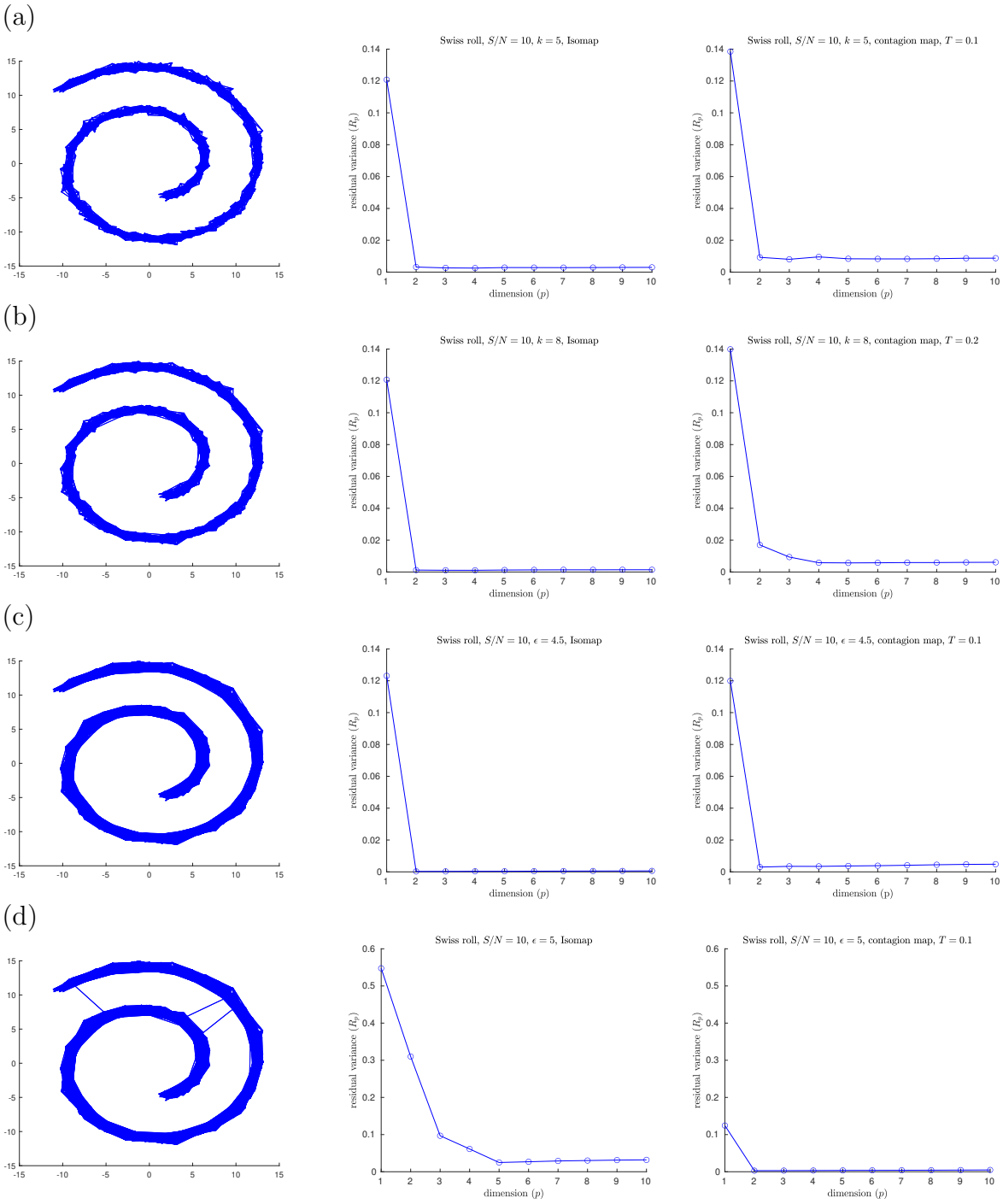


Figure 5.8: Neighbourhood graphs on 2000 points of the Swiss roll data set from [72] with white Gaussian noise added with $S/N = 10$ (left column). The residual variances of MDS projections to dimensions 1 to 10 resulting from (middle column) the approximate pairwise geodesic distances between nodes based on shortest paths on the weighted graph (i.e. Isomap) and (right column) the activation times in a threshold contagion on the unweighted graph (i.e. contagion maps). The neighbourhood graphs are (a) the 5-nearest neighbour graph, (b) the 8-nearest neighbour graph, (c) the 4.5-neighbourhood graph, and (d) the 5-neighbourhood graph.

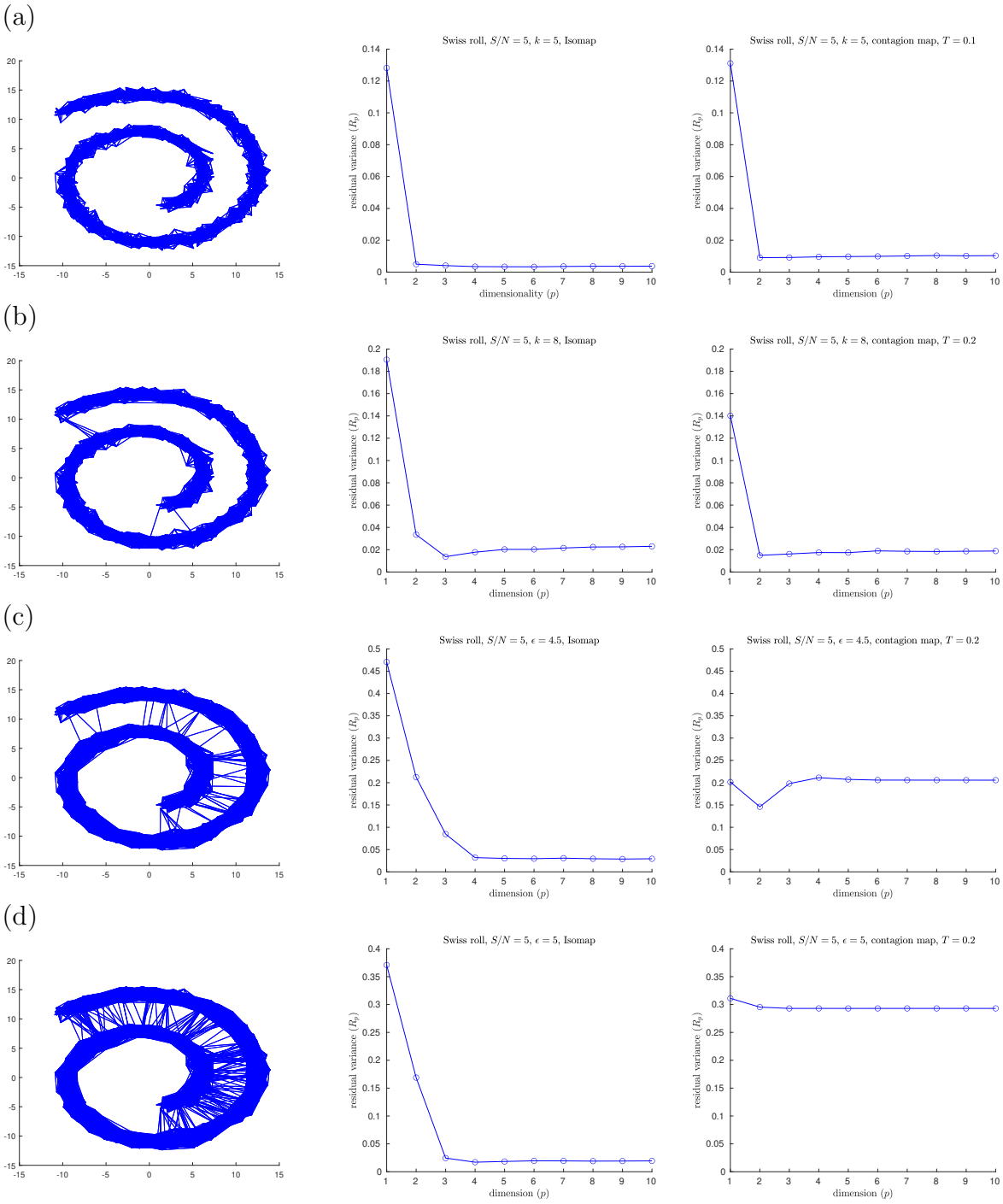


Figure 5.9: Neighbourhood graphs on 2000 points of the Swiss roll data set from [72] with white Gaussian noise added with $S/N = 5$ (left column). The residual variances of MDS projections to dimensions 1 to 10 resulting from (middle column) the approximate pairwise geodesic distances between nodes based on shortest path on the weighted graph (i.e. Isomap) and (right column) the activation times in a threshold contagion on the unweighted graph (i.e. contagion maps). The neighbourhood graphs are (a) the 5-nearest neighbour graph, (b) the 8-nearest neighbour graph, (c) the 4.5-neighbourhood graph, and (d) the 5-neighbourhood graph.

5.3.3 Torus-based networks

We consider the torus-based model described in Section 3.2.1 with $N = 2500$ nodes and a geometric degree of $d^{\mathrm{G}} = 8$ (see Figure 5.10). We add first 2 and then 4 non-geometric edges per node uniformly at random (i.e. $d^{\mathrm{NG}} = 2$ or 4, and $\gamma = 0$) and apply versions of both Isomap and contagion maps (with thresholds $T = 0, 0.1, \ldots, 1$) to the resulting networks. Namely, we calculate the shortest-path lengths between pairs of nodes in these two unweighted networks, as well as the activation times in all realizations of the threshold contagion defined in Section 3.2.2 that are seeded at the direct neighbourhoods of individual nodes (the red nodes in Figure 5.10). We thus obtain two dissimilarity matrices, one holding the shortest-path lengths (D_{iso}) and one holding the symmetrized activation times (D_{cont}). We analyse the information held in these two dissimilarity matrices geometrically, topologically, and in terms of dimensionality in two ways each. We first analyse the estimated pairwise geodesic distances held in the dissimilarity matrices directly, and we then consider the point cloud that results from taking columns (or, equivalently, rows) of each dissimilarity matrices as the coordinate vectors of points in $\mathbb{R}^{2500}$.

In detail, we perform the following analyses:

- In terms of dimensionality: We perform MDS based on the entries in each dissimilarity matrix to dimensions 1 to 10 and record the residual variance for each dimension. We also perform MDS on the point cloud that results from taking columns (or, equivalently, rows) of each dissimilarity matrix as the coordinate vectors of points in $\mathbb{R}^{2500}$. In both cases, we identify the approximate embedding dimension as the lowest dimension such that the residual variance when projecting down to that dimension via MDS is below 5%.

- Topologically: We build Vietoris–Rips filtrations based on the approximations to the pairwise geodesic distances (i.e. the entries in each dissimilarity matrix)

and compute their persistent homologies. We also build Vietoris–Rips filtrations on the points cloud that results from taking columns (or, equivalently, rows) of each dissimilarity matrix as the coordinate vectors of points in $\mathbb{R}^{2500}$ and compute their persistent homologies.

- Geometrically: We calculate the Pearson correlation coefficient between the entries in each dissimilarity matrix and the corresponding pairwise distances between regularly spaced points on a torus (see Formula 4.2). We also calculate the Pearson correlation coefficient between the pairwise distances between points in the point cloud that results from taking columns (or, equivalently, rows) of each dissimilarity matrix as the coordinate vectors of points in $\mathbb{R}^{2500}$ and the corresponding pairwise distances between regularly spaced points on a torus (see Formula 4.2).

We find that Isomap is unable to infer the underlying torus structure from these networks. Contagion map, however, detects the characteristics of the torus when using a threshold of $T \approx 0.2$. The results in this section illustrate the utility of contagion maps for spatial network data that incorporates noisy edges and its potential to outperform Isomap in such scenarios, but they also highlight that a careful choice of T is critical.

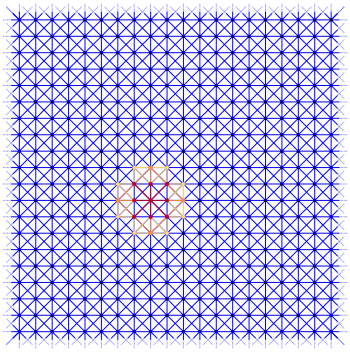


Figure 5.10: Illustration of the purely geometric base network of the networks that we consider in this section. This network is similar to Kleinberg's small-world like network [46]; see Section 3.2.1 for a formal definition. Each node has 8 geometric neighbours; nodes on opposite sides of the illustration are adjacent. The set of dark red nodes consists of one node and its immediate neighbours. We consider realizations of the threshold contagion defined in Section 3.2.2 that are seeded at such a set of nodes. The set of medium colour (respectively, light colour) nodes consists of the nodes that are activated in the first (respectively, second) time step when $T < \frac{3}{8}$. We consider this network (on 2500 nodes) with 2 and 4 non-geometric edges per node added uniformly at random.

5.3.3.1 Dimensionality

MDS based on Isomap does not identify the embedding dimension 4 of a torus for either $d^{\mathrm{NG}} = 2$ or $d^{\mathrm{NG}} = 4$ (see Figure 5.11(a,f) and Figure 5.12(a,f)). The residual variance based on contagion maps does have a dip at dimension 4 for the threshold $T = 0.2$ (see Figure 5.11(d,i) and Figure 5.12(d,i)), but does not when the threshold is $T = 0.1$ or $T = 0.3$. Note that for Isomap, as well as contagion maps with all considered thresholds, the residual variances are smaller for all considered dimensions when the analysis is done on the point cloud, making the results for contagion map with $T = 0.2$ look sharper.

We identify the approximate embedding dimensions for Isomap and contagion maps with thresholds $T = 0, 0.1, \ldots, 1$ and show the results in Figure 5.13 (for the torus-based network with $d^{\mathrm{NG}} = 2$) and Figure 5.14 (for the torus-based network with $d^{\mathrm{NG}} = 4$). Both versions of Isomap (the one performing MDS based on the entries in D_{iso} and the one performing MDS based on the distances between the rows

of D_{iso}) vastly overestimate the embedding dimension for the torus-based network with $d^{\text{NG}} = 2$ and the one with $d^{\text{NG}} = 4$. When performing MDS based on the entries in D_{iso}, the approximate embedding dimension is at least 100 (which is the dimension at which we cap our computations). When performing MDS based on the distances between the rows of D_{iso}, the embedding dimensions are still very large: It is 61 for the network with $d^{\text{NG}} = 2$, and 95 for the network with $d^{\text{NG}} = 4$. Note that these results are practically identical to those for contagion map with $T = 0$, as we are working with the same unweighted graphs in both Isomap and contagion maps. For contagion maps with varying thresholds T, the approximate embedding dimension has a dip around $T = 0.2$ (see Figure 5.13 (b) and Figure 5.14 (b)), except when performing MDS based on the entries in D_{cont} for the network with $d^{\text{NG}} = 4$, in which case contagion maps return an embedding dimension of at least 100 for all thresholds (see Figure 5.14 (a)). This suggests that, for thresholds close to 0.2, the contagion spreads predominantly via WFP along the underlying torus, making it possible to identify the underlying low-dimensional structure.

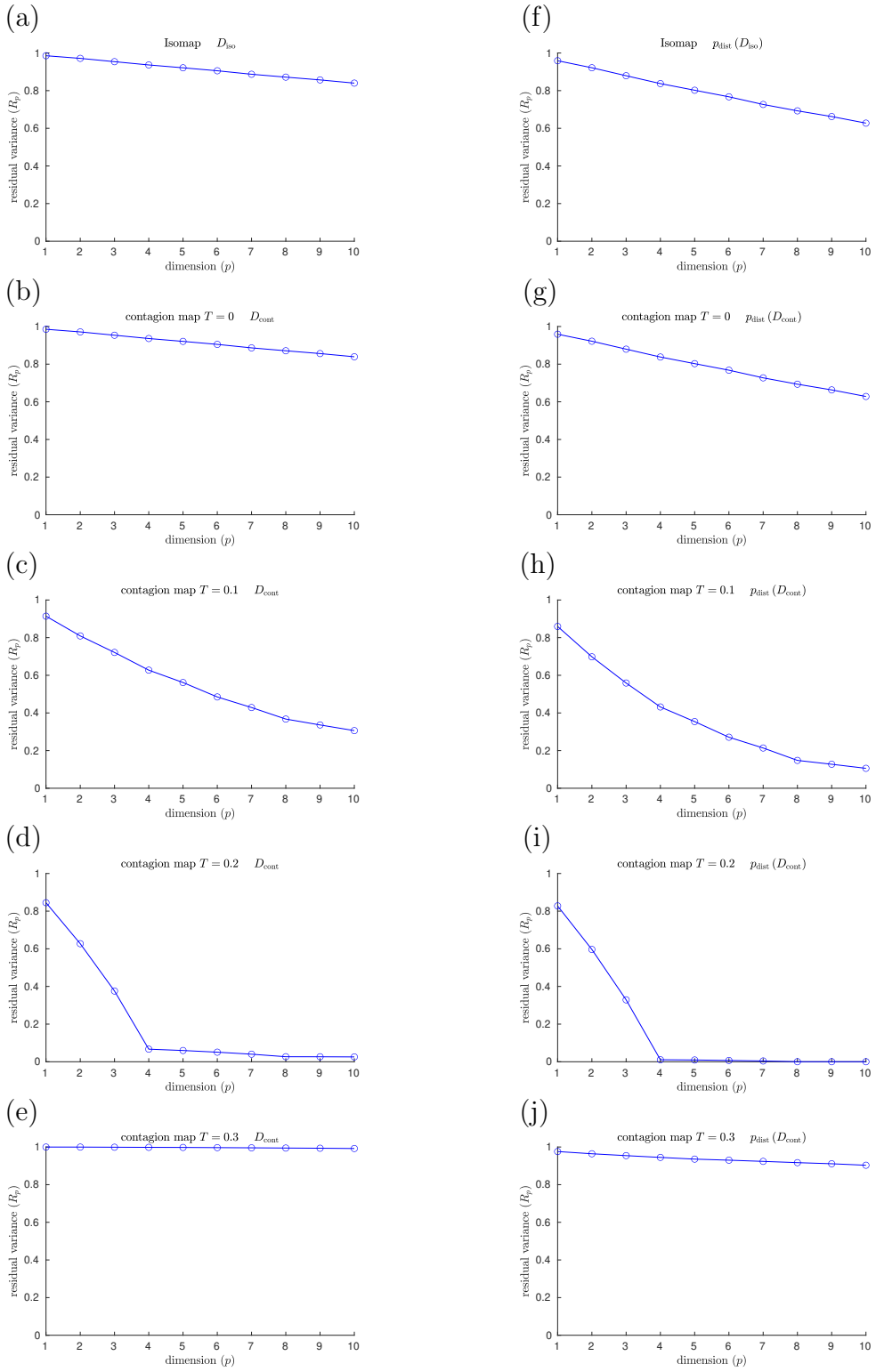


Figure 5.11: Dimensionality results on our torus-based network with $d^{\text{NG}} = 2$. (Left column) Residual variances of MDS based on the estimated geodesic distances (i.e. the entries in D_{iso} and D_{cont}) according to (a) Isomap, and according to contagion maps with thresholds (b) $T = 0$, (c) $T = 0.1$, (d) $T = 0.2$, and (e) $T = 0.3$. (Right column) Residual variances of MDS based on the point cloud (i.e. the rows of the D_{iso} and D_{cont}) according to (f) Isomap, and according to contagion maps with thresholds (g) $T = 0$, (h) $T = 0.1$, (i) $T = 0.2$, and (j) $T = 0.3$.

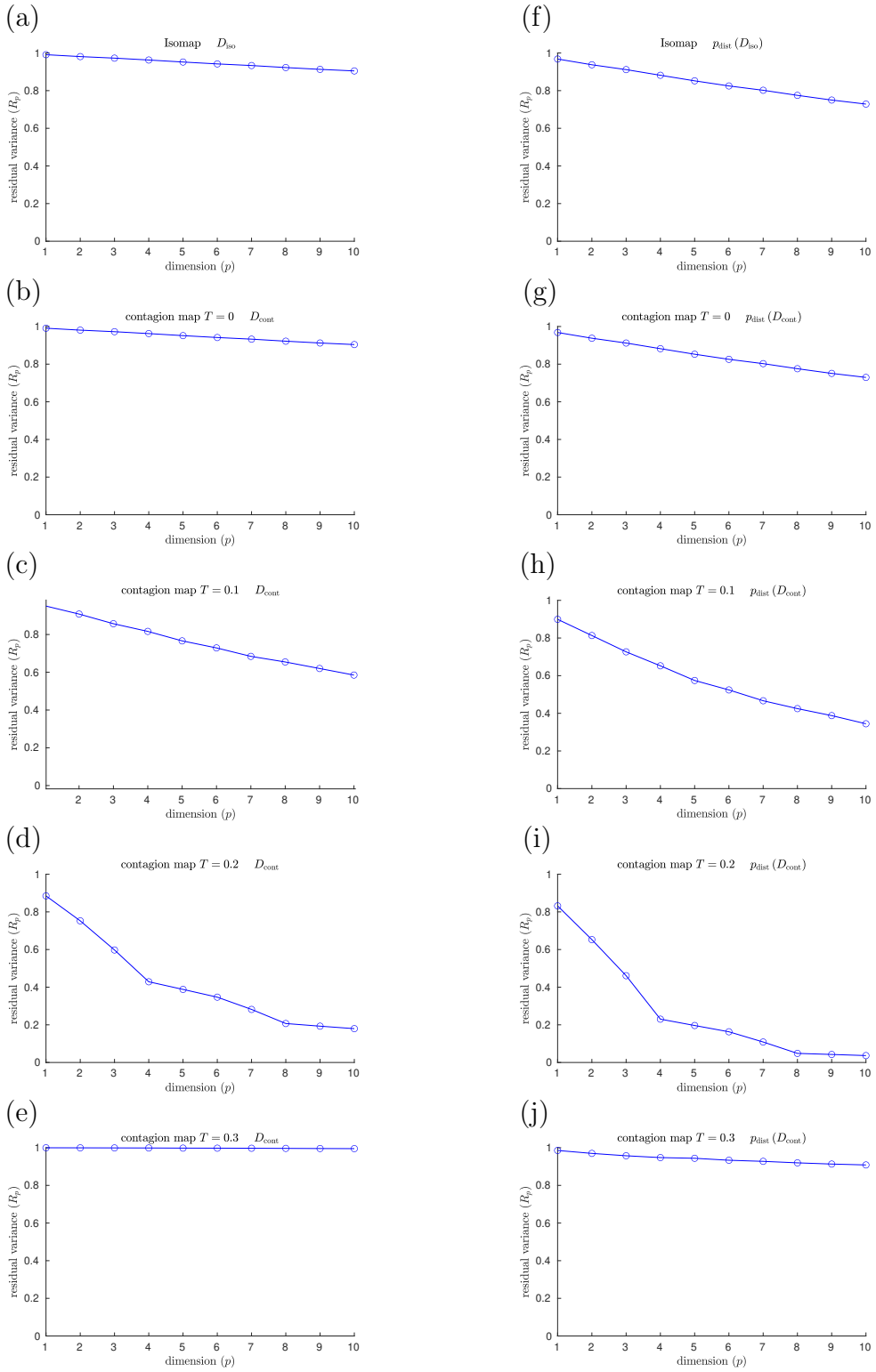


Figure 5.12: Dimensionality results on our torus-based network with $d^{\text{NG}} = 4$. (Left column) Residual variances of MDS based on the estimated geodesic distances (i.e. the entries in D_{iso} and D_{cont}) according to (a) Isomap, and according to contagion maps with thresholds (b) $T = 0$, (c) $T = 0.1$, (d) $T = 0.2$, and (e) $T = 0.3$. (Right column) Residual variances of MDS based on the point cloud (i.e. the rows of the D_{iso} and D_{cont}) according to (f) Isomap, and according to contagion maps with thresholds (g) $T = 0$, (h) $T = 0.1$, (i) $T = 0.2$, and (j) $T = 0.3$.

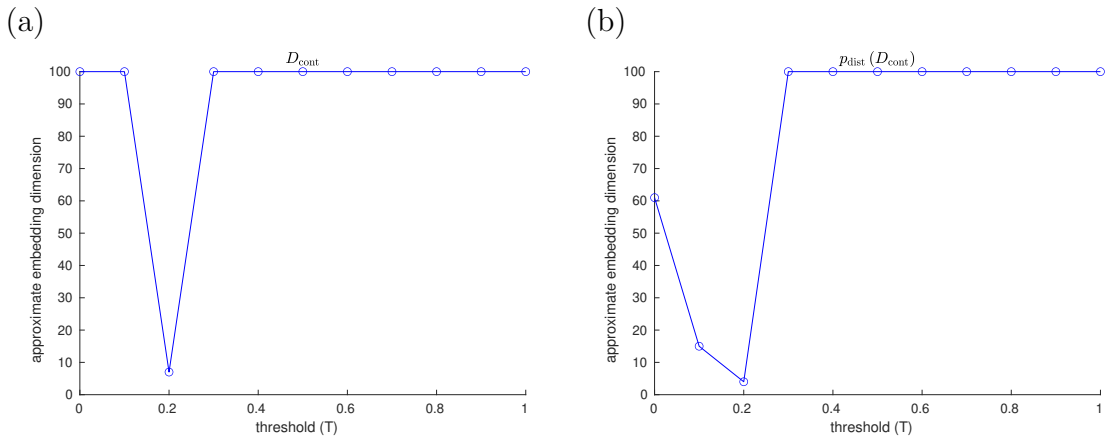


Figure 5.13: Dimensionality results on our torus-based network with $d^{\text{NG}} = 2$. Approximate embedding dimension according to the contagion map with different values of threshold T and critical value 5% using (a) the dissimilarity matrix D_{cont}, and (b) the point cloud whose coordinate vectors are the row in D_{cont}. The results for Isomap are practically identical to those for contagion map with $T = 0$: The approximate embedding dimension is at least 100 (which is the dimension at which we cap our computations) when based on the entries in D_{iso}; and it is 61 when based on the point cloud whose coordinate vectors are given by the rows of D_{iso}. The approximate embedding dimension according to contagion maps reaches a minimum value for $T = 0.2$ both when working with D_{cont} and when working with $p_{\text{dist}}(D_{\text{cont}})$. This minimal value is 7 in the former case (see (a)), and is 4 in the latter case (see (b)).

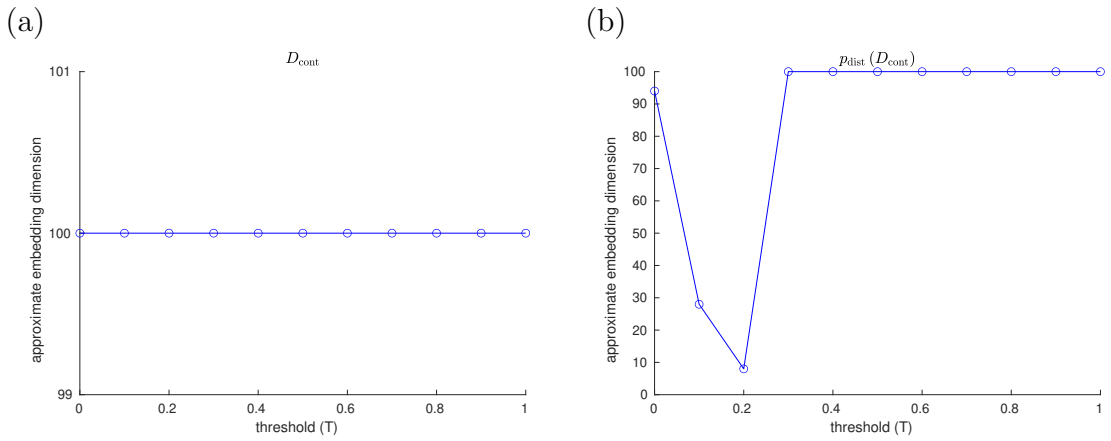


Figure 5.14: Dimensionality results on our torus-based network with $d^{\text{NG}} = 4$. Approximate embedding dimension according to the contagion map with different values of threshold T and critical value 5% using (a) the dissimilarity matrix D_{cont}, and (b) the point cloud whose coordinate vectors are the row in D_{cont}. The results for Isomap are practically identical to those for contagion map with $T = 0$: The approximate embedding dimension is at least 100 (which is the dimension at which we cap our computations) when based on the entries in D_{iso}; and it is 95 when based on the point cloud whose coordinate vectors are given by the rows of D_{iso}. The approximate embedding dimension according to contagion maps reaches a minimum value of 8 for $T = 0.2$ when working with $p_{\text{dist}}(D_{\text{cont}})$ (see (b)).

5.3.3.2 Topology

Figures 5.15 and 5.16 show the barcodes corresponding to the persistent homology in dimension 1 of the Vietoris–Rips filtrations built according to the different versions of Isomap and contagion maps. The barcodes in dimension 1 of the Vietoris–Rips filtrations based on the estimated geodesic distances (i.e. the entries in D_{iso} and D_{cont}) do not seem to reveal any significant features for either Isomap or contagion maps. (See panels (a–e) of Figures 5.15 and 5.16). The barcode in dimension 1 of the Vietoris–Rips filtration on the point cloud based on Isomap (i.e. given by the rows of D_{iso}) on the network with $d^{\text{NG}} = 2$ does feature two dominant bars (see Figure 5.15(f)), as do the barcodes corresponding to point clouds based on contagion maps for $T = 0$, $T = 0.1$, and $T = 0.2$ (i.e. given by the rows of D_{cont}) (see Figure 5.15 (g–i)). Note, however, that in panels (f–h), due to the order of the bars, the dominance appears slightly stronger than it actually is. For the network with $d^{\text{NG}} = 4$

the barcode in dimension 1 of the Vietoris–Rips filtration on the point cloud based on Isomap does not have any dominant bars, whereas the one based on contagion maps for $T = 0.2$ does feature two dominant bars, as is characteristic of the torus (see Figure 5.16 (i)).

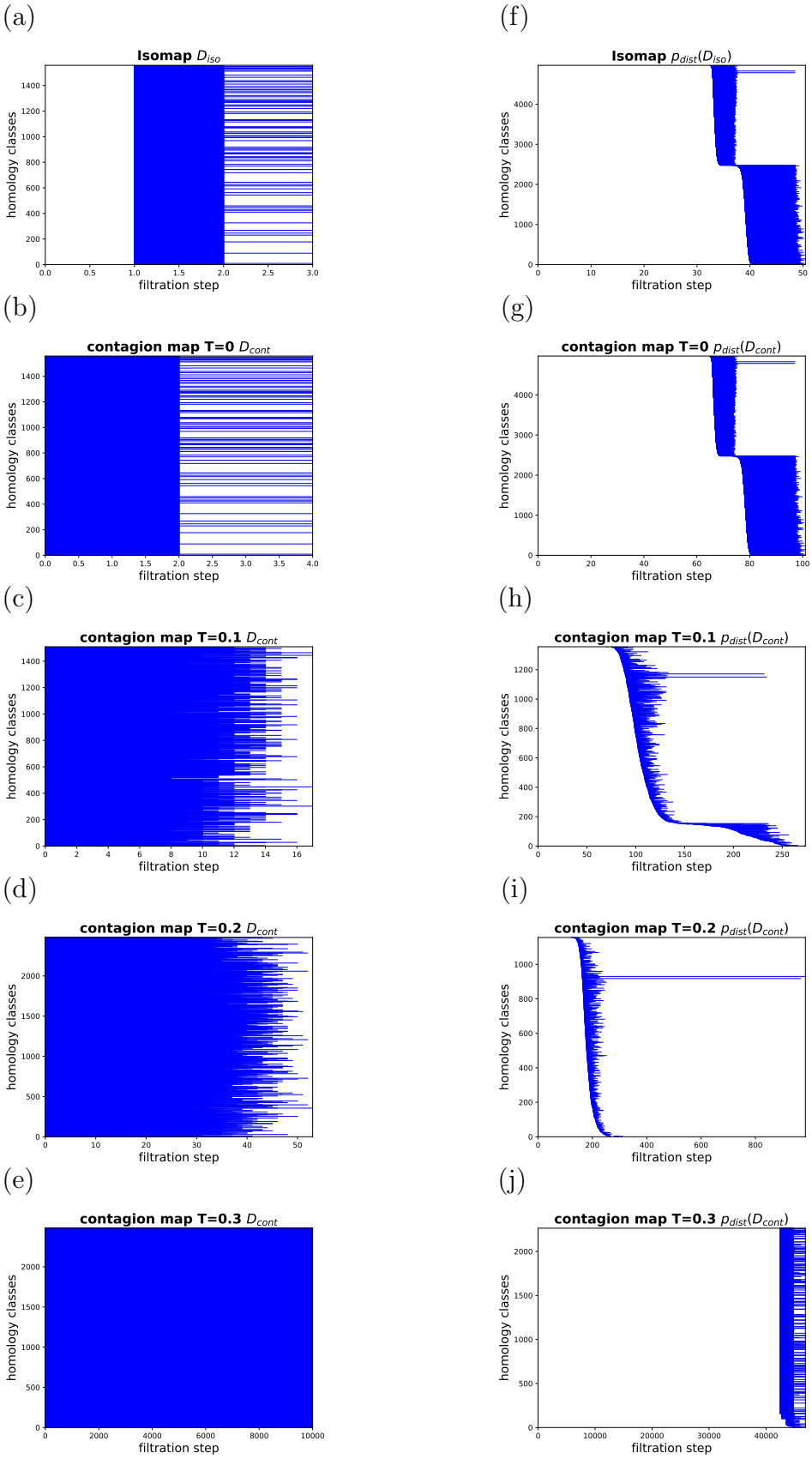


Figure 5.15: Topology results on our torus-based network with $d^{\mathrm{NG}} = 2$. (Left column) Dimension 1 barcodes of the Vietoris–Rips filtrations based on the estimated geodesic distances (i.e. the entries in D_{iso} and D_{cont}) according to (a) Isomap, and according to contagion maps with (b) $T = 0$, (c) $T = 0.1$, (d) $T = 0.2$, and (e) $T = 0.3$. (Right column) Barcodes of the Vietoris–Rips filtrations on the point clouds according to (f) Isomap, and according to contagion maps with (g) $T = 0$, (h) $T = 0.1$, (i) $T = 0.2$, and (j) $T = 0.3$.

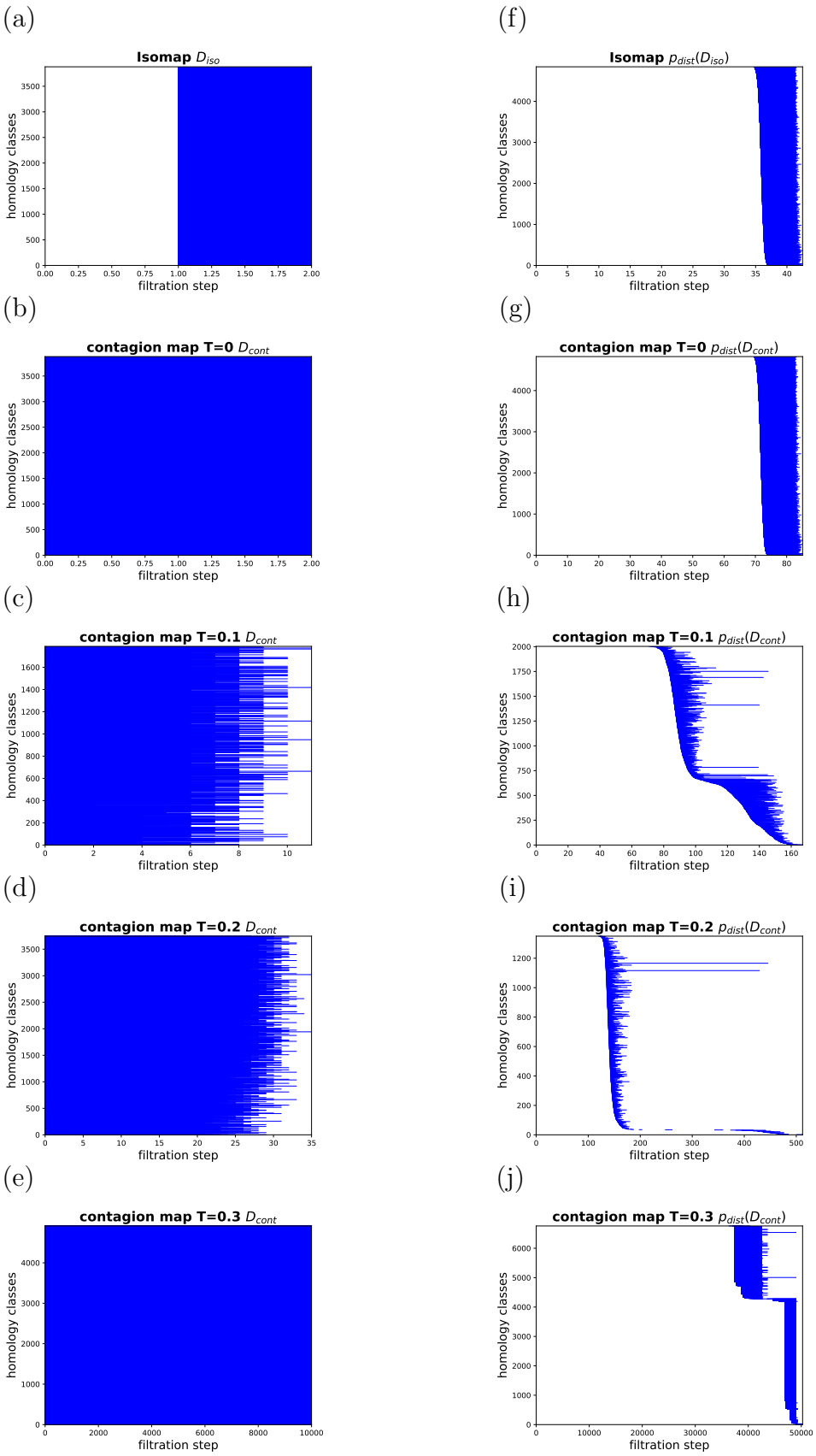


Figure 5.16: Topology results on our torus-based network with $d^{\mathrm{NG}} = 4$. (Left column) Dimension 1 barcodes of the Vietoris–Rips filtrations based on the estimated geodesic distances (i.e. the entries in D_{iso} and D_{cont}) according to (a) Isomap, and according to contagion maps with (b) $T = 0$, (c) $T = 0.1$, (d) $T = 0.2$, and (e) $T = 0.3$. (Right column) Barcodes of the Vietoris–Rips filtrations on the point clouds according to (f) Isomap, and according to contagion maps with (g) $T = 0$, (h) $T = 0.1$, (i) $T = 0.2$, and (j) $T = 0.3$.

5.3.3.3 Geometry

We examine the geometry of our Isomap and contagion map results by comparing the entries of D_{iso} and D_{cont}, as well as the point clouds given by the rows of these dissimilarity matrices, to the regularly spaced points on a torus (4.2) via the Pearson correlation coefficient. See Figure 5.17 for the results for the torus-based network with non-geometric degree $d^{\text{NG}} = 2$ and Figure 5.18 for the results for the torus-based network with non-geometric degree $d^{\text{NG}} = 4$. Isomap returns a low Pearson correlation in all cases, suggesting that the shortest-path distances are (as expected, given the large number of non-geometric edges) not good estimates for the distances along the torus that underlies these networks. Note that these results are practically identical to those for contagion map with $T = 0$, as we are working with the same unweighted graphs in both Isomap and contagion maps. For contagion maps with varying thresholds T, the Pearson correlation coefficient peaks around $T = 0.2$, suggesting that, for thresholds close to $T = 0.2$, the contagion spreads predominantly via WFP, making the activation times good estimates for the distance along the torus that underlies these networks.

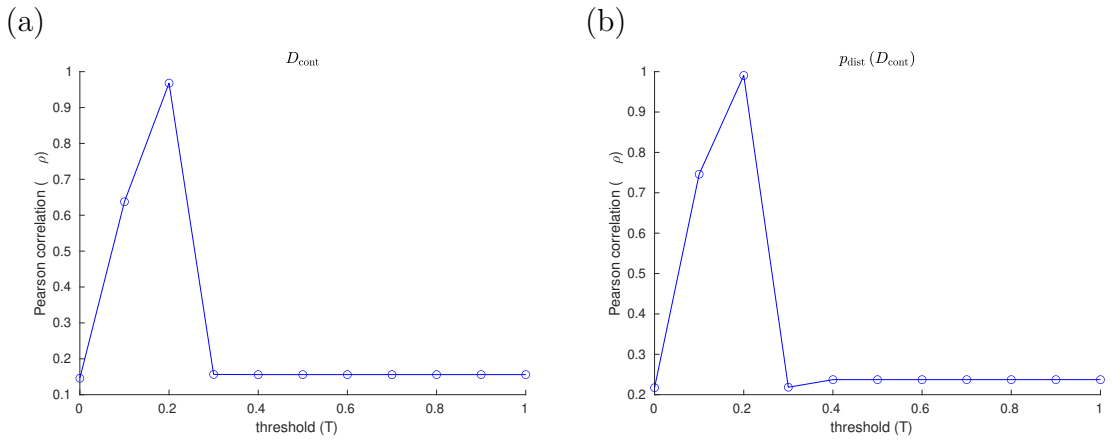


Figure 5.17: Geometry results on our torus-based network with $d^{\text{NG}} = 2$. Pearson correlation coefficient between the pairwise distances between regularly spaced points on a torus (4.2) and the following sets: (a) The estimated geodesic distances (i.e. the entries in D_{cont}) according to contagion maps with thresholds $T = 0, 0.1, \ldots, 1$. (b) The pairwise distances between points in $\mathbb{R}^{2500}$ whose coordinate vectors are the rows of D_{cont} according to contagion maps with thresholds $T = 0, 0.1, \ldots, 1$. The results for Isomap are practically identical to those for contagion map with $T = 0$: The Pearson correlation coefficient is 0.1458 when based on the entries in D_{iso}; and it is 0.2175 when based on the point cloud whose coordinate vectors are given by the rows of D_{iso}.

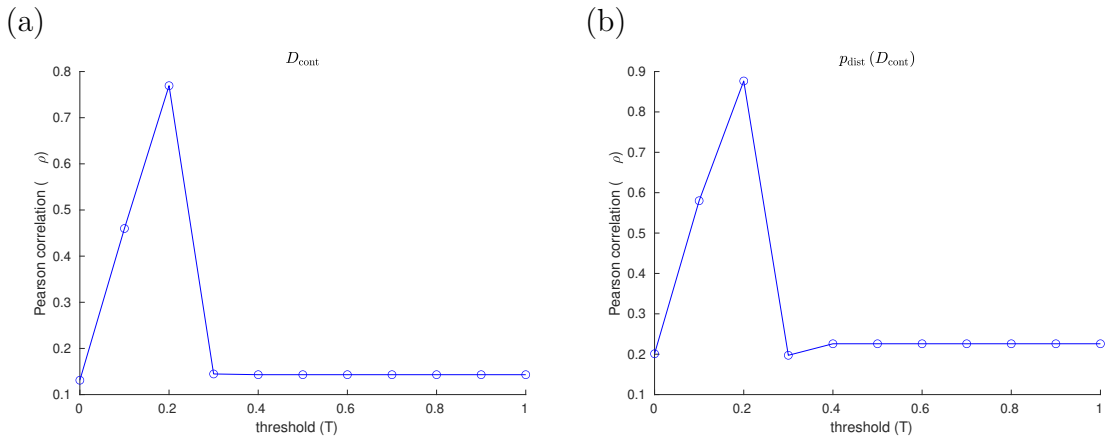


Figure 5.18: Geometry results on our torus-based network with $d^{\text{NG}} = 4$. Pearson correlation coefficient between the pairwise distances between regularly spaced points on a torus (4.2) and the following sets: (a) The estimated geodesic distances (i.e. the entries in D_{cont}) according to contagion maps with thresholds $T = 0, 0.1, \ldots, 1$. (b) The pairwise distances between points in $\mathbb{R}^{2500}$ whose coordinate vectors are the rows of D_{cont} according to contagion maps with thresholds $T = 0, 0.1, \ldots, 1$. The results for Isomap are practically identical to those for contagion map with $T = 0$: The Pearson correlation is 0.1311 when based on the entries in D_{iso}; and it is 0.2013 when based on the point cloud whose coordinate vectors are given by the rows of D_{iso}.

5.3.4 Conformation space of the cyclo-octane molecule

The cyclo-octane molecule $(CH_2)_8$ consists of a ring of eight carbon atoms, each bonded with two hydrogen atoms.

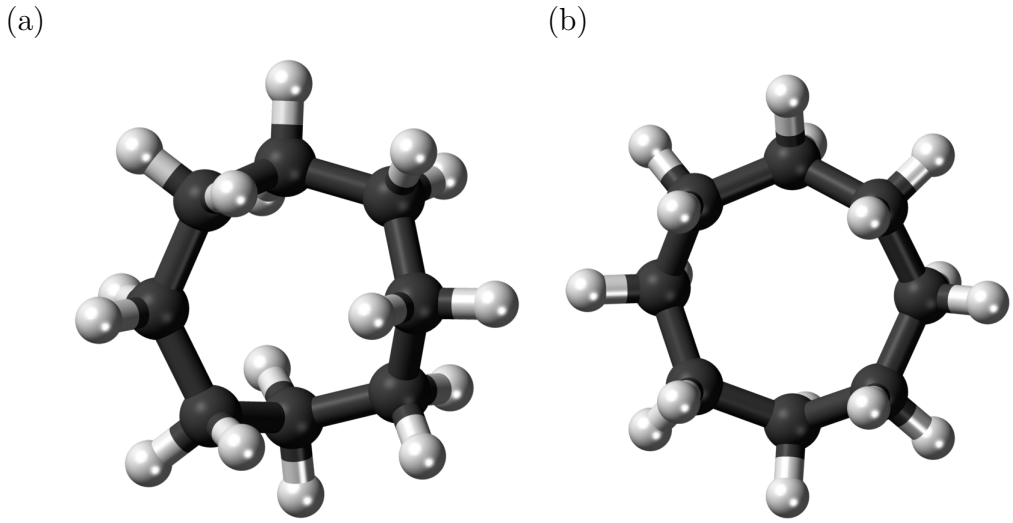


Figure 5.19: Ball-and-stick model representations of the cyclo-octane molecule in (a) its boat-chair conformation and (b) its crown conformation [58]. (Figure taken from [79].)

A *conformation* of a molecule is a possible spatial arrangement of its atoms (modulo rotation and translation) [53]. The conformation of a molecule can be specified by the coordinates of each of its atoms in three-dimensional space, giving a point in $\mathbb{R}^{3a}$, where a is the number of atoms in the molecule. (In this case, each coordinate of each atom in three-dimensional space is a feature and $\mathbb{R}^{3a}$ is the feature space.) The set of such points for all conformations of a molecule is called its *conformation space*. Each conformation is accompanied by a state of potential energy of the molecule, and a conformation is more likely to occur the lower its associated potential energy. The cyclo-octane molecule has many conformations of comparable potential energy, and its conformation space has been studied in computational chemistry for over 50 years [41, 58]. Given the locations of the eight carbon atoms in a conformation of the cyclo-octane molecule, the locations of the hydrogen atoms are determined to mini-

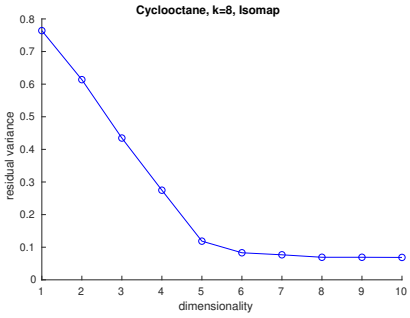
mize energy: The two covalent hydrogen atoms of each carbon atom are positioned to form a tetrahedral arrangement with the two neighbouring carbon atoms that minimizes the potential energy of that subunit of the molecule. The conformation space of cyclo-octane thus lies in $\mathbb{R}^{3\times 8} = \mathbb{R}^{24}$. It is generally assumed that conformation spaces form low-dimensional manifolds, so identifying the structure of the conformation space of a molecule is essentially a manifold-learning problem. The conformation space of cyclo-octane has been shown to be the union of a sphere with a Klein bottle intersecting in two circles of singularities [10,51], forming a two-dimensional manifold with singularities.

Martin *et al.* [51] analysed a data set of 6040 points in the conformation space of cyclo-octane, subsampled from a larger data set consisting of 1 031 644 cyclo-octane conformations. This data set is publicly available as part of the javaPlex software package [70]. To visualize this set of points, Martin *et al.* mapped the points from $\mathbb{R}^{24}$ to $\mathbb{R}^3$ via Isomap. We explore different versions of both Isomap and contagion maps on it.

Figure 5.20 shows the residual variances for projections via MDS onto dimensions 1 to 10 based on shortest-path distances (i.e. based on the entries in D_{iso}), and the visualization of the projection to 3D. Figure 5.21 shows the residual variances for projections via MDS onto dimensions 1 to 10 based on activation times in contagions with thresholds $T = 0.1, 0.2, 0.3$, and 0.4 (i.e. based on the entries in D_{cont}), and the visualizations of the projection to 3D. We see that Isomap and contagion maps with low thresholds ($T = 0.1$ and 0.2) detect the embedding dimension of the underlying space, suggesting the absence of noisy edges in the 8-nearest-neighbour graph on this data set. Contagion maps with higher thresholds ($T = 0.3$ and 0.4) do not seem to reveal any meaningful structure. This is likely due to the contagion stabilizing before much of the graph has been activated, leading to many 'infinite' activation times.

In Figure 5.22 we show barcodes corresponding to the persistent homology in dimension 1 of various Vietoris–Rips filtrations based on the cyclo-octane data set of 6040 points in $\mathbb{R}^{24}$. The barcode in panel (a) corresponds to the Vietoris–Rips filtration built directly on the data points in their ambient space $\mathbb{R}^{24}$. This barcode has one dominant bar, suggesting that the Klein bottle and the sphere (whose union is the conformation space of cyclo-octane) intersect in a way that makes the 1-dimensional loop that is present in the homology of the Klein bottle over $\mathbb{Z}/2\mathbb{Z}$, but not over $\mathbb{Z}$, nullhomotopic. The other panels show barcodes corresponding to various Vietoris–Rips filtrations that, in a sense, mimic the Vietoris–Rips filtration built according to the intrinsic metric on the underlying manifold. Namely, they are Vietoris–Rips filtrations based on different versions of Isomap and contagion map, that is, based on estimates for the geodesic distance on the underlying manifold. Panel (b) shows the barcode corresponding to the shortest-path distances in the 8-nearest-neighbour graph (i.e. based on the entries in D_{iso}). Panel (c) shows the Vietoris–Rips filtration based on the points whose coordinate vectors are the columns of D_{iso}. Panel (d) shows the Vietoris–Rips filtration based on the activation times of the contagion with threshold $T = 0.2$ on the 8-nearest-neighbour graph (i.e. based on the entries in D_{cont}). Panel (e) shows the Vietoris–Rips filtration based on the points whose coordinate vectors are the columns of D_{cont}. All of these barcodes have one dominant bar, consistent with the homology of the underlying manifold. The barcode in panel (d) has many bars with identical birth and death. This stems from the fact that activation times (in our contagion model) have integer values between 0 and $2N$ (where N is the number of node, or, equivalently, data points), and so a Vietoris–Rips filtration based on these values has only few filtration steps at which simplices are added.

(a)



(b)

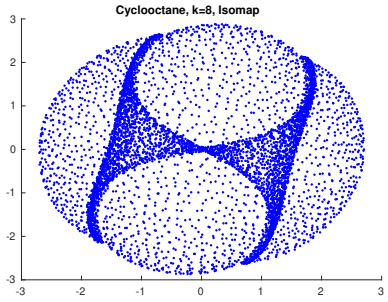


Figure 5.20: Results for Isomap on the data set of 6040 points in the conformation space of cyclo-octane (using an 8-nearest neighbour neighbourhood graph). (a) Residual variances for projections via MDS onto dimensions 1 to 10 in the original Isomap algorithm. (b) Visualization of Isomap in 3D.

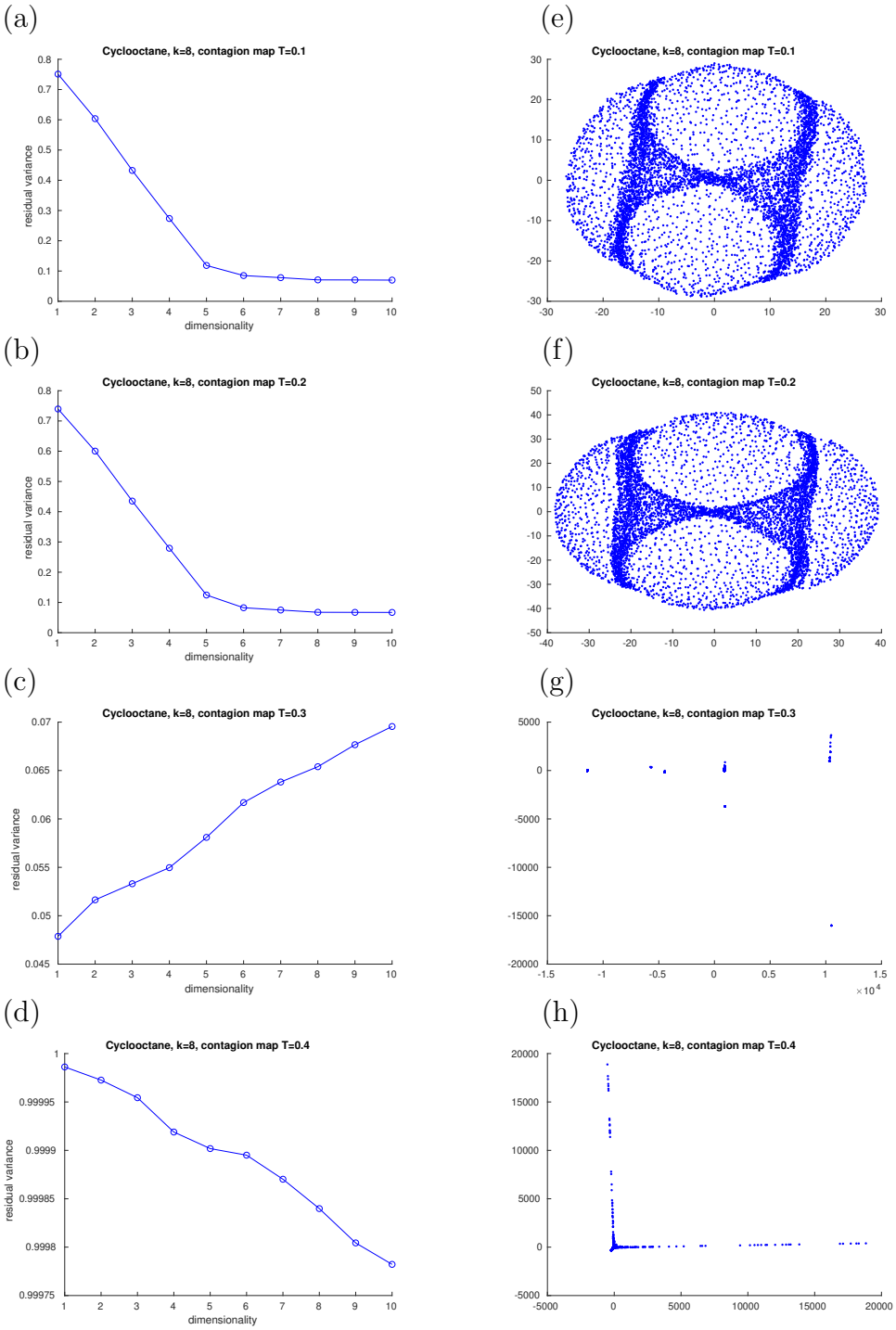


Figure 5.21: Results for contagion maps for different values of T on the data set of 6040 points in the conformation space of cyclo-octane (using an 8-nearest-neighbour graph). (a–d) Residual variances for projections onto dimensions 1 to 10 in the contagion-map algorithm for (a) $T = 0.1$, (b) $T = 0.2$, (c) $T = 0.3$, (b) $T = 0.4$. (e–h) Visualizations of projections to 3D for (e) $T = 0.1$, (f) $T = 0.2$, (g) $T = 0.3$, (h) $T = 0.4$.

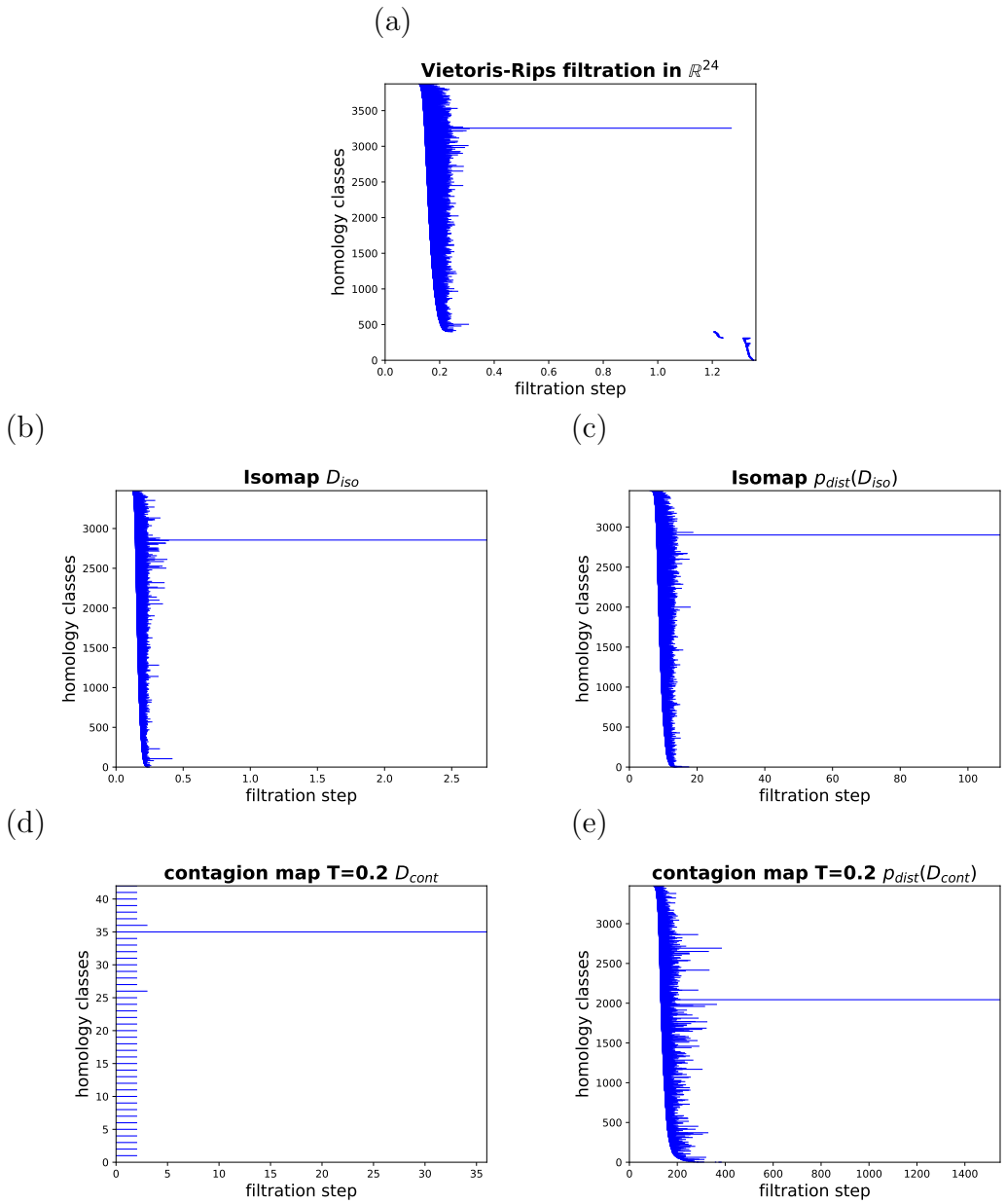


Figure 5.22: Barcodes for the persistent homology in dimension 1 of various Vietoris–Rips filtrations based on the cyclo-octane data set. (a) Vietoris–Rips filtration directly on the data points in $\mathbb{R}^{24}$. (b) Vietoris–Rips filtration based on the shortest-path distances in the 8-nearest-neighbour graph (i.e. based on the entries in D_{iso}). (c) Vietoris–Rips filtration based on the points whose coordinate vectors are the rows of D_{iso} (corresponding to the 8-nearest-neighbour graph). (d) Vietoris–Rips filtration based on the activation times of the contagion with threshold $T = 0.2$ on the 8-nearest-neighbour graph (i.e. based on the entries in D_{cont}). (e) Vietoris–Rips filtration based on the points whose coordinate vectors are the rows of D_{cont} (corresponding to the contagion with threshold $T = 0.2$ on the 8-nearest-neighbour graph).

5.4 Conclusions and future work

Isomap is a well established manifold-learning tool and is useful for many data sets. It can successfully handle curvature of data in many cases, and the freedom of choosing the parameter k or ϵ when creating a neighbourhood graph allows it to handle noise to some extent. However, when faced with particularly sparse and noisy data, Isomap is prone to so-called 'short-circuit errors', which, in some cases, cannot be avoided regardless of the choice of k or ϵ. For such data, contagion maps can be a more adequate tool. For a suitable choice of the threshold parameter T, single noisy edges that occur in a neighbourhood graph do not carry a contagion and thus do not distort the estimate of the geodesic distances via activation times significantly. In other words, with the right choice of T, contagion maps are able to 'exploit social reinforcement to silence noise'.

There are manifold avenues to continue and build on the work in this chapter. Our results leave us with multiple open questions: What effect does the operation p_{dist} of taking pairwise distances between columns have on a dissimilarity matrix, and on our various measures of geometry, topology, and dimensionality based on such a matrix? We have seen in some instances that this operation seems to accentuate our results (see e.g. Figure 5.12 (d,i), or Figure 5.13). Moreover, in our method for determining topological features of a data set, the p_{dist} operation does not only emphasize results, but seems to be — in some cases — a necessary methodological step. Building the Vietoris–Rips filtration based on a dissimilarity matrix D can lead to degenerate barcodes in cases where building the Vietoris–Rips filtration based on $p_{\text{dist}}(D)$ seems to bring out meaningful features (see Figures 5.15 and 5.16). This might, in part, be due to the entries in a dissimilarity matrix D not necessarily forming a metric on the data points. Applying p_{dist} to D remedies this issue, while (at least intuitively) preserving the information held in D. How exactly p_{dist} transforms D is another

question to ask.

To establish contagion maps as a manifold-learning tool, it should be useful to test them under various conditions of data other than just noisiness and curvature of the underlying structure. Such conditions include traits such as non-uniform density and the presence of outliers, both of which are common in real-world data. In particular, it would be interesting to test how contagion maps perform under such conditions compared to Isomap as well as other dimensionality-reduction techniques, such as diffusion maps [21], commute-time distance approaches [25], and locally linear embeddings (LLE) [65]. Furthermore, one could explore a 'weighted version' of contagion maps that is defined through a contagion model that takes edge weights in a network into account in some way. Likewise, one could check how forgetting edge weights in a neighbourhood graph affects the performance of Isomap.

Appendix A

Persistent Homology of a Torus with various Choices of Metric

Consider the torus as the Cartesian product of two circles: $\mathbb{T} = \mathbb{S}^1 \times \mathbb{S}^1$. Each copy of $\mathbb{S}^1 \subset \mathbb{C}$ can be viewed as the subset of points in $\mathbb{C}$ that are at a fixed distance r from 0 according to the Euclidean metric on $\mathbb{C}$. Therefore, $\mathbb{T}$ can be viewed as lying in $\mathbb{C} \times \mathbb{C} \cong \mathbb{R}^4$, constituting a subset of the 3-sphere $\mathbb{S}^3$.

Conventionally, the radius of a circle $\mathbb{S}^1$ lying in $\mathbb{C}$ is often taken to be $r = 1$ according to the Euclidean metric. This makes the torus $\mathbb{T} = \mathbb{S}^1 \times \mathbb{S}^1$ a subset of the 3-sphere $\mathbb{S}^3$ of radius $\sqrt{2}$ (according to the Euclidean metric on $\mathbb{R}^4$). However, we will consider a circle $\mathbb{S}^1$ lying in $\mathbb{C}$ to have a circumference of 1, so its radius is $\frac{1}{2\pi}$. To emphasize this, we note the factor of $\frac{1}{2\pi}$ and write the torus as

$$\mathbb{T} = \frac{1}{2\pi}\mathbb{S}^1 \times \frac{1}{2\pi}\mathbb{S}^1 \subset \mathbb{C} \times \mathbb{C} \cong \mathbb{R}^4 \,. \tag{A.1}$$

We are interested in Vietoris–Rips filtrations on clouds of points that are evenly distributed on such a torus. In order to determine locations of points on the torus that are evenly distributed and to construct the Vietoris–Rips filtration, one needs to specify a metric on $\mathbb{T}$. We will consider a number of different such metrics.

Recall that, for two metric spaces (X, d_X) and (Y, d_Y), one can define a metric on the product $X \times Y$ for any $p \in \mathbb{N}$ as follows. Let $(x_1, y_1), (x_2, y_2) \in (X \times Y)$, then

$$d_p\left[(x_1, y_1), (x_2, y_2)\right] = \left[d_X(x_1, x_2)^p + d_Y(y_1, y_2)^p\right]^{\frac{1}{p}} .$$

Also,

$$d_\infty\left[(x_1,y_1),(x_2,y_2)\right]=\max\left\{d_X(x_1,x_2),d_Y(y_1,y_2)\right\}.$$

Clearly, this construction can be extended to products of more than two metric spaces. Most of the metrics on $\mathbb{T}$ that we will consider are product metrics of this kind.

A.1 The $d_{2,1}$ metric

Consider $\mathbb{T}$ as a subspace of $\mathbb{C}\times\mathbb{C}$, which can be identified with the product $\mathbb{R}\times\mathbb{R}\times\mathbb{R}\times\mathbb{R}$. The absolute difference is a metric on $\mathbb{R}$. Taking the d_2 product metric on $\mathbb{C}\cong\mathbb{R}^2$ induced by the absolute difference yields the standard Euclidean metric on $\mathbb{C}$. Taking the d_1 product metric induced by the Euclidean metric on $\mathbb{C}$ induces a metric on $\mathbb{C}^2$ that we will denote by $d_{2,1}$ and that forms a metric on $\mathbb{T}$ as a subspace of $\mathbb{C}\times\mathbb{C}$.

Now, $N=n^2$ points can be chosen to be evenly distributed on $\mathbb{T}$ according to the $d_{2,1}$ metric such that their coordinates are

$$\frac{1}{2\pi}\left(\cos\left(\frac{2\pi i}{n}\right),\sin\left(\frac{2\pi i}{n}\right),\cos\left(\frac{2\pi j}{n}\right),\sin\left(\frac{2\pi j}{n}\right)\right)$$

for $i,j\in\{0,1\ldots n-1\}$. Note that two points

$$q_1=\left(\cos\left(\theta_1\right),\sin\left(\theta_1\right)\right),\ \ q_2=\left(\cos\left(\theta_2\right),\sin\left(\theta_2\right)\right)\in\mathbb{S}^1\subset\mathbb{R}^2$$

are at a distance of

$$\begin{aligned}
d_2(q_1,q_2)&=\left[(\cos(\theta_1)-\cos(\theta_2))^2+(\sin(\theta_1)-\sin(\theta_2))^2\right]^{\frac{1}{2}} \qquad \text{(A.2)}\\
&=\left[\cos^2\left(\theta_1\right)-2\cos\left(\theta_1\right)\cos\left(\theta_2\right)+\cos^2\left(\theta_2\right)+\sin^2\left(\theta_1\right)-2\sin\left(\theta_1\right)\sin\left(\theta_2\right)+\sin^2\left(\theta_2\right)\right]^{\frac{1}{2}}\\
&=\left[2-2\cos\left(\theta_1\right)\cos\left(\theta_2\right)-2\sin\left(\theta_1\right)\sin\left(\theta_2\right)\right]^{\frac{1}{2}}\\
&=\sqrt{2}\left[1-\cos\left(\theta_1-\theta_2\right)\right]^{\frac{1}{2}}\\
&=\sqrt{2}\left[2\sin^2\left(\frac{1}{2}\left(\theta_1-\theta_2\right)\right)\right]^{\frac{1}{2}}\\
&=2\left|\sin\left(\frac{1}{2}(\theta_1-\theta_2)\right)\right|
\end{aligned}$$

from one another. Therefore, the distance between two points

$$p_1 = \frac{1}{2\pi}\left(\cos\left(\frac{2\pi i}{n}\right), \sin\left(\frac{2\pi i}{n}\right), \cos\left(\frac{2\pi j}{n}\right), \sin\left(\frac{2\pi j}{n}\right)\right)$$

and

$$p_2 = \frac{1}{2\pi}\left(\cos\left(\frac{2\pi k}{n}\right), \sin\left(\frac{2\pi k}{n}\right), \cos\left(\frac{2\pi l}{n}\right), \sin\left(\frac{2\pi l}{n}\right)\right)$$

is

$$d_{2,1}\left(p_1, p_2\right) = \frac{1}{\pi}\left(\sin\left(\frac{|i-k|\pi}{n}\right) + \sin\left(\frac{|j-l|\pi}{n}\right)\right).$$

The Vietoris–Rips filtration on this evenly distributed point cloud has a number of distinct times where several homology classes are born or die in its persistent homology. To find these distinct points in time, one needs to find the distances between points that are the closest together.

At time 0 there are $N = n^2$ components, no one-dimensional loops and no voids. That is, the first three Betti numbers are $\beta_0 = n^2, \beta_1 = 0$, and $\beta_2 = 0$. This state remains until the time when the balls around the points that are the closest together merge. The points that are at a minimal distance from each other are points that lie at the same angle in the first circle $\mathbb{S}^1$ (i.e. the first two coordinates coincide and the points lie on the same second 'copy' of $\mathbb{S}^1$) and are next to each other (at an angle of $\frac{2\pi}{n}$) on (this copy of) the second circle $\mathbb{S}^1$, or vice versa. Such pairs of points have coordinates

$$p_1 = \frac{1}{2\pi}\left(\cos\left(\frac{2\pi i}{n}\right), \sin\left(\frac{2\pi i}{n}\right), \cos\left(\frac{2\pi j}{n}\right), \sin\left(\frac{2\pi j}{n}\right)\right)$$

and

$$p_2 = \frac{1}{2\pi}\left(\cos\left(\frac{2\pi i}{n}\right), \sin\left(\frac{2\pi i}{n}\right), \cos\left(\frac{2\pi (j+1)}{n}\right), \sin\left(\frac{2\pi (j+1)}{n}\right)\right),$$

or

$$p_2' = \frac{1}{2\pi}\left(\cos\left(\frac{2\pi (i+1)}{n}\right), \sin\left(\frac{2\pi (i+1)}{n}\right), \cos\left(\frac{2\pi j}{n}\right), \sin\left(\frac{2\pi j}{n}\right)\right).$$

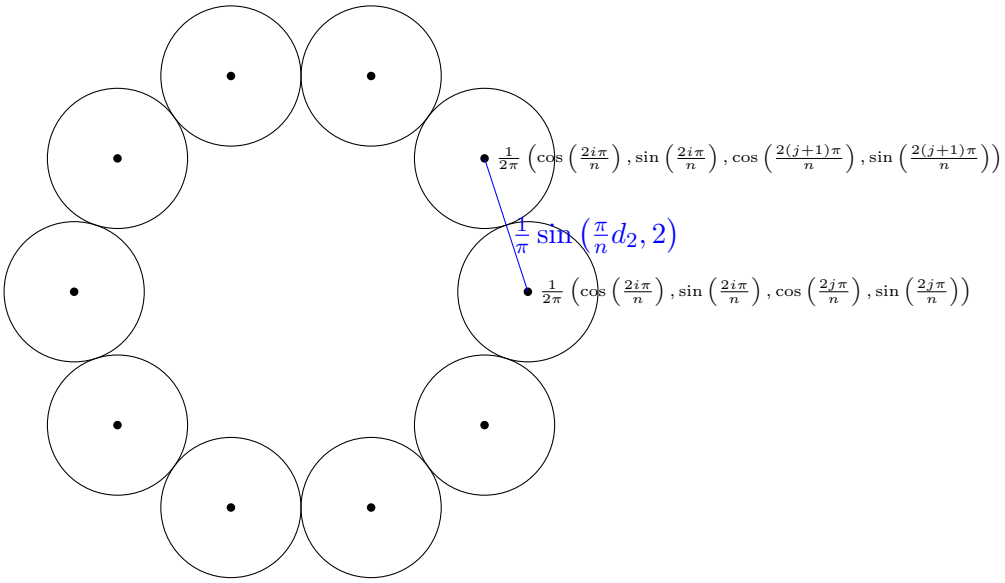


Figure A.1: Points in $\left\{\frac{1}{2\pi}\left(\cos\left(\frac{2i\pi}{n}\right),\sin\left(\frac{2i\pi}{n}\right)\right)\right\}\times\frac{1}{2\pi}S^1$. The two toroidal loops are formed at time $t=\frac{1}{\pi}\sin\left(\frac{\pi}{n}\right)$.

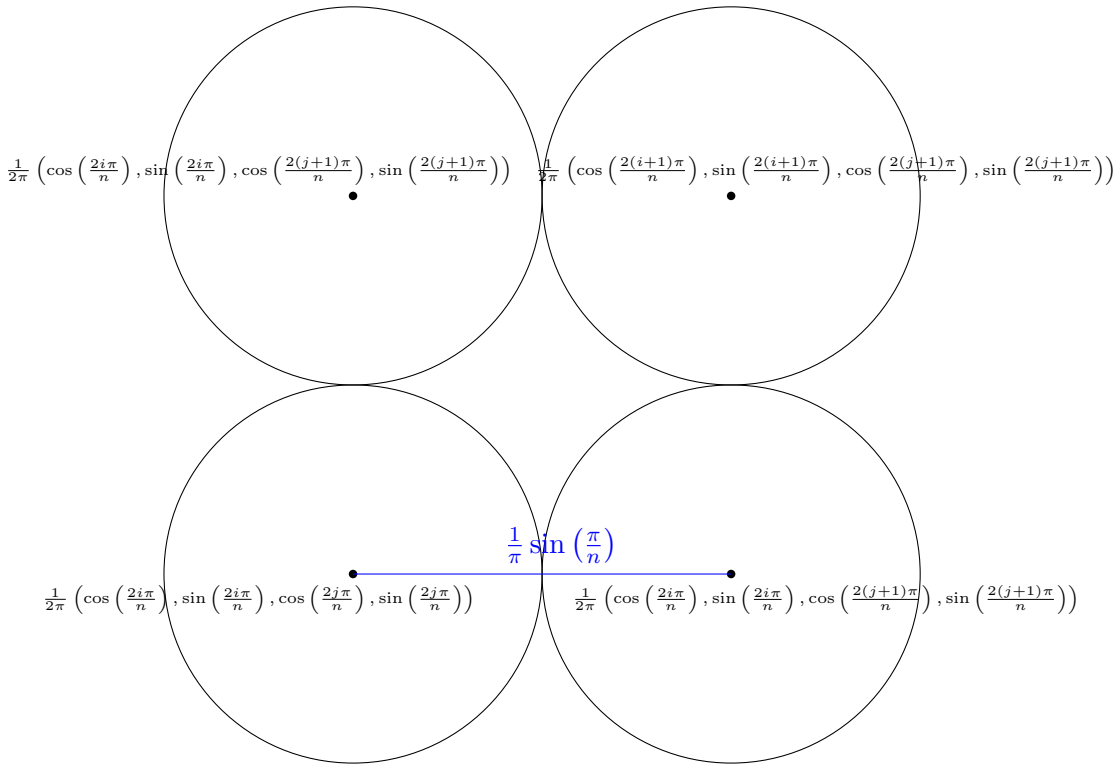


Figure A.2: Loops in the toroidal 'surface' appear at time $t=\frac{1}{\pi}\sin\left(\frac{\pi}{n}\right)$.

The distance between them is

$$d_{2,1}\left(p_1, p_2\right) = \frac{1}{\pi}\sin\left(\frac{\pi}{n}\right) \ .$$

Two points that lie at this minimal distance from each other are, for example, the points $\frac{1}{2\pi}\left(1,0,1,0\right)$ and $\frac{1}{2\pi}\left(1,0,\cos\left(\frac{2\pi}{n}\right),\sin\left(\frac{2\pi}{n}\right)\right)$, both lying at the zero angle in the first circle $\mathbb{S}^1$, and at the zero angle (the first point) and at a positive $\frac{2\pi}{n}$ angle (the second point) in the second circle $\mathbb{S}^1$, respectively.

Note that

$$\mathbb{T}\backslash\{k \text{ points}\} \simeq \mathbb{S}^1 \vee \mathbb{S}^1 \vee \bigvee_{k-1} \mathbb{S}^1 \ ,$$

so $H_1\left(\mathbb{T}\backslash\{k \text{ points}\}\right) \cong \mathbb{Z}^{k+1}$ and $\beta_1\left(\mathbb{T}\backslash\{k \text{ points}\}\right) = k+1$. Therefore, at time $t = \frac{1}{\pi}\sin\left(\frac{\pi}{n}\right)$, the $N = n^2$ components in the Vietoris–Rips filtration merge to form n^2+1 one-dimensional loops. The Betti numbers are then $\beta_0 = 1, \beta_1 = n^2+1$, and $\beta_2 = 0$.

To find the next point in time when the topological features of the Vietoris–Rips filtration change qualitatively, one needs to find the distance between points that are second closest to each other. These are the pairs of points that are separated by an angle of $\frac{2\pi}{n}$ in each of the two circles that make up the torus as a Cartesian product $\mathbb{T} = \frac{1}{2\pi}\mathbb{S}^1 \times \frac{1}{2\pi}\mathbb{S}^1$. That is, they have coordinates

$$p_3 = \frac{1}{2\pi}\left(\cos\left(\frac{2\pi i}{n}\right), \sin\left(\frac{2\pi i}{n}\right), \cos\left(\frac{2\pi j}{n}\right), \sin\left(\frac{2\pi j}{n}\right)\right)$$

and

$$p_4 = \frac{1}{2\pi}\left(\cos\left(\frac{2\pi(i+1)}{n}\right), \sin\left(\frac{2\pi(i+1)}{n}\right), \cos\left(\frac{2\pi(j+1)}{n}\right), \sin\left(\frac{2\pi(j+1)}{n}\right)\right) \ .$$

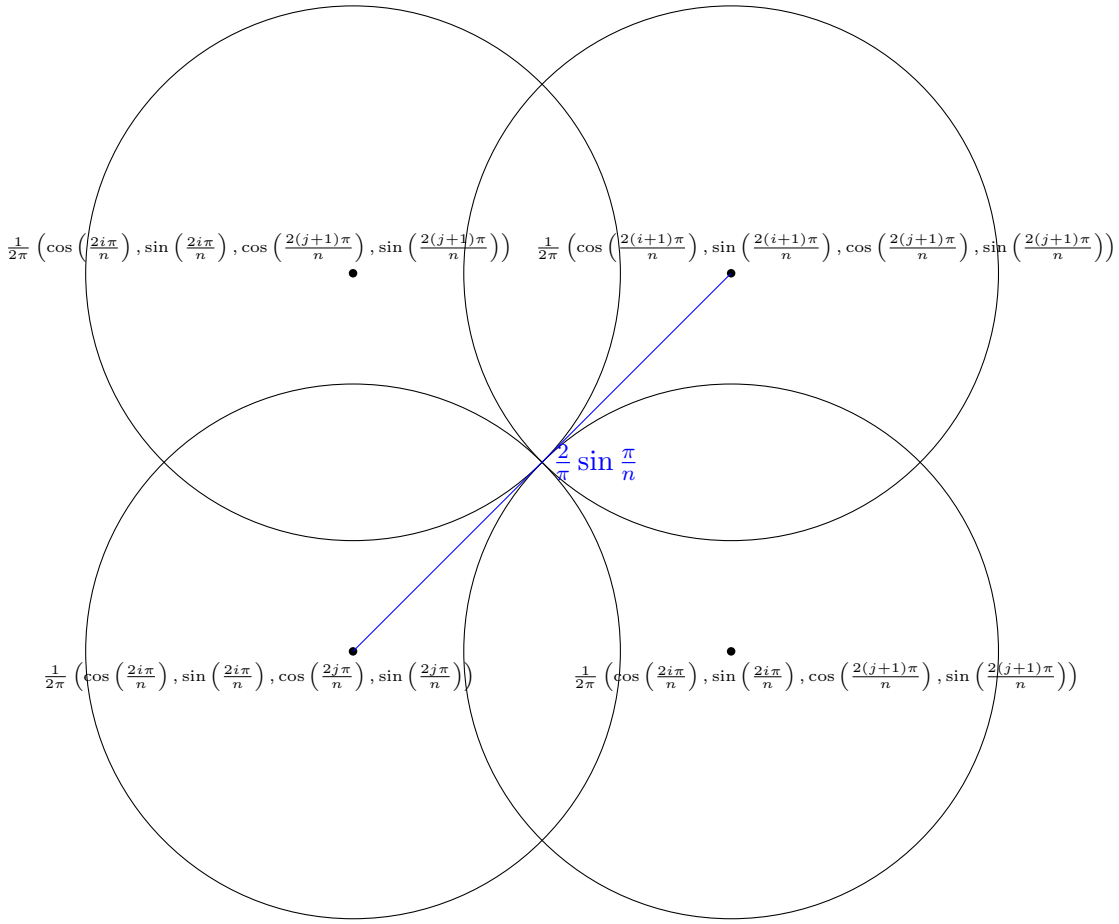


Figure A.3: At time $t = \frac{2}{\pi}\sin\left(\frac{\pi}{n}\right)$, the n^2 toroidal 'surface holes' close, creating a torus.

The distance between them is

$$d_{2,1}\left(p_3, p_4\right) = \frac{2}{\pi}\sin\frac{\pi}{n}\ .$$

So at time $t = \frac{2}{\pi}\sin\left(\frac{\pi}{n}\right)$, the n^2 'surface holes' fill out, killing $n^2 - 1$ of the $n^2 + 1$ homology classes in dimension 1 and creating one homology class in dimension 2. The Betti numbers are now $\beta_0 = 1, \beta_1 = 2$, and $\beta_2 = 1$, i.e. those of a torus.

The two remaining loops disappear simultaneously with the one void at the moment in time when the balls around two points that are at the same point in one of the circles $\mathbb{S}^1$ (i.e. in the same 'copy' of the second circle $\mathbb{S}^1$) and at a maximal distance in the second circle $\mathbb{S}^1$ merge.

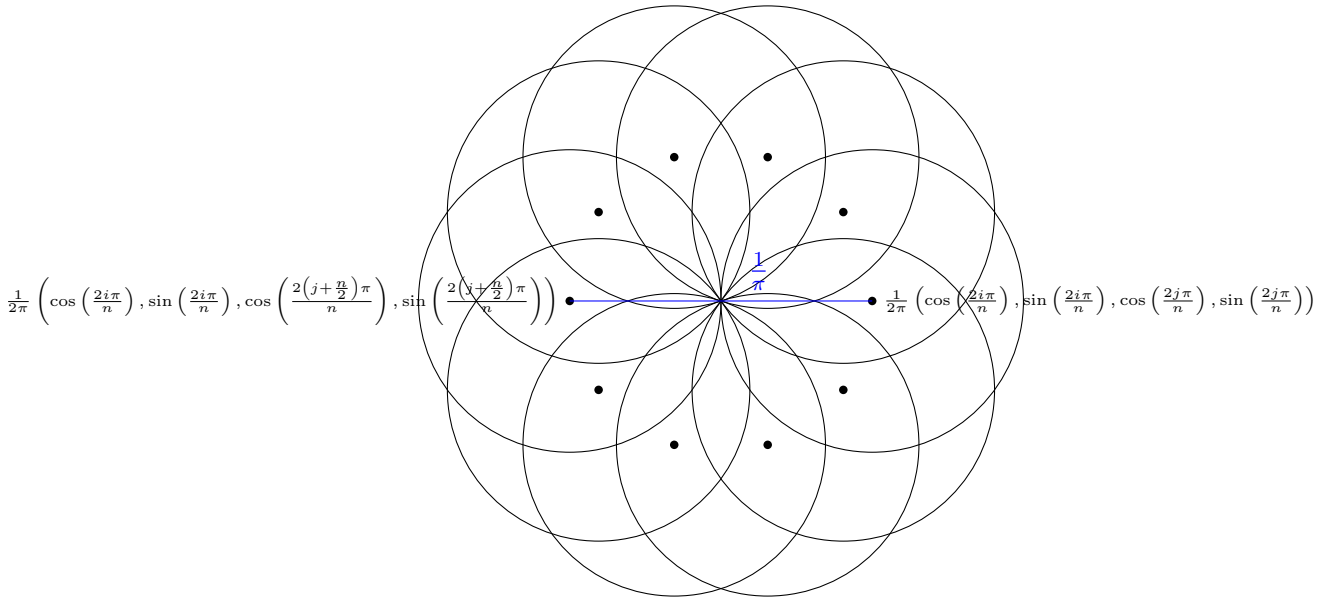


Figure A.4: At time $t = \frac{1}{\pi}$ for n even, and $t = \frac{1}{\pi}\sin\left(\frac{\pi(n-1)}{2n}\right)$ for n odd, all balls of pairs touch and the Vietoris–Rips complex becomes an n^2-dimensional simplex.

Finding the distance between two such points means finding the maximal value of $\frac{1}{\pi}\sin\left(\frac{\pi i}{n}\right)$ for $i \in \mathbb{N}$. This maximal value is achieved when $\frac{\pi i}{n}$ is closest to $\frac{\pi}{2}$, that is, when $i = \lfloor \frac{n}{2} \rfloor$. For n even, the maximal value is achieved when $i = \frac{n}{2}$ and is $\frac{1}{\pi}\sin\left(\frac{\pi}{2}\right) = \frac{1}{\pi}$. For n odd, the maximal value is achieved when $i = \lfloor \frac{n}{2} \rfloor = \frac{n-1}{n}$ and is $\frac{1}{\pi}\sin\frac{\pi(n-1)}{2n}$. Roughly speaking, such two points are on opposite sides of the circle. Therefore, at time $t = \frac{1}{\pi}$ for n even, and $t = \frac{1}{\pi}\sin\left(\frac{\pi(n-1)}{2n}\right)$ for n odd, the balls around the points all merge into one component and both the two loops and the void disappear. The barcodes corresponding to the persistent homology of this Vietoris–Rips filtration in dimensions $0, 1$ and 2 are thus as in Figure A.5.

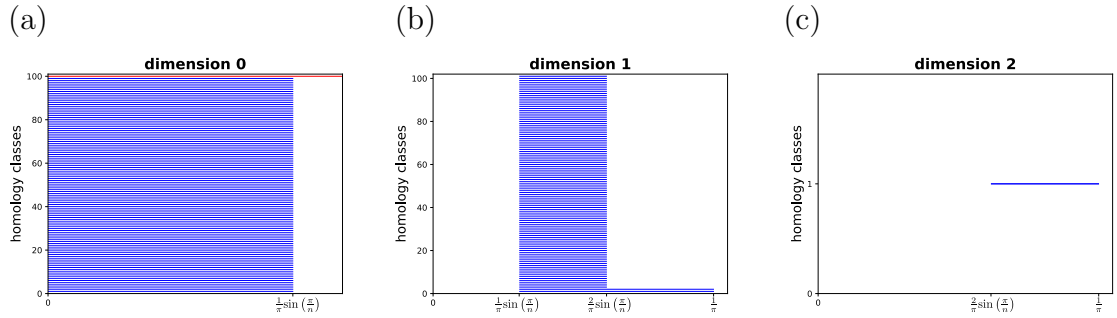


Figure A.5: Barcodes for the persistent homology in dimension (a) 0, (b) 1, and (c) 2 of the Vietoris–Rips filtration according to the $d_{2,1}$-metric. (The red bar in (a) never dies.)

A.2 The $d_{2,2}$ metric

The $d_{2,2}$ metric on the torus $\mathbb{T} = \frac{1}{2\pi}\mathbb{S}^1 \times \frac{1}{2\pi}\mathbb{S}^1 \subset \mathbb{C} \times \mathbb{C} \cong \mathbb{R}^4$ is, similarly to the $d_{2,1}$ metric described in the previous section, built as the product metric induced by the product metric induced by the absolute difference on $\mathbb{R}$. In this case we take the d_2 product metric on $\mathbb{C} \cong \mathbb{R}^2$ induced by the absolute difference and then the d_2 product metric on $\mathbb{C}^2$ induced by this metric. This yields the standard Euclidean metric on $\mathbb{R}^4$ and induces a metric on the torus $\mathbb{T}$ as a subspace of $\mathbb{R}^4$.

The arguments and calculations needed to calculate the persistent homology of the Vietoris–Rips filtration according to the $d_{2,2}$ metric are very similar to those according to the $d_{2,1}$ metric done in the previous section. The persistent homology here only differs in the time step where all but two loops die and a void appears, which occurs earlier by a factor of $\sqrt{2}$ in the case of the $d_{2,2}$ metric. This means that the topological features of a torus are more pronounced in this case. In summary we have the following.

A cloud of $N = n^2$ points can be chosen to be evenly distributed on $\mathbb{T}$ according to the $d_{2,2}$ metric such that their coordinates are (just as when working with the $d_{2,1}$ metric)

$$\frac{1}{2\pi}\left(\cos\left(\frac{2\pi i}{n}\right), \sin\left(\frac{2\pi i}{n}\right), \cos\left(\frac{2\pi j}{n}\right), \sin\left(\frac{2\pi j}{n}\right)\right)$$

for $i, j \in \{0, 1, \ldots, n-1\}$.

Using A.2, we can see that the distance between two points

$$p_1 = \frac{1}{2\pi}\left(\cos\left(\frac{2\pi i}{n}\right), \sin\left(\frac{2\pi i}{n}\right), \cos\left(\frac{2\pi j}{n}\right), \sin\left(\frac{2\pi j}{n}\right)\right)$$

and

$$p_2 = \frac{1}{2\pi}\left(\cos\left(\frac{2\pi k}{n}\right), \sin\left(\frac{2\pi k}{n}\right), \cos\left(\frac{2\pi l}{n}\right), \sin\left(\frac{2\pi l}{n}\right)\right)$$

is

$$d_{2,2}(p_1, p_2) = \frac{1}{\pi}\left(\sin^2\left(\frac{|i-k|\pi}{n}\right) + \sin^2\left(\frac{|j-l|\pi}{n}\right)\right)^{\frac{1}{2}}.$$

The point in time when the initial n^2 components of the Vietoris–Rips filtration merge and create n^2+1 one-dimensional loops is (as with the $d_{2,1}$ metric) $t = \frac{1}{\pi}\sin\left(\frac{\pi}{n}\right)$. The point in time when n^2-1 of these one-dimensional loops die leaving 2 loops and creating a void is $t = \frac{\sqrt{2}}{\pi}\sin\left(\frac{\pi}{n}\right)$. The point in time when all balls merge, leaving one component and no higher-dimensional features is (as with the $d_{2,1}$ metric) $t = \frac{1}{\pi}$ for n even, and $t = \frac{1}{\pi}\sin\left(\frac{\pi(n-1)}{2n}\right)$ for n odd.

The barcodes corresponding to the persistent homology of this Vietoris–Rips filtration in dimensions $0, 1$ and 2 are thus as in Figure A.6.

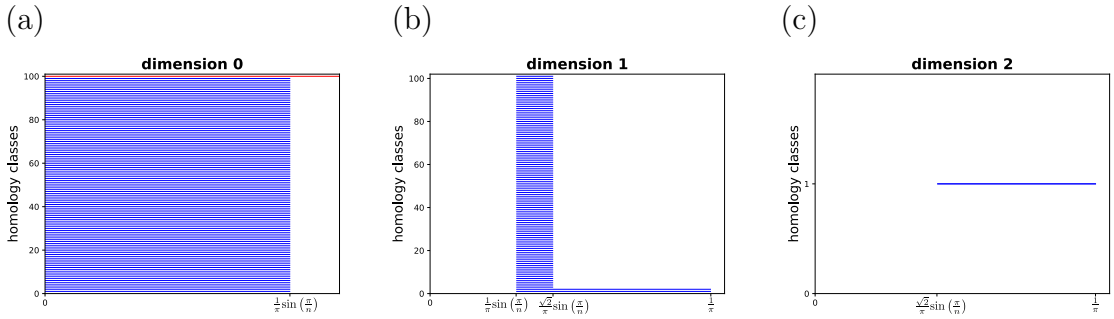


Figure A.6: Barcodes for the persistent homology in dimension (a) 0, (b) 1, and (c) 2 of the Vietoris–Rips filtration according to the $d_{2,2}$-metric. (The red bar in (a) never dies.)

A.3 The $d_{\text{arc},1}$ metric

The metric on the torus used in this section is again a product metric on $\mathbb{T} = \frac{1}{2\pi}\mathbb{S}^1 \times \frac{1}{2\pi}\mathbb{S}^1$, but this time the two circles, the Cartesian product of which forms the torus, are not considered as subspaces of copies of $\mathbb{C}$, inheriting its Euclidean metric. Instead, we denote points on $\mathbb{S}^1$ by the angle at which they lie on the circle and take the distance between two points on the circle to be the arc length between them. That is, points on $\frac{1}{2\pi}\mathbb{S}^1$ are $\{\theta \mid 0 \leq \theta < 2\pi\}$ and the distance between two points $\theta_1, \theta_2 \in \frac{1}{2\pi}\mathbb{S}^1$ is

$$d_{\text{arc}}(\theta_1, \theta_2) = \frac{1}{2\pi}|\theta_1 - \theta_2|_{\text{per}} \ .$$

where

$$|\theta|_{\text{per}} = \min\{\phi \in \mathbb{R}_{\geq 0} \mid \phi = 2\pi k + \theta \ \text{ or } \ \phi = 2\pi k - \theta \text{ for some } k \in \mathbb{Z}\} \, .$$

We then take the metric on the torus $\mathbb{T} = \frac{1}{2\pi}\mathbb{S}^1 \times \frac{1}{2\pi}\mathbb{S}^1$ to be the d_1 product metric and denote it by $d_{\text{arc},1}$. That is, the distance between two points

$$p_1 = (\theta_1, \phi_1), p_2 = (\theta_2, \phi_2) \in \mathbb{T}$$

is

$$d_{\text{arc},1}(p_1, p_2) = \frac{1}{2\pi}\left(|\theta_1 - \theta_2|_{\text{per}} + |\phi_1 - \phi_2|_{\text{per}}\right) .$$

Notice that this metric is really a 'continuous version' of the *periodic lattice distance*, i.e. the lattice distance between points on a regular lattice where opposite sides are thought to be adjacent (see Section 3.2.1). Also note that the d_2-metric and the d_{arc} are related by $d_2 = 2\sin\left(\frac{1}{2}d_{\text{arc}}\right)$.

Remark: Usually Vietoris–Rips fitrations are built on point clouds that are thought to be sampled from spaces that are proper subspaces of $\mathbb{R}^p$ for some $p \in \mathbb{N}$. That is, usually one builds a Vietoris–Rips fitration in an ambient space that contains

points that are not in the space under consideration, and usually the ambient space is a p-dimensional Euclidean space for some $p \in \mathbb{N}$. We have already worked with an example where the ambient space of the point cloud is not Euclidean (in Section A.1 on the $d_{2,1}$ metric). We have not yet considered a scenario where the torus $\mathbb{T}$ is not assumed to lie in some ambient space whose metric it inherits and that contains points that are not on $\mathbb{T}$. In a sense, the metric $d_{\text{arc},1}$ provides such a scenario, forming an intrinsic metric on $\mathbb{T}$. However, we will see that, while $d_{\text{arc},1}$ is an intrinsic metric, its associated Vietoris–Rips filtration does not recover the intrinsic topology of $\mathbb{T}$, which would entail homology classes in dimensions 1 and 2 that live forever. The reason for this is that the Nerve Lemma (2.4) does not hold in this case, as unions of balls in this metric can have non-contractible intersections.

Choosing $N = n^2$ points to be evenly distributed on $\mathbb{T}$ according to the $d_{\text{arc},1}$ gives the cloud of points of coordinates

$$\left(\frac{2\pi i}{n}, \frac{2\pi j}{n}\right)$$

for $i, j \in \{0, 1 \dots n-1\}$. The distance between two of these points is exactly $\frac{1}{n}$ of the periodic lattice distance between the two corresponding points on a lattice of side length 1. We now construct a Vietoris–Rips fitration around these points according to the $d_{\text{arc},1}$ metric and consider its persistent homology. All components merge to one and $n^2 + 1$ one-dimensional loops appear at $t = \frac{1}{n}$. All but two one-dimensional loops disappear and a two-dimensional void appears at $t = \frac{2}{n}$. All non-trivial features disappear at $t = \frac{1}{2}$ for n even, and $t = \frac{n-1}{2n}$ for n odd. The barcodes corresponding to the persistent homology of this Vietoris–Rips filtration in dimensions $0, 1$ and 2 are thus as in Figure A.7.

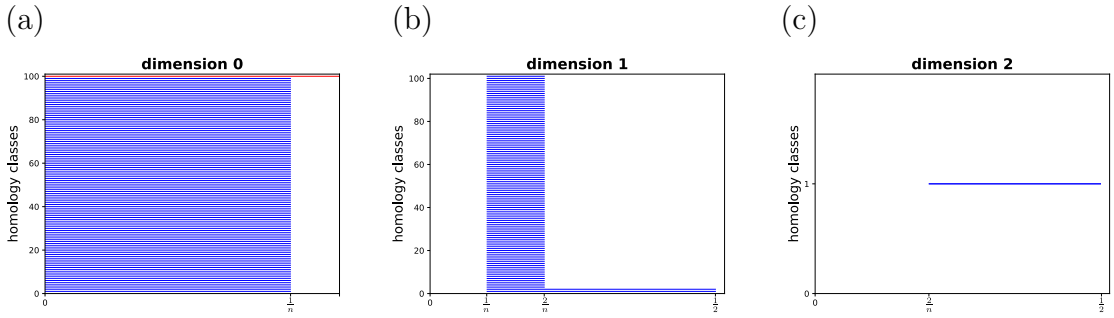


Figure A.7: Barcodes for the persistent homology in dimension (a) 0, (b) 1, and (c) 2 of the Vietoris–Rips filtration according to the $d_{\mathrm{arc},1}$-metric. (The red bar in (a) never dies.)

A.4 The $d_{\mathrm{arc},2}$ metric

The $d_{\mathrm{arc},2}$ metric is very similar to the $d_{\mathrm{arc},1}$ metric with the only difference being that it is the d_2 product metric on $\mathbb{T} = \frac{1}{2\pi}S^1 \times \frac{1}{2\pi}S^1$ induced by the arc length on $\frac{1}{2\pi}S^1$ rather than the d_1 product metric. This metric is basically the metric on a square of side length 1 with sides appropriately identified to form a torus. Notice that this is again a scenario where a Vietoris–Rips filtration is built on a torus $\mathbb{T}$ that does not lie in some ambient space whose metric it inherits and which contains points that are not on $\mathbb{T}$. The statements about the intrinsic nature of $d_{\mathrm{arc},1}$ in the remark in the previous sections also apply to $d_{\mathrm{arc},2}$.

A cloud of $N = n^2$ points evenly distributed according to the $d_{\mathrm{arc},2}$ metric is, just as according to the $d_{\mathrm{arc},1}$ the set of points with coordinates

$$\left(\frac{2\pi i}{n}, \frac{2\pi j}{n}\right)$$

for $i, j \in \{0, 1 \dots n-1\}$. We now construct a Vietoris–Rips filtration on this point cloud according to the $d_{\mathrm{arc},2}$ metric and consider its persistent homology. All components merge to one and $n^2 + 1$ one-dimensional loops appear at $t = \frac{1}{n}$. All but two one-dimensional loops disappear and a two-dimensional void appears at $t = \frac{\sqrt{2}}{n}$. All non-trivial features disappear at $\frac{1}{2}$ for n even, and $\frac{n-1}{2n}$ for n odd. The barcodes corresponding to the persistent homology of this Vietoris–Rips filtration in dimensions

$0, 1$ and 2 are thus as in Figure A.8.

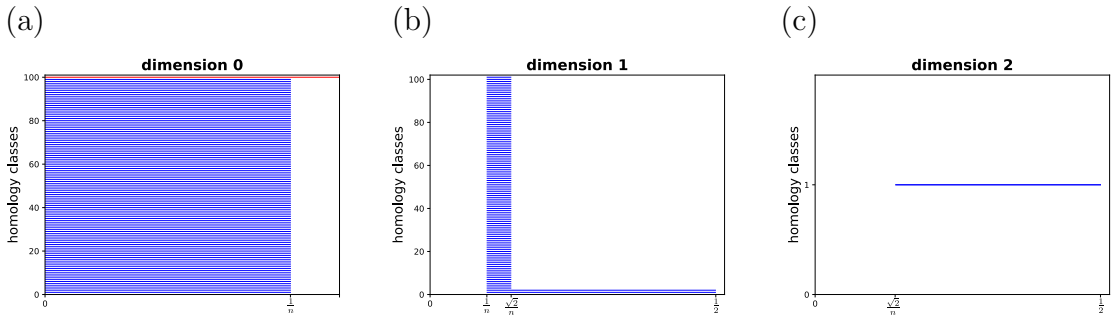


Figure A.8: Barcodes for the persistent homology in dimension (a) 0, (b) 1, and (c) 2 of the Vietoris–Rips filtration according to the $d_{\mathrm{arc},2}$-metric. (The red bar in (a) never dies.)

Remark: Notice that, with all four metrics $d_{2,1}, d_{2,2}, d_{\mathrm{arc},a}$ and $d_{\mathrm{arc},2}$, the more points the point cloud contains (i.e. the greater $N = n^2$ is) the more pronounced the toroidal features of the Vietoris–Rips filtration are (the shorter the noisy bars and the longer the persistent bars in the associated barcodes). In particular, as $n \to \infty$, the lengths of the noisy bars tend to 0, and the birth times of the bars representing the toroidal features tend to 0 (while their death times are independent of n).

The figures in this appendix represent the case when $n = 10$ (i.e. when the point cloud has 100 points).

Bibliography

[1] D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, and A. Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21484–21489, 2009.

[2] M. Barthelemy. Spatial networks. *Physics Reports*, 499(1-3):1–101, 2011.

[3] M. Barthelemy. *Morphogenesis of Spatial Networks*. Springer International Publishing, Cham, Switzerland, 2018.

[4] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems*, pages 585–591, 2002.

[5] R. Bellman. *Dynamic Programming*. Rand Corporation research study. Princeton University Press, 1957.

[6] M. Bernstein, V. de Silva, J. C. Langford, and Tenenbaum J. B. Graph approximations to geodesics on embedded manifolds. *Technical report, Department of Psychology, Stanford University*, 2000.

[7] M. Boguñá, F. Papadopoulos, and D. Krioukov. Sustaining the Internet with hyperbolic mapping. *Nature Communications*, 1(6):62, 2010.

[8] D. Brockmann and D. Helbing. The hidden geometry of complex, network-driven contagion phenomena. *Science*, 342(6164):1337–1342, 2013.

[9] J. Brown and T. Gedon. Structure of the afferent terminals in terminal ganglion of a cricket and persistent homology. *PLoS ONE*, 7, 2012.

[10] W. M. Brown, S. Martin, S. N. Pollock, E. A. Coutsias, and J.-P. Watson. Algorithmic dimensionality reduction for molecular structure analysis. *The Journal of chemical physics*, 129(6):064118, 2008.

[11] Scott J. Bubenik P., de Silva V. Metrics for generalized persistence modules. *Found Comput Math*, 15:1501–1531, 2014.

[12] Scott J. A. Bubenik P. Categorification of persistent homology. *Discrete Comput Geom*, 51:600–627, 2014.

[13] F. Cagliari, M. Ferri, and P. Pozzi. Size functions from a categorical viewpoint. *Acta Applicandae Matematicae*, 67:225–235, 2001.

[14] G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[15] D. Centola. The spread of behavior in an online social network experiment. 329(September):1194–1198, 2010.

[16] D. Centola, M. W. Macy, and V. M. Eguíluz. Cascade dynamics of complex propagation. *Physica A*, 374(1):449–456, 2007.

[17] J. Chalupa, P. L. Leath, and G. R. Reich. Bootstrap percolation on a Bethe lattice. *Journal of Physics C: Solid State Physics*, 12(1):L31–L35, 1979.

[18] F. Chazal, D. Cohen-Steiner, M. Glisse, L. J. Guibas, and Oudot S. Y. Proximity of persistence modules and their diagrams. *Proceedings of the twenty-fifth annual*

symposium on computational geometry, SoCG 09. ACM, New York, pages 237–246, 2009.

[19] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete Comput Geom*, 37:103–120, 2007.

[20] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences of the United States of Americal*, 102(21):7426–7431, 2005.

[21] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

[22] D. J. Corsi, M. H. Boyle, S. A. Lear, C. K. Chow, K. K. Teo, and S. V. Subramanian. Trends in smoking in Canada from 1950 to 2011: Progression of the tobacco epidemic according to socioeconomic status and geography. *Cancer Causes and Control*, 25(1):45–57, 2014.

[23] M. A. Cox, T. F. & Cox. *Multidimensional Scaling*. CRC Press, Boca Raton, FL, USA, 2010.

[24] V. De Silva and R. Ghrist. Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology*, 7(1):339–358, 2007.

[25] P. G. Doyle and J. L. Snell. Random walks and electric networks. *arXiv preprint math/0001057*, 2000.

[26] R. Ebrahimi, J. Gao, G. Ghasemiesfeh, and G. Schoenebeck. Complex Contagions in Kleinberg's Small World Model. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 63–72. ACM, 2015.

[27] H. Edelsbrunner and J. L. Harer. Persistent homology — A survey. In E. G. Jacob, P. János, and R. Pollak, editors, *Surveys on Discrete and Computational Geometry. Twenty years later*, volume 453 of *Contemporary Mathematics*, pages 257–282. American Mathematical Society, 2008.

[28] H. Edelsbrunner and J. L. Harer. *Computational Topology: An Introduction.* American Mathematical Society, Providence, RI, USA, 2010.

[29] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel. On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29:551–559, 1983.

[30] D. J. Elzinga and D. W. Hearn. The minimum covering sphere problem. *Management Science*, 19(1):96–104, 1972.

[31] D. J. Fellemann and D. C. Van Essen. Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cereb Cortex*, 1:1–47, 1991.

[32] R. W. Floyd. Algorithm 97: shortest path. *Communications of the ACM*, 5(6):345, 1962.

[33] P. Frosini and C. Landi. Size functions from a categorical viewpoint. *Pattern Recognition and Image Analysis*, 9:596–603, 1999.

[34] S. Gerber, T. Tasdizen, and R. Whitaker. Robust non-linear dimensionality reduction using successive 1-dimensional Laplacian Eigenmaps. *Proceedings of the 24th international conference on Machine learning — ICML '07*, pages 281–288, 2007.

[35] G. Ghasemiesfeh, R. Ebrahimi, and J. Gao. Complex contagion and the weakness of long ties in social networks: revisited. *Proceedings of the fourteenth ACM Conference on Electronic Commerce*, 1(212):507–524, 2013.

[36] R. Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

[37] M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443., 1978.

[38] G. H. Hardy. On the expression of a number as the sum of two squares. *Quarterly Journal of Mathematics*, 46:263–283, 1915.

[39] A. Hatcher. Algebraic topology. 2001.

[40] P. Hedström. Contagious collectivities: On the spatial diffusion of Swedish trade unions, 1890-1940. *American Journal of Sociology*, 99(5):1157–1179, 1994.

[41] J. B. Hendrickson. Molecular geometry. v. evaluation of functions and conformations of medium rings. *Journal of the American Chemical Society*, 89(26):7036–7043, 1967.

[42] P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

[43] M. Kerber, D. Morozov, and A. Nigmetov. Geometry helps to compare persistence diagrams. 22(1):1–20, 2017.

[44] I. Z. Kiss, J. C. Miller, and P. L. Simon. *Mathematics of Epidemics on Networks: From Exact to Approximate Models*. Springer-Verlag, Berlin, Germany, 2017.

[45] J. M. Kleinberg. Navigation in a small world. *Nature*, 406(6798):845, 2000.

[46] J. M. Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, pages 163–170. ACM, 2000.

[47] J. A. Lee and M. Verleysen. *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.

[48] Y. Lee, S. D. Barthel, P. Dłotko, S. M. Moosavi, K. Hess, and B. Smit. Quantifying similarity of pore-geometry in nanoporous materials. *Nature Communications*, 8:15396, 2017.

[49] location = Berlin, Germany Year = 2018 Lehmann, S. and Ahn, Y.-Y., Publisher = Springer-Verlag. *Complex Spreading Phenomena in Social Systems*.

[50] K. V. Mardia, J. T. Kent, and J. M. Bibby. Multivariate analysis. *Probability and mathematical statistics. Academic Press Inc*, 1979.

[51] S. Martin, A. Thompson, E. A. Coutsias, and J.-P. Watson. Topology of cyclooctane energy landscape. *The Journal of Chemical Physics*, 132(23):234115, 2010.

[52] S. Melnik, J. A. Ward, J. P. Gleeson, and M. A. Porter. Multi-stage complex contagions. *Chaos*, 23(1):013124, 2013.

[53] G. P. Moss. Basic terminology of stereochemistry (iupac recommendations 1996). *Pure and applied chemistry*, 68(12):2193–2222, 1996.

[54] M. E. J. Newman. Models of the small world. *Journal of Statistical Physics*, 101(3-4):819–841, 2000.

[55] M. E. J. Newman. *Networks*. Oxford University Press, second edition, 2018.

[56] M. Nicolau, A. J. Levine, and G. Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108:7265–7270, 2011.

[57] N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, and H. A. Harrington. A roadmap for the computation of persistent homology. *European Physical Journal — Data Science*, 2017.

[58] P. W. Pakes, T. C. Rounds, and H. L. Strauss. Conformations of cyclooctane and some related oxocanes. *The Journal of Physical Chemistry*, 85(17):2469–2475, 1981.

[59] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani. Epidemic processes in complex networks. *Reviews of Modern Physics*, 87(3):925–979, 2015.

[60] M. A. Porter. Small-world network. *Scholarpedia*, 7(2):1739, 2012.

[61] M. A. Porter and J. P. Gleeson. *Dynamical Systems on Networks: A Tutorial*. Frontiers in Applied Dynamical Systems: Reviews and Tutorials. Springer International Publishing, 4 edition, 2016.

[62] C. J. Rhodes and R. M. Anderson. Epidemic thresholds and vaccination in a lattice model of disease spread. *Theoretical Population Biology*, 52:101–118, 1997.

[63] V. Robins. Toward computing homology from finite approximations. *Topology Proceedings*, 24:503–532, 1999.

[64] E. M. Rogers. *Diffusion of Innovations*. Simon and Schuster, New York City, NY, USA, 2010.

[65] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[66] M. Serrano, M. Boguñá, and F. Sagués. Uncovering the hidden geometry behind metabolic networks. *Molecular BioSystems*, 8:843–850, 2012.

[67] G. W. Shannon, G. F. Pyle, and R. L. Bashshur. The geography of AIDS: Origins and course of an epidemic. 1991.

[68] G. Singh, F. Memoli, T. Ishkhanov, G. Sapiro, G. Carlsson, and D. Ringach. Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Journal of Vision*, 8, 2008.

[69] C. O. S. Sorzano, J. Vargas, and A. P. Montano. A survey of dimensionality reduction techniques. *http://arxiv.org/abs/1403.2877*, pages 1–35, 2014.

[70] A. Tausz, M. Vejdemo-Johansson, and H. Adams. JavaPlex: A research software package for persistent (co)homology. In Han Hong and Chee Yap, editors, *Proceedings of ICMS 2014*, Lecture Notes in Computer Science 8592, pages 129–136, 2014. Software available at `http://appliedtopology.github.io/javaplex/`.

[71] 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. *Nature Communications*, 6:7723, 2015.

[72] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[73] W. S. Torgerson. Theory and methods of scaling. 1958.

[74] T. W. Valente. Network models of the diffusion of innovations. *Computational & Mathematical Organization Theory*, 2(2):163–164, 1996.

[75] S. Warshall. A theorem on boolean matrices. In *Journal of the ACM*. Citeseer, 1962.

[76] D. J. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9):5766–5771, 2002.

[77] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[78] K. Q. Weinberger and L. K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *International journal of computer vision*, 70(1):77–90, 2006.

[79] Wikimedia. Ball-and-stick model of the cyclooctane molecule, 2011. Files: Cyclooctane-boat-chair-3D-balls.png, Cyclooctane-crown-3D-balls.png.

[80] X. J. Xu, X. Zhang, and J. F. F. Mendes. Impacts of preference and geography on epidemic spreading. *Physical Review E*, 76(5):1–4, 2007.

[81] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete and Computational Geometry*, 33:249–274, 2005.