http://topology.auburn.edu/tp/

TOPOLOGY
PROCEEDINGS

Volume 56, 2020

Pages 1–24

http://topology.nipissingu.ca/tp/

Persistent Homology and Random Models of the Gaussian Primes

by

Carl Hammarsten, Rae Helmreich, Anchala Krishnan, John Meier, and Nathan Schmitz

Electronically published on May 30, 2019

Topology Proceedings

Web: http://topology.auburn.edu/tp/
Mail: Topology Proceedings
Department of Mathematics & Statistics
Auburn University, Alabama 36849, USA
E-mail: topolog@auburn.edu
ISSN: (Online) 2331-1290, (Print) 0146-4124

http://topology.auburn.edu/tp/
http://topology.nipissingu.ca/tp/

TOPOLOGY PROCEEDINGS
Volume 56 (2020)
Pages 1-24

E-Published on May 30, 2019

PERSISTENT HOMOLOGY AND RANDOM MODELS OF THE GAUSSIAN PRIMES

CARL HAMMARSTEN, RAE HELMREICH, ANCHALA KRISHNAN, JOHN MEIER, AND NATHAN SCHMITZ

Abstract. Ilan Vardi introduced a probabilistic model for the distribution of Gaussian primes in 1998 (see *Prime percolation*, Experiment. Math. **7** (1998), no. 3, 275–289). We use persistent homology to test how well Vardi's model and a symmetrized variant of Vardi's model capture the geometry of the Gaussian primes. An analysis of the persistent first homology of point clouds produced by these models provides statistical evidence that the models miss some fundamental geometric features of the Gaussian primes.

1. Introduction

The Gaussian primes are the irreducible elements in the Gaussian integers (the ring $\mathbb{Z}[i]$). There are elementary, open questions about the geometry of the Gaussian primes thought of as a point cloud in $\mathbb{R}^2$. In 1962 at the International Congress of Mathematicians in Stockholm, Basil Gordon asked if you can "walk to infinity" stepping only on Gaussian primes. Said a bit more formally, for any positive real number D, define $\mathcal{G}_D$ to be the graph whose vertices consist of Gaussian primes and whose edges are the pairs of primes $\{p, q\}$ where $d(p, q) \leq D$. (The case when

2010 *Mathematics Subject Classification.* Primary: 57Q05; Secondary: 62F03.
Key words and phrases. Gaussian primes, persistent homology, Vietoris–Rips complexes.
This project has been supported by NSF Grant #1560222, REU Site: Undergraduate Research in Mathematics, Applied Mathematics, and Statistics at Lafayette College.

$D = \sqrt{5}$ is shown in Figure 1.) Gordon's question asks if there is a $D > 0$ such that the connected component of $\mathcal{G}_D$ containing $1 + \mathrm{i}$ is unbounded; a common variant asks if any connected component is unbounded. It is known that $D = \sqrt{26}$ is not sufficient [7]. We note that, just as there are arbitrarily wide gaps between adjacent primes in $\mathbb{Z}$, there are regions in $\mathbb{R}^2$ of arbitrarily large diameter where no Gaussian primes are present, but this does not resolve Gordon's question as it may be possible to walk around these large, prime-free zones.

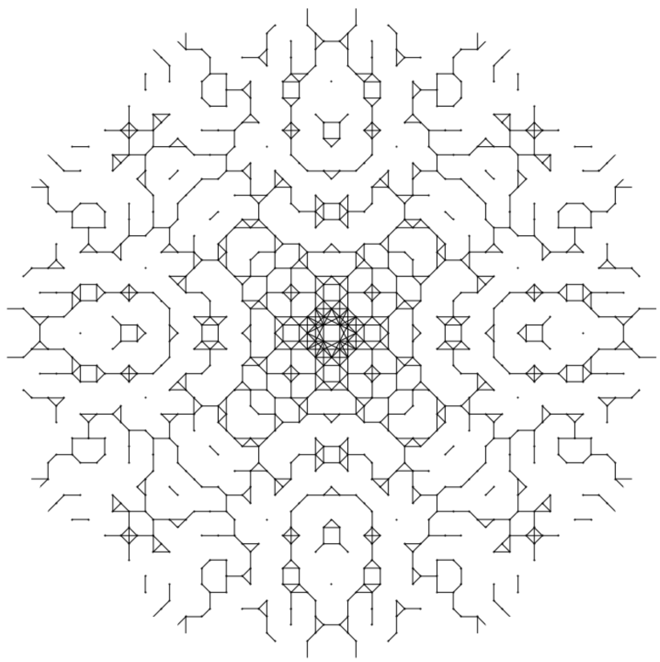


FIGURE 1. The graph $\mathcal{G}_{\sqrt{5}}$ joins Gaussian primes if they are at most a distance $\sqrt{5}$ apart. This figure includes the Gaussian primes where $|z| \le 50$.

Ilan Vardi [15] examined Gordon's question using percolation theory, introducing a probabilistic model for the distribution of Gaussian primes. His random model is inspired by Cramér's classical model of the integral primes, which lead to Cramér's Conjecture that

$$p_{n+1} - p_n = \mathcal{O}((\log p_n)^2)$$

where p_n denotes the n^{th} integral prime. Using an approach similar to Cramér's, Vardi proves a result about the size of prime-free regions in his random models, leading to the following conjecture.

Conjecture 1.1 ([15, Proposition 4.1]). *Let $d(R)$ be the radius of the largest Gaussian prime free zone a distance at most R from the origin in $\mathbb{R}^2$. Then*

$$\limsup_{R\to\infty} \frac{d(R)}{\log(R)} = 1.$$

It is natural to explore the questions of Gordon and Vardi via the tools of topological data analysis, particularly persistent homology. Persistent homology is a tool for detecting topological and geometric features of a finite point cloud in $\mathbb{R}^n$. For example, consider the restriction of $\mathcal{G}_D$ (the graph with vertex set the Gaussian primes and edges joining those that are at most a distance D apart) to the Gaussian primes where $|z| \leq R$ for some fixed $R > 0$ (as in Figure 1). If we consider the vertices and edges of $\mathcal{G}_D$ as 0- and 1-dimensional simplices, respectively, we may then extend $\mathcal{G}_D$ to a simplicial complex $\hat{\mathcal{G}}_D$ by adding a k-simplex ($k > 1$) for every complete subgraph on k vertices. Persistent homology then measures how the homology groups $H_i(\hat{\mathcal{G}}_D)$ change as one varies the parameter D. Since $H_0(\hat{\mathcal{G}}_D)$ measures the number of connected components in the complex, it is directly related to Gordon's question on the existence of an unbounded component in $\mathcal{G}_D$. Similarly, given a prime free region in $\mathbb{R}^2$, it is likely that there are values of the parameter D such that a homologically non-trivial cycle of edges in $\hat{\mathcal{G}}_D$ bounds that prime free region. Examples of such 1-cycles can be seen in Figure 1. Further, as we explain in more detail later, the larger the prime free region is, the more persistent this homology class will be. Thus, Vardi's conjecture is related to the persistence of homology classes in $H_1(\hat{\mathcal{G}}_D)$.

In this paper, we use the tools of persistent homology to compare the point cloud consisting of the Gaussian primes within a bounded region with point clouds produced by Vardi's random model of the Gaussian primes in the same region. We also compare the Gaussian primes with point clouds produced by a symmetric version of Vardi's random model. A statistical analysis of these comparisons gives evidence that neither of the two models accurately captures the geometry of the Gaussian primes.

We present definitions and background information on Gaussian primes, Vardi's model, and persistent homology in §2 and §3. Our statistical analysis, comparing the persistent homology associated to the Gaussian primes with that associated to point clouds produced by each of the models, is presented in §4.

2. Modeling the Gaussian Primes

We briefly recount some key characteristics of the Gaussian primes before introducing Vardi's random model for the distribution of Gaussian primes.

2.1. The Gaussian primes.

A complex number $\alpha = a + b\mathrm{i}$ is a *Gaussian integer* if $a, b \in \mathbb{Z}$. The irreducible Gaussian integers—those that cannot be written as a non-trivial product of Gaussian integers—are the *Gaussian primes*. The characterization of Gaussian primes stated below is a key result in elementary number theory and proofs can be readily found, including in [9].

Theorem 2.1. *Let $\alpha = a + b\mathrm{i} \in \mathbb{Z}[\mathrm{i}]$. Then α is a Gaussian prime if and only if one of the following is true.*

1. *If a or $b = 0$, $|\alpha| \equiv 3 \mod 4$, and $|\alpha|$ is prime in $\mathbb{Z}$.*
2. *Otherwise, α is a Gaussian prime if and only if $a^2 + b^2$ is prime in $\mathbb{Z}$.*

We note that this characterization implies that the Gaussian primes have eight-fold dihedral symmetry. If $a + b\mathrm{i}$ is a Gaussian prime, then so are $\pm a \pm b\mathrm{i}$ and $\pm b \pm a\mathrm{i}$. Thus, the set of Gaussian primes is preserved by reflections across the real and imaginary axes, as well as across the diagonals. These transformations generate D_4, the dihedral group of order 8.

Corollary 2.2. *The Gaussian primes are preserved under the action of D_4, generated by the reflection across the real axis and the diagonal $y = x$.*

2.2. The Vardi–Cramér model.

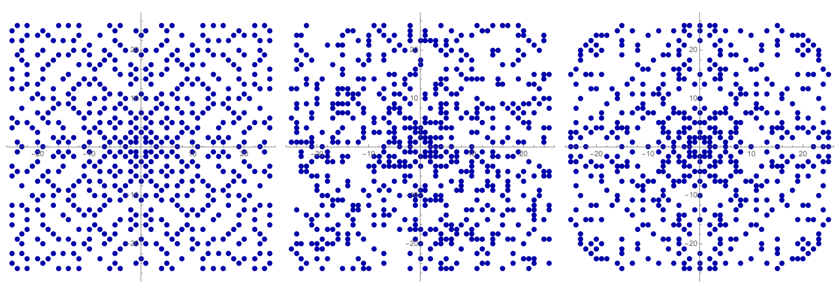
Vardi's random model for the Gaussian primes is based on Cramér's random model for the integral primes. Cramér's model is a random selection of positive integers where the natural number n is included with probability $\frac{1}{\log(n)}$. In other words, Cramér's model is governed by the prime number theorem. The characterization of Gaussian primes combined with the prime number theorem determines the density of the Gaussian primes among the Gaussian integers. Using this, Vardi introduces a random model of the Gaussian primes where each Gaussian integer $z = a + b\mathrm{i}$ is included in the model with probability $\frac{2}{\pi \log |z|}$. (For the details, see [15].) When we need to distinguish this random model from the next, we call this model the *basic* Vardi–Cramér model.

2.3. The symmetric Vardi–Cramér model.

One weakness in the basic Vardi–Cramér model is that it does not generally enforce the eightfold symmetry of the Gaussian primes; it only accounts for their density. By Corollary 2.2, one can produce all Gaussian primes by first starting with those that are in the $\frac{\pi}{4}$-sector $\{z = a + b\mathrm{i} \mid 0 \leq b \leq a\}$ and applying the action of D_4. We introduce this dihedral symmetry to the basic Vardi–Cramér model. The *symmetric* Vardi–Cramér model begins by including each Gaussian integer $z = a+b\mathrm{i}$, where $0 \leq b \leq a$, with probability $\frac{2}{\pi \log |z|}$. Then one applies the D_4 action to these chosen points to create a point cloud in the plane with D_4 symmetry.

We expect point clouds created under the symmetric model will have better correlation to the genuine point cloud of primes because any homological feature arising from the random selection of points in the $\pi/4$-sector will be replicated eight times under the D_4 symmetry. Furthermore, there may also be features that arise when the $\pi/4$-sector is reflected over $y = x$ or the x-axis; the D_4 symmetry then replicates these features four times. It follows that homological features will frequently occur in quartets or octets, similar to how they likely appear in the Gaussian primes due to the same D_4 symmetry.

In Figure 2 we show the distribution of the Gaussian primes and compare it with example point clouds produced by the basic and symmetric Vardi–Cramér (V–C) models.



The Gaussian Primes Basic V–C Model Symmetric V–C Model

Figure 2. The Gaussian primes, a point cloud created using the basic Vardi–Cramér model, and a point cloud created using the symmetric Vardi–Cramér model are shown. Each is contained in the square with corners $\pm 25 \pm 25\mathrm{i}$.

3. Persistent Homology

Our presentation of persistent homology is terse and focused on the main concepts that we rely upon in our analysis of the Vardi–Cramér models of the Gaussian primes. In particular, we assume the reader has a basic familiarity with simplicial homology, at the level of [10]. While we do not always explicitly mention it, we are computing homology over the real numbers. The fundamental theory underlying persistent homology requires field coefficients.

A short introduction to persistent homology, that does not include persistence diagrams, can be found in [8]. The reader wanting an accessible and more detailed introduction to persistent homology should consult [5].

3.1. Vietoris–Rips complexes.

Given a point cloud $X \subset \mathbb{R}^n$ and a real number $\epsilon \geq 0$, the *Vietoris–Rips complex* is the simplicial complex $VR_\epsilon(X)$ that is obtained by including a k-simplex for every subset of $(k+1)$ points in X where the distance between any two is at most 2ϵ. As is shown in Figure 3, for a given point cloud X and $\epsilon \geq 0$, one can imagine the collection of balls of radius ϵ about each $x \in X$. In this picture, the simplices of the Vietoris–Rips complex $VR_\epsilon(X)$ are then subsets $\sigma \subset X$ where, for each pair of points $x_i, x_j \in \sigma$, the ϵ-balls centered at x_i and x_j have non-empty intersection.

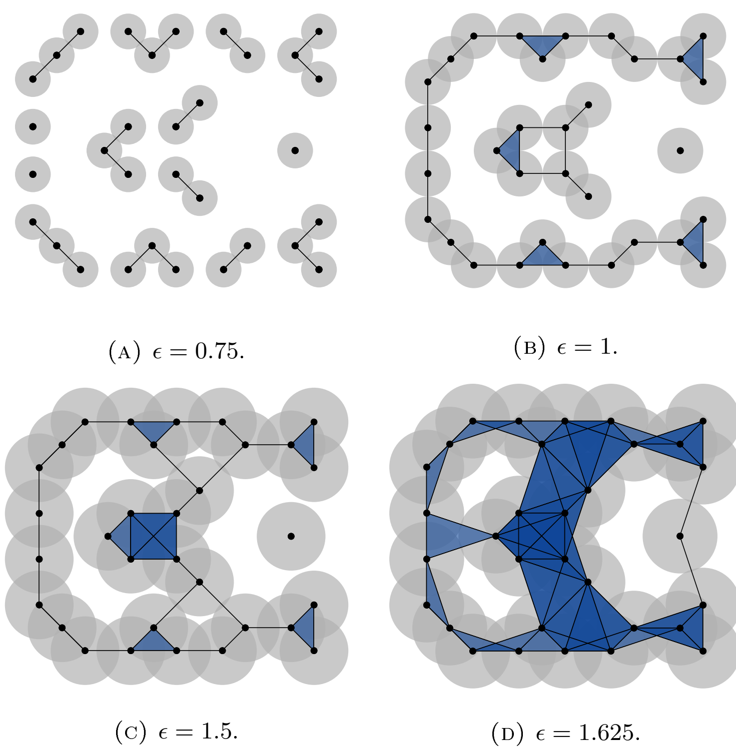
The sequence of Vietoris–Rips complexes shown in Figure 3 illustrates what happens as the value of ϵ increases. The ground set for all four figures are the Gaussian primes contained in the rectangle with corners at $20 \pm 5\mathrm{i}$ and $32 \pm 5\mathrm{i}$. We note that because we are working with a finite set of points in the plane, we can compute the rank of $H_0(VR_\epsilon(X))$ and $H_1(VR_\epsilon(X))$ just by examining the figures above and determining the number of connected components and visible holes formed by the planar projection of $VR_\epsilon(X)$ suggested by these pictures (see [4]).

(A) In the first figure, $\epsilon = 0.75$. Only a handful of primes that are a distance $\sqrt{2}$ apart are connected by edges. No 1-dimensional homology classes have been formed.

(B) In the second figure, $\epsilon = 1$. There are fewer connected components and a cycle of 4 edges gives rise to a non-trivial first homology class.

(C) In the third figure, $\epsilon = 1.5$. Here there are only two connected components and the cycle of 4 edges from the previous complex now bounds a 2-complex. There are also new edge cycles that give non-trivial homology classes in dimension 1.

(D) In the fourth figure, $\epsilon = 1.625$. Here we see quite a few higher dimensional simplices and we now have a single connected component. Again, some 1-cycles from the previous complex are now boundaries and there are also some new 1-cycles that are not boundaries. Interestingly, we also see that the leftmost 1-cycle from the previous complex has been split into the sum of two distinct non-trivial 1-cycles.



(A) $\epsilon = 0.75$. (B) $\epsilon = 1$.

(C) $\epsilon = 1.5$. (D) $\epsilon = 1.625$.

FIGURE 3. Four Vietoris–Rips complexes. Each vertex set consists of the Gaussian primes in the rectangle with corners at $20 \pm 5\mathrm{i}$ and $32 \pm 5\mathrm{i}$.

The Betti numbers of a simplicial complex K are the ranks of its homology groups:

$$b_i(K) = \operatorname{rank}\{H_i(K, \mathbb{R})\}.$$

The values of b_0 and b_1 for these four Vietoris–Rips complexes are shown in the table below.

TABLE 1. The values of b_0 and b_1 in four Vietoris–Rips complexes as shown in Figure 3.

ϵ	$b_0(VR_\epsilon(X))$	$b_1(VR_\epsilon(X))$
0.750	14	0
1.000	3	1
1.500	2	3
1.625	1	3

Of course, for any finite set $X \subset \mathbb{R}^n$, $VR_0(X) = X$ and the only non-zero Betti number is $b_0 = \#(X)$. If D is the diameter of X, then $VR_D(X)$ is a $(\#(X) - 1)$-simplex and has no non-trivial homological features. The interesting observation is thus to track how the homology of these complexes evolves as ϵ varies between these two extremes.

3.2. Persistence diagrams and bottleneck distance.

The persistence diagrams for the Vietoris–Rips complexes $VR_\epsilon(X)$ provide a summary of how the homology of these complexes evolves as ϵ increases. If $\alpha < \omega$, then $VR_\alpha(X) \subseteq VR_\omega(X)$, and so there is an induced homomorphism

$$H_k(VR_\alpha(X)) \to H_k(VR_\omega(X)).$$

This kind of map allows us to keep track of the homology class associated with a k-cycle as the index ϵ increases. While ϵ is defined as a unit of distance, it is common to think of ϵ simultaneously as a unit of time and to discuss homological features present at time $t = \epsilon$. In particular, we can associate a *birth-death pair* (α, ω) with each homology class, where $\epsilon = \alpha$ is the first time the homology class appears and $\epsilon = \omega$ is the time it becomes a boundary of a $(k + 1)$-chain.

The Vietoris–Rips complexes $VR_\epsilon(X)$ induce a directed system of simplicial chain complexes. The structure theorem for the homology of the resulting filtered chain complex is determined by the birth-death pairs of a generating set of chains (see [16]). The *k-dimensional persistence diagram* consists of all the birth-death pairs for these generators (counted with multiplicity when distinct homology classes share the same lifespan) in dimension k. It is commonly displayed as a collection of points in the first quadrant, each of which lies on or above the line $y = x$, since every homology class must be born before it can die. There is a direct connection between the Betti numbers we mentioned above and the persistence diagrams. The Betti number $b_k(VR_\epsilon(X))$ is the rank of $H_k(VR_\epsilon(X), \mathbb{R})$, which is generated by k-cycles that have been born by time ϵ and have not yet died. That is, $b_k(VR_\epsilon(X))$ is the number of points (x, y) in the k-dimensional persistence diagram where $x \leq \epsilon$ and $y > \epsilon$.

In Figure 4 we show the 0- and 1-dimensional persistence diagrams for the point cloud X, the subset of Gaussian primes used in Figure 3. There are over thirty 0-dimensional homology classes. There are far fewer points shown as all these classes are born when $\epsilon = 0$ and all but one share their time of death with other points in X. Note also that, following standard conventions, we plot a point at $(0, 3)$ in the 0-dimensional persistence diagram. This corresponds to the fact that, for all $\epsilon > D$ (the diameter of X), there will be a single 0-dimensional feature that never dies. That is, the full $(\#(X) - 1)$-simplex always has a single generator in 0-dimensional homology. The 1-dimensional persistent homology has six homology classes. However, only five points are visible in Figure 3 because two of these features share the same birth and death times. In contrast to the 0-dimensional case, here the full $(\#(X) - 1)$-simplex has no 1-dimensional homology. So everything of interest is completely and accurately indicated in the persistence diagram.

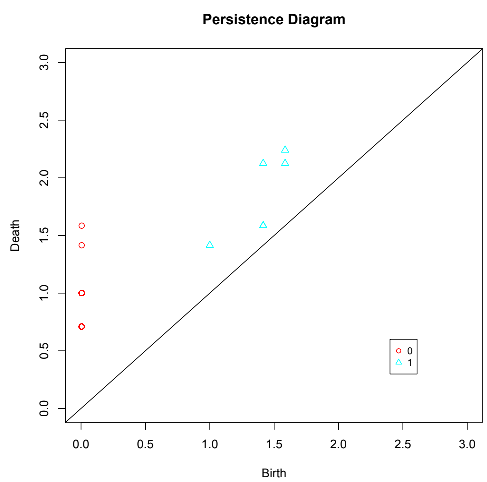


FIGURE 4. The 0- and 1-dimensional persistence diagrams for the subset of Gaussian primes considered in Figure 3.

Because the questions of Gordon and Vardi are connected to the 0- and 1-dimensional homology of Vietoris–Rips complexes associated to the Gaussian primes, we focus on 0- and 1-dimensional persistence diagrams.

We note that, in general, 0-dimensional persistence diagrams are somewhat boring, as all the connected components are born at the same time, $\epsilon = 0$.

We want to compare the persistence diagrams produced when X is a subset of the Gaussian primes with the persistence diagrams arising from point clouds X produced by our random models. The bottleneck distance is one of the more commonly used metrics on the space of persistence diagrams. The bottleneck distance between two persistence diagrams is defined in terms of bijections between their sets of birth-death pairs. When constructing bijections between two persistence diagrams, it may be that the two persistence diagrams have a different number of birth-death pairs. For this reason, the convention is to allow the inclusion of extra points along the diagonal $y = x$, together with the original birth-death pairs. Indeed, even when the two persistence diagrams have the same number of birth-death pairs, it is still allowed that an optimal bijection would match a birth-death pair to some point on the diagonal.

Definition 3.1. The *bottleneck distance* between two persistence diagrams $\mathcal{P}_0$ and $\mathcal{P}_1$ is

$$d_B(\mathcal{P}_0, \mathcal{P}_1) = \inf_{\forall \beta \in \Sigma} \left(\sup_{\forall p \in \mathcal{P}_0} ||p - \beta(p)||_\infty \right).$$

Here, Σ denotes the set of all bijections between $\mathcal{P}_0$ and $\mathcal{P}_1$.

The bottleneck distance gives a bound on how far points in one persistence diagram need to move when being matched with the points of another persistence diagram. Think of the two persistence diagrams as sitting simultaneously in the first quadrant. If the bottleneck distance between $\mathcal{P}_0$ and $\mathcal{P}_1$ is d, then, given any $\epsilon > 0$, there is a bijection from $\mathcal{P}_0$ to $\mathcal{P}_1$ where no point of $\mathcal{P}_0$ is moved more than $d + \epsilon$ units (in the ℓ_∞ metric). That is, the individual (x, y)-coordinates are never changed by more than a difference of d.

We are now equipped with the necessary tools to discuss the differences between pairs of persistence diagrams. However, there is unfortunately no canonical method to compute the mean of two (or more) persistence diagrams. So, while persistence diagrams provide succinct, and pairwise comparable, descriptions of the persistent homology of individual point clouds, they are insufficient to describe the mean behavior of persistent homology across a collection of point clouds.

3.3. Rank functions and "mean persistent homology".

Another method of examining the differences between persistence diagrams is to consider the related rank functions. In contrast to persistence diagrams, rank functions have a canonical method to compute means across a collection. Hence with rank functions, we can describe the persistent homological behavior observed, "on average," across a collection of point clouds.

Definition 3.2. Let $\mathcal{P}$ be the k-dimensional persistence diagram for the Vietoris–Rips complexes associated with a finite point cloud X. The *rank function* of $\mathcal{P}$ is a function ρ from $\{(x, y) \in \mathbb{R}^2 \mid x \geq 0 \text{ and } y \geq x\}$ to $\mathbb{Z}$ where the value of $\rho(x, y)$ is the number of elements in the basis of the **persistent** homology that have been born by time x but have not yet died by time y.

Rank functions for persistence diagrams are introduced in [12] and they are closely related to the landscape functions defined in [3].

Given a finite collection of persistence diagrams $\{\mathcal{P}_1, \ldots, \mathcal{P}_n\}$ and the associated rank functions $\{\rho_1, \ldots, \rho_n\}$, the *mean rank function* is

$$\mu(x, y) = \frac{1}{n} \sum_{i=1}^{n} \rho_i(x, y).$$

Notice that the mean rank function will frequently take non-integer values and hence, it is almost always the case that the mean rank function is not actually the rank function of any persistence diagram.

The distance between two rank functions is the integral of their squared differences. And, in particular, the distance between a rank function ρ and the mean rank function μ is

$$\int_{\mathbb{R}^2} (\rho(x, y) - \mu(x, y))^2 \, dA.$$

4. Results

Using the approaches described above, we compare the distribution of Gaussian primes with point clouds produced by the Vardi–Cramér and symmetric Vardi–Cramér models. Specifically, if

$$\mathcal{R}_{200} = \{(x, y) \in \mathbb{R}^2 \mid |x| \text{ and } |y| \leq 200\},$$

then we compare the persistent homology of the Vietoris–Rips complex associated with the Gaussian primes inside $\mathcal{R}_{200}$ with the persistent homology of the Vietoris–Rips complexes associated with point clouds produced by each of the models in $\mathcal{R}_{200}$. We note that, on this scale, there are approximately $17,000$ points in each point cloud considered. This is a non-trivial number of points when performing persistent homology computations.

Our focus will be on comparisons and conclusions derived from considering 1-dimensional persistent homology. In §4.2, we examine the evolution of the first Betti numbers as ϵ increases. In §4.3, we examine 1-dimensional persistence diagrams and bottleneck distances, as well as mean rank functions and their associated distances. In the latter case, we are able to apply null hypothesis testing to conclude that both models are statistically unlikely to produce the Gaussian primes.

In §4.1, we briefly discuss our work on 0-dimensional persistent homology, where the models appear to align more closely with the Gaussian primes. However, in this case due to the large number of points being considered, computations of bottleneck and/or rank function distances are impractical and so we do not present a statistical analysis.

Finally, in §4.4, we include a discussion on what our techniques observe and why the associated statistical analysis is not applicable with respect to 2-dimensional persistent homology.

In all of the following comparisons, we have used 100 randomly generated point clouds produced by the indicated model.

4.1. Connected components in the Vietoris–Rips complexes.

For a fixed point cloud $X \subset \mathbb{R}^n$, recall that the zeroth Betti number, $b_0 = \text{Rank}\,[H_0(VR_\epsilon(X), \mathbb{R})]$, records the number of connected components of the Vietoris–Rips complex $VR_\epsilon(X)$. For $\epsilon = 0$, this is simply the number of points in the point cloud X. Furthermore, the rate at which b_0 decreases to 1 as ϵ increases is a measure of how difficult it is to connect the points in X by edges of bounded lengths.

Observation. In Figure 5 we compare values of b_0 when X is the set of Gaussian primes in $\mathcal{R}_{200}$ with the distribution of values of b_0 obtained for 100 random point clouds generated using both the basic and the symmetric Vardi–Cramér models (shown on the left and right, respectively). The Gaussian primes are indicated with a thick black line; the distribution of Betti numbers for the randomly generated point clouds are indicated by overlapping gray lines.

Discussion. The most notable difference between the two models is that, for any choice of ϵ, the symmetric model has a greater variation within the distribution of observed zeroth Betti numbers.

While there is not exact alignment in either model, there are several instances where the predicted values of b_0 from the models are close to the true value of b_0 for the Gaussian primes. Notably, the overall rates of decrease for b_0 from the Gaussian primes and from the models appear

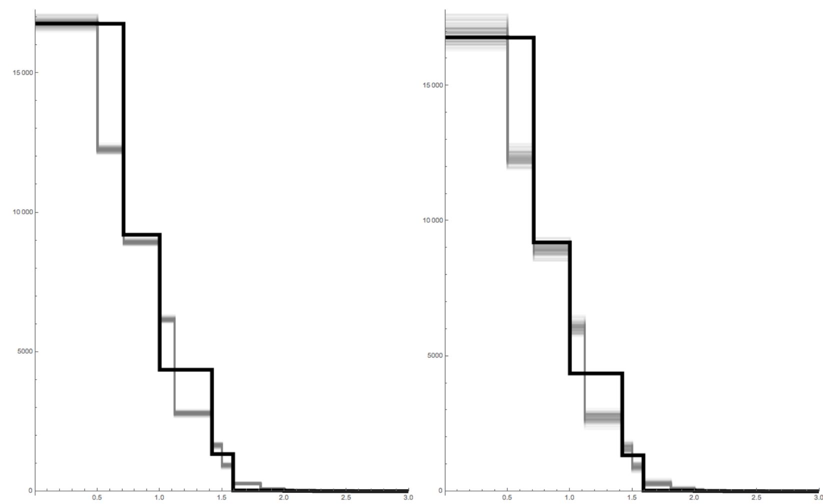


FIGURE 5. The evolution of the number of connected components, as a function of ϵ, in the Vietoris–Rips complexes based on the Gaussian primes (thick black) compared with the evolution of the number of connected components in the complexes built from 100 random point clouds created via the basic (gray on left) and symmetric (gray on right) Vardi–Cramér methods.

comparable. Moreover, the increased variability in the symmetric model only marginally improves its alignment with the true Gaussian curve.

Given that Vardi's model and our symmetric variant take the density of the Gaussian primes as a key input, it is not at all surprising that the complexes built on these point clouds become connected at similar rates. Additionally, the forced symmetry of our variant has very little impact on the radially-outward density used in Vardi's model, so it is also not surprising that we see very little difference in the analysis at this level.

4.2. The evolution of first Betti numbers.

For a fixed point cloud $X \subset \mathbb{R}^n$, we let b_1 denote the first Betti number of the simplicial complex $VR_\epsilon(X)$. That is, $b_1 = \text{Rank}\,[H_1(VR_\epsilon(X), \mathbb{R})]$. Just as we did in the previous subsection, for a fixed X, we graph b_1 as a function of ϵ. This allows us to see how the first Betti number changes as the parameter ϵ increases. It also allows us to compare this collection of Betti numbers for the Gaussian primes with the Betti numbers for point clouds generated by the random models. While the basic shape of the Betti curves produced by the models is consistent with the pattern produced by the Gaussian primes, there are notable differences.

Observation. In Figure 6 we compare the evolution of first Betti numbers for the Vietoris–Rips complexes built off of the Gaussian primes in $\mathcal{R}_{200}$ with the distribution of first Betti numbers obtained from Vietoris–Rips complexes built off of 100 random point clouds generated by the basic Vardi–Cramér model. The Betti numbers for the Gaussian primes are indicated with a thick black line; the distribution of Betti numbers for the randomly generated point clouds are indicated by overlapping gray lines.

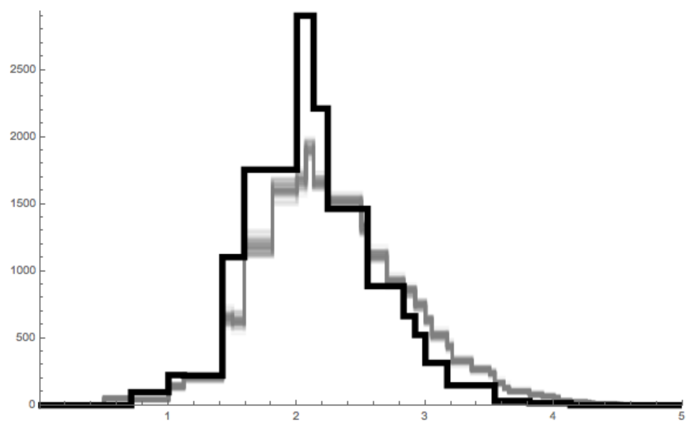


FIGURE 6. The first Betti number, as a function of ϵ, for Vietoris–Rips complexes based on the Gaussian primes (black) compared with the distribution of the first Betti numbers for complexes based on 100 random point clouds created via the basic Vardi–Cramér model (gray).

Discussion. If the Vardi–Cramér model was indeed producing accurate approximations to the Gaussian primes, one would anticipate that, for any ϵ, the first Betti numbers arising from the Gaussian primes would generally fall within the distribution of first Betti numbers arising from the model. However, the Gaussian primes produce a significantly higher maximum first Betti number and their first Betti numbers decay to zero faster than the model predicts.

Observation. In Figure 7 we compare the evolution of first Betti numbers for the Vietoris–Rips complexes built off of the Gaussian primes in

$\mathcal{R}_{200}$ with the distribution of first Betti numbers obtained from Vietoris–Rips complexes built off of 100 random point clouds generated by the symmetric Vardi–Cramér model. Again, the Betti numbers for the Gaussian primes are indicated with a thick black line and the distribution of Betti numbers for the randomly generated point clouds are indicated by overlapping gray lines.

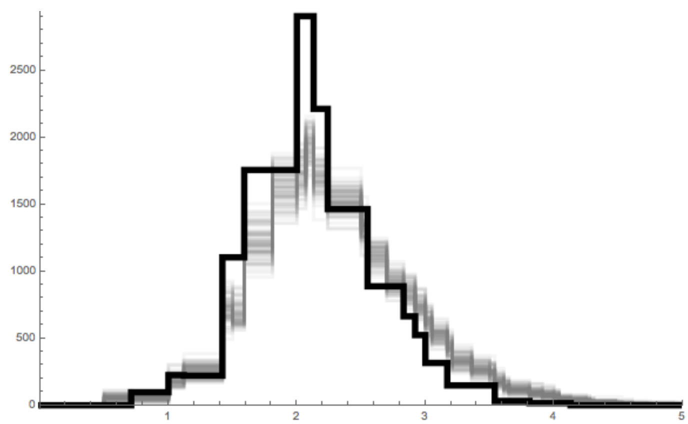


FIGURE 7. The first Betti number, as a function of ϵ, for Vietoris–Rips complexes based on the Gaussian primes (black) compared with the distribution of the first Betti numbers for complexes based on 100 random point clouds created via the symmetric Vardi–Cramér model (gray).

Discussion. The most notable difference from the previous case is that, for any choice of ϵ, the symmetric model has greater variation within the distribution of observed first Betti numbers. As discussed in §2.3, this increased variation stems from the fact that any homological feature arising from the random selection of points in the $\pi/4$-sector will be replicated eight times under the 8-fold symmetry. Furthermore, there are also features that arise when the $\pi/4$-sector is reflected over $y = x$ or the x-axis; the D_4 symmetry then replicates these features four times. It follows that homological features are frequently observed in quartets or octets, similar to how they appear in the Gaussian primes. Hence, in contrast to the 0-dimensional case, the higher variation in the Betti numbers produced by the symmetric model does allow for greater agreement with the Betti numbers arising from the Gaussian primes. Specifically, the growth and decay in the first Betti numbers for the Gaussian primes are closer to the distributions produced by the symmetric model. However, it is still the case that the primes produce a higher maximum and decay to zero faster than this symmetric model predicts.

4.3. Persistence diagrams.

Observation. Figure 8 compares the 1-dimensional persistence diagram for the Vietoris–Rips complexes associated with the Gaussian primes with the union of all 1-dimensional persistence diagrams for the Vietoris–Rips complexes associated with 100 point clouds randomly generated by the basic and symmetric Vardi–Cramér models.

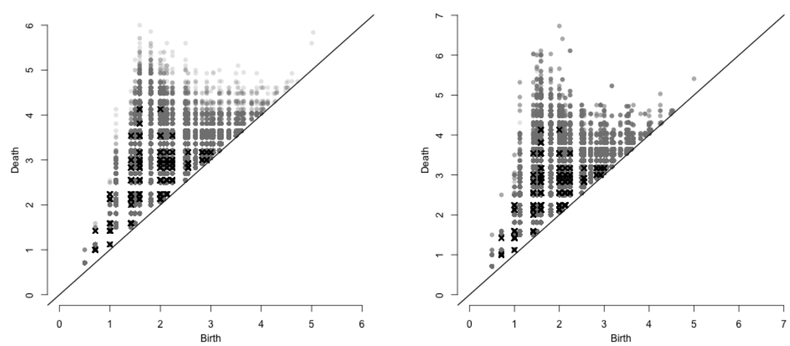


FIGURE 8. Comparing the 1-dimensional persistence diagrams arising from 100 random point clouds generated by the basic (gray on left) and symmetric (gray on right) Vardi–Cramér models with the 1-dimensional persistence diagram arising from the Gaussian primes (black).

Discussion. As with the comparison of the first Betti numbers, one sees some similarity between the persistence diagram produced by the Gaussian primes and the collection of diagrams produced by each model. However, there are once again some distinguishing features. Using either the basic or the symmetric model, one observes a large number of 1-dimensional homology classes that are born later than any that occur in the Gaussian primes. This is particularly true for the point clouds produced by the basic Vardi–Cramér model. Further, both models give rise to 1-dimensional homology classes that die later than any that occur in the Gaussian primes. This is particularly true for the point clouds produced by the symmetric Vardi–Cramér model. The impact of these distinguishing features is quantified by the two distance functions we described in §3.

4.3.1. Analysis of bottleneck distances.

Observation. We compute the bottleneck distance between the 1-dimensional persistence diagram for the Vietoris–Rips complexes associated with the Gaussian primes and the 1-dimensional persistence diagrams for the Vietoris–Rips complexes associated with 100 point clouds randomly generated by the basic Vardi–Cramér model, as well as 100 point clouds generated by the symmetric Vardi–Cramér model. The descriptive statistics for the distributions of observed bottleneck distances are shown in Table 2.

Table 2. Distribution of bottleneck distances between the persistence diagram associated with the Gaussian primes and 100 diagrams associated with random point clouds generated by each of the basic and symmetric Vardi–Cramér models.

	Basic Vardi–Cramér	**Symmetric Vardi–Cramér**
minimum	0.56	0.50
median	0.94	0.91
maximum	1.87	2.37
mean	1.01	1.06

Discussion. To give some context for the numerical significance of these statistics, one should note that all of these persistence diagrams have almost all of their birth and death times in the range $0 \leq \epsilon \leq 6$. So, having the mean bottleneck distance be roughly 1 suggests that, in every matching, there exists at least one birth-death coordinate pair which must be matched with a point that is roughly 16% of the total range of values away. Subsequently, this implies there are fundamental differences, specifically in the 1-dimensional homological structure, between the Gaussian primes and the point clouds generated by either model.

4.3.2. Analysis of rank functions.

In this section we will use rank functions to allow us to apply a null hypothesis test and conclude that the difference between the Gaussian primes and the point clouds arising from either model is statistically significant.

Observation. Given 100 random point clouds generated by the basic Vardi–Cramér model, we construct the mean rank function μ_b from all 100 rank functions arising from 1-dimensional persistence diagrams for each individual point cloud. We then compute the distance between μ_b and each of the 100 individual rank functions. This creates a distribution of possible distances observed within the basic model. Finally, we also compute the distance from μ_b to the rank function arising from the 1-dimensional persistence diagram for the Gaussian primes. This procedure is repeated for 100 point clouds randomly generated by the symmetric Vardi–Cramér model. We construct the mean rank function μ_s and compare the distribution of observed distances within the model to the distance from the Gaussian primes. Table 3 shows the descriptive statistics of both distributions, as well as the distance to the Gaussian primes. You will note that the distance from the mean rank function to the rank function for the Gaussian primes is considerably larger than the distance to any of the random point clouds.

TABLE 3. Distribution of distances between the mean rank function and the individual rank functions of 100 random point clouds generated by the basic and symmetric Vardi–Cramér models.

	Basic Vardi–Cramér	**Symmetric Vardi–Cramér**
minimum	242	2104
median	759	5944
maximum	3370	20205
mean	906	6908
standard deviation	522	3701
to Gaussian primes	189218	176051

Discussion. Now, our null hypothesis is that the Gaussian primes could be produced by either of the probabilistic models. Or more accurately, the distance from the mean rank function to the rank function of the Gaussian primes could be produced by either of these models. If the distribution of distances from the rank functions for the random point clouds generated by the models to the mean rank function was nearly normal, the probability of producing a distance as large as that produced between the mean rank function and the Gaussian primes would be essentially zero. However, our set of sample point clouds produces a skewed distribution of distances, so applying statistical methodology that relies on normality is inappropriate.

Chebyshev's Inequality provides a broadly applicable method of finding the maximum probability of a value arising n standard deviations from the mean for almost any distribution. If we assume that our sample size of 100 random point clouds generated by each model is sufficiently large to accurately determine each underlying distribution's true mean and standard deviation, we can then apply Chebyshev's Inequality to produce the upper bounds on probability indicated in Table 4. For example, the distance between the mean rank function for the symmetric Vardi–Cramér model and the rank function generated by the Gaussian primes is over 45 standard deviations from the mean. The chance of producing a distance this far from the mean, in almost any distribution, is less than 0.05%.

TABLE 4. Chebyshev's Inequality implies that it is highly unlikely for the models to produce a rank function with distance as far from the mean rank function as that achieved by the Gaussian primes.

	Basic Vardi–Cramér	**Symmetric Vardi–Cramér**
Standard deviations from the mean	360.47	45.70
Chebyshev's upper bound on probability	0.0000077	0.000479

A far more conservative approach would be to use the bound established in [14] for when the population mean and standard deviation are unknown and must be approximated by a sample mean and sample standard deviation. Again using our sample size of $n = 100$, we find that this conservative bound still gives the probability of seeing distances as far from the mean rank function as that produced by the Gaussian primes is less than 1% in both the basic and symmetric models.

Conclusion. In the language of hypothesis testing, we reject the null hypothesis at the $\alpha = 0.01$ level of significance, even under the most conservative assumptions about our distribution. Thus, we conclude that neither model is producing point clouds whose 1-dimensional persistent homology is consistent with the Gaussian primes.

4.4. Looking at 2-dimensional homology.

While H_2 is not motivated by the questions of Gordon and Vardi, higher dimensional **persistent** homology is connected to the geometry of the base point cloud. (See [1] and [2].) We will briefly describe what our techniques can, and cannot, say about this case.

Let $b_2 = \text{Rank}\,[H_2(VR_\epsilon(X), \mathbb{R})]$. That is, b_2 denotes the second Betti number of the simplicial complex $VR_\epsilon(X)$. Just as we did for b_0 and b_1, we will graph b_2 as a function of ϵ for a fixed X.

Observation. In Figure 9 we compare the evolution of b_2 when X is the set of Gaussian primes in $\mathcal{R}_{200}$ with the distribution of values of b_2 obtained from 100 random point clouds generated using both the basic (shown on the left) and the symmetric (shown on the right) Vardi–Cramér models. The Betti numbers for the Gaussian primes are indicated with a thick black line; the distribution of Betti numbers for the randomly generated point clouds are indicated by overlapping gray lines.

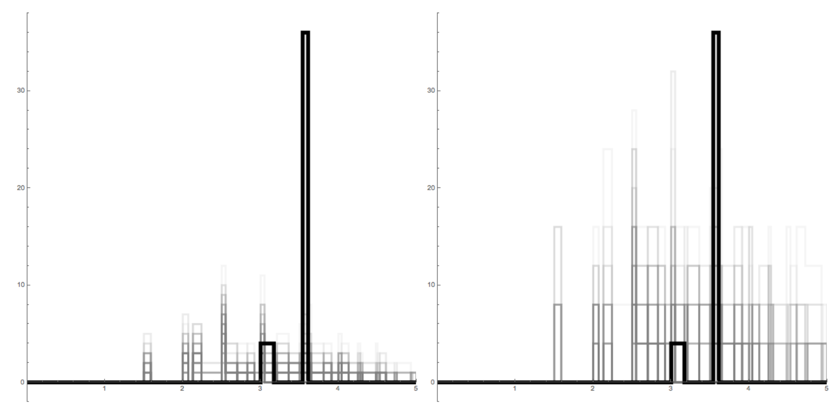


Figure 9. The second Betti number, as a function of ϵ, for Vietoris–Rips complexes based on the Gaussian primes (black) compared with the distribution of second Betti numbers for complexes based on 100 random point clouds created via the basic (gray on left) and symmetric (gray on right) Vardi–Cramér model.

Discussion. Most notably, the Gaussian prime Betti number curve exhibits exactly two non-zero features: a thin and tall spike near $\epsilon = 3.5$ and a shorter, wider feature near $\epsilon = 3$. None of the point clouds created under the basic model come anywhere near achieving the same height as the former and very few of these point clouds have any features with the same width as the latter. On the other hand, some of the random point clouds generated by the symmetric model have Betti curves that achieve similarly tall and thin spikes and other random point clouds have Betti curves with features of notable widths. Unfortunately, these desirable features generally occur at different values of ϵ from those in the Gaussian primes.

Observation. In Figure 10 we present the 2-dimensional persistence diagram for the Vietoris–Rips complexes associated with the Gaussian primes with the union of all 2-dimensional persistence diagrams for the Vietoris–Rips complexes associated with point clouds generated by the basic and symmetric Vardi–Cramér models.

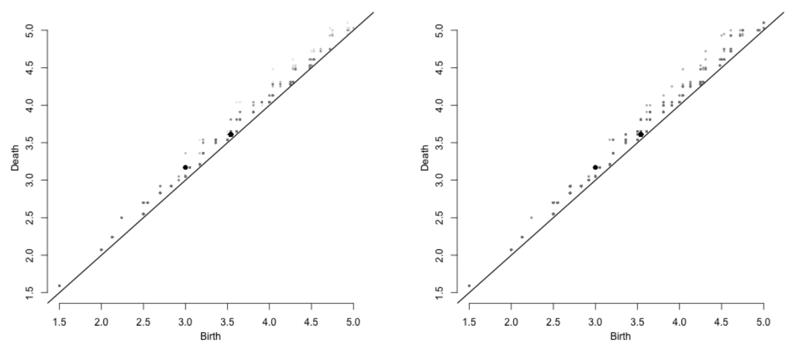


FIGURE 10. Comparing the 2-dimensional persistence diagrams arising from 100 random point clouds generated by the basic (gray on left) and symmetric (gray on right) Vardi–Cramér models with the 2-dimensional persistence diagram arising from the Gaussian primes (black).

Discussion. Quantifying the impact of these distinguishing features is beyond the scope of bottleneck distances. One can see that all of the birth-death pairs occur very close to the diagonal $y = x$. For this reason, the optimal matching used to compute pairwise bottleneck distances will almost always pair every point with points along the diagonal. So the bottleneck distances will not specifically measure the difference between diagrams, but instead merely report that both diagrams are very close to the diagonal. Essentially, we have that all persistence diagrams considered are "small" under the bottleneck metric and hence must be close to each other, even if they contain drastically different birth-death points.

Similarly, these distinguishing features are not well quantified by rank functions. In general, when using persistent homology techniques, one considers homological features which have birth-death pairs near the diagonal to be noise. However, following the work in [1] and [2], the indication is that some of these short-lived features could be directly arising due to the geometry of the point cloud, rather than just as artifacts of the topological methods. Unfortunately, in our data, one sees great variation across all point clouds constructed by either the basic or symmetric Vardi–Cramér models. Specifically, the precise location of the birth-death pairs travels up and down the diagonal from around $\epsilon = 2$ to $\epsilon = 5$ with no clear preference indicated across all trials. For this reason, when constructing the mean rank functions, any relevant properties that may exist will almost certainly be lost in the noise since they both exist on similar scales. Hence, comparisons to the Gaussian prime rank function are uninformative. This is especially true in the symmetric case, where the Betti number graphs from the model exhibit the "tall and thin" spike behavior at a variety of different ϵ values.

So while there are observable differences between the persistent second homology of the Gaussian primes and the persistent second homologies of point clouds generated by either the basic or symmetric Vardi–Cramér models, the techniques of this paper are not well suited to specifically quantify or analyze their significance. We would like to note, however, it appears that once again the inclusion of symmetry into the model induces a greater variability in the observed persistent homology which broadly translates to a higher level of agreement with the true Gaussian primes.

5. Concluding Remarks

We have used statistical techniques, applied to persistent homology, to argue that the basic Vardi–Cramér model and the symmetric Vardi–Cramér model are not producing point clouds in $\mathcal{R}_{200}$ that accurately describe the geometry of the Gaussian primes. In doing so we have essentially used a null hypothesis methodology, which, as is nicely described in [13], is not straightforward but is reasonable. We also acknowledge that, while we are working with a large number of points in the plane and our samples are about as large as our computing power will allow us to analyze, we are using finitary methods to analyze the fitness of models and thus give us insights into conjectures about asymptotic behavior. Even with these caveats in the foreground, we believe that the visual differences in the first Betti numbers, the bottleneck distance statistics, and especially the rank function distances, give evidence to suggest that neither Vardi–Cramér's probabilistic method of modeling the Gaussian primes nor its symmetric variant are accurately approximating the geometry of the Gaussian primes.

Acknowledgments. We thank Vidit Nanda for helpful comments and for his program Perseus, developed from the ideas in [11], which we used in all of our persistent homology computations. Further data analysis was completed in part using the Topological Data Analysis library (TDA, [6]) in R. We also thank Henry Adams for giving us the opportunity to present some of this work at a special session on topological data analysis at the Joint Mathematics Meetings in 2018 and for suggesting that we should examine persistent 2-dimensional homology.

References

[1] Michał Adamaszek and Henry Adams, *The Vietoris–Rips complexes of a circle*, Pacific J. Math. **290** (2017), no. 1, 1–40.

[2] Michał Adamaszek, Henry Adams, and Samadwara Reddy, *On Vietoris–Rips complexes of ellipses.* J. Topol. Anal. in press. https://doi.org/10.1142/S1793525319500274

[3] Peter Bubenik, *Statistical topological data analysis using persistence landscapes*, J. Mach. Learn. Res. **16** (2015), 77–102.

[4] Erin W. Chambers, Vin de Silva, Jeff Erickson, and Robert Ghrist, *Vietoris–Rips complexes of planar point sets*, Discrete Comput. Geom. **44** (2010), no. 1, 75–90.

[5] Herbert Edelsbrunner, *A Short Course in Computational Geometry and Topology.* SpringerBriefs in Applied Sciences and Technology. Cham: Springer, 2014.

[6] Brittany T. Fasy, Jisu Kim, Fabrizio Lecci, Clement Maria, David L. Millman, and Vincent Rouvreau. The included GUDHI is authored by Clement Maria, Dionysus by Dmitriy Morozov, and PHAT by Ulrich Bauer, Michael Kerber, and Jan Reininghaus. *TDA: Statistical Tools for Topological Data Analysis.* Version 1.6.5. 2017. Available at https://rdrr.io/cran/TDA/

[7] Ellen Gethner, Stan Wagon, and Brian Wick, *A stroll through the Gaussian primes*, Amer. Math. Monthly **105** (1998), no. 4, 327–337.

[8] Robert Ghrist, *Barcodes: The persistent topology of data*, Bull. Amer. Math. Soc. (N.S.) **45** (2008), no. 1, 61–75.

[9] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers.* 5th ed. New York: The Clarendon Press, Oxford University Press, 1979.

[10] Allen Hatcher, *Algebraic Topology.* Cambridge: Cambridge University Press, 2002.

[11] Konstantin Mischaikow and Vidit Nanda, *Morse theory for filtrations and efficient computation of persistent homology*, Discrete Comput. Geom. **50** (2013), no. 2, 330–353.

[12] Vanessa Robins and Katharine Turner, *Principal component analysis of persistent homology rank functions with case studies of spatial point patterns, sphere packing and colloids*, Phys. D **334** (2016), 99–117.

[13] Andrew Robinson and Katharine Turner, *Hypothesis testing for topological data analysis*, J. Appl. Comput. Topol. **1** (2017), no. 2, 241–261.

[14] John G. Saw, Mark C. K. Yang, and Tse Chin Mo, *Chebyshev inequality with estimated mean and variance*, Amer. Statist. **38** (1984), no. 2, 130–132.

[15] Ilan Vardi, *Prime percolation*, Experiment. Math. **7** (1998), no. 3, 275–289.

[16] Afra Zomorodian and Gunnar Carlsson, *Computing persistent homology*, Discrete Comput. Geom. **33** (2005), no. 2, 249–274.

(Hammarsten) Lafayette College; Easton PA 18042
Email address: `hammarsc@lafayette.edu`

(Helmreich) Wheaton College; Norton MA 02766
Email address: `helmreich_rae@wheatoncollege.edu`

(Krishnan) University of Washington Bothell; Bothell WA 98011
Email address: `an97@uw.edu`

(Meier) Lafayette College; Easton PA 18042
Email address: `meierj@lafayette.edu`

(Schmitz) University of Minnesota, Minneapolis MN 55455
Email address: `schm5112@umn.edu`