Foundations of Data Science
doi:10.3934/fods.2021007

THE (HOMOLOGICAL) PERSISTENCE OF GERRYMANDERING

Moon Duchin

Department of Mathematics
Tufts University
Medford, MA 02155, USA

Tom Needham*

Department of Mathematics
Florida State University
Tallahassee, FL 32306, USA

Thomas Weighill

Department of Mathematics and Statistics
UNC Greensboro
Greensboro, NC 27402, USA

(Communicated by Gunnar Carlsson)

Abstract. We apply persistent homology, the dominant tool from the field of topological data analysis, to study electoral redistricting. We begin by combining geographic and electoral data from a districting plan to produce a persistence diagram. Then, to see beyond a particular plan and understand the possibilities afforded by the choices made in redistricting, we build methods to visualize and analyze large ensembles of alternative plans. Our detailed case studies use zero-dimensional homology (persistent components) of filtered graphs constructed from voting data to analyze redistricting in Pennsylvania and North Carolina. We find that, across large ensembles of partitions, the features cluster in the persistence diagrams in a way that corresponds strongly to geographic location, so that we can construct an average diagram for an ensemble, with each point identified with a geographical region. Using this localization lets us produce zonings of each state at Congressional, state Senate, and state House scales, show the regional non-uniformity of election shifts, and identify attributes of partitions that tend to correspond to partisan advantage.

The methods here are set up to be broadly applicable to the use of TDA on large ensembles of data. Many studies will benefit from interpretable summaries of large sets of samples or simulations, and the work here on localization and zoning will readily generalize to other partition problems, which are abundant in scientific applications. For the mathematically and politically rich problem of redistricting in particular, TDA provides a powerful and elegant summarization tool whose findings will be useful for practitioners.

2020 *Mathematics Subject Classification.* Primary: 62R40, 55N31; Secondary: 68T09.

Key words and phrases. Topological data analysis, Fréchet means, Markov chain ensembles, political redistricting.

The first and third author were supported by NSF Convergence Accelerator Grant No. OIA-1937095.

* Corresponding author: Tom Needham.

1. **Introduction.** In this paper, we bring the techniques of topological data analysis to bear on electoral redistricting. We do this through the novel combination of several mathematical techniques, applied with in-depth knowledge of the application domain to address its unique challenges and important questions. We begin by introducing the key notions.

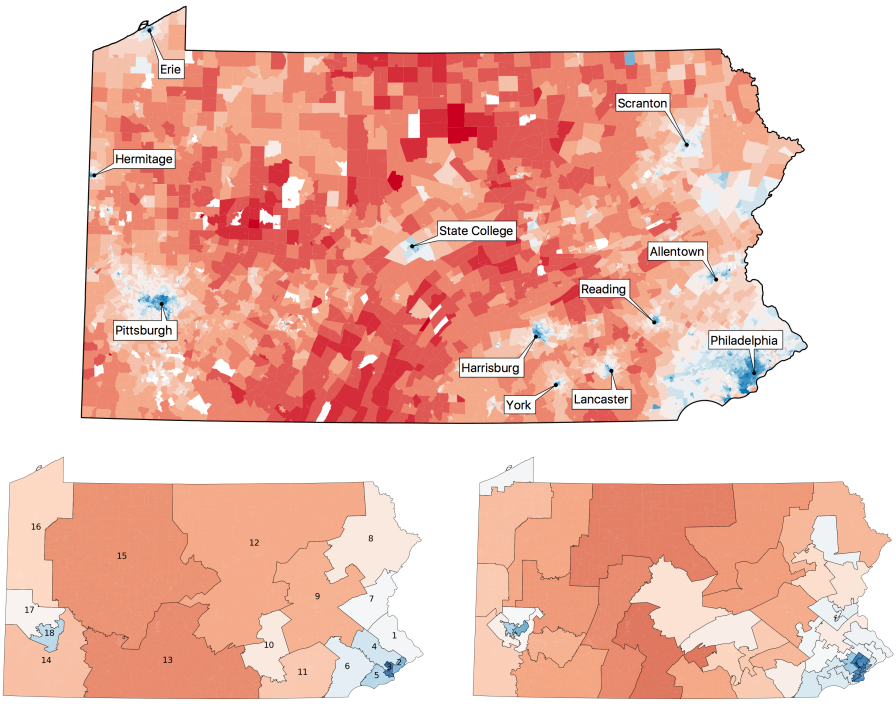
1.1. **Redistricting.** In electoral politics, *redistricting* is the process of drawing new boundary lines for electoral regions called *districts*, in which an election will be conducted to select one or more people as representatives for a governing body. For instance, the United States Congress has a House of Representatives that has had 435 seats for over a hundred years. These seats are apportioned to the states after every decennial Census—for instance, the 2010 Census left Pennsylvania with 18 seats and North Carolina with 13, in rough proportion to their population—and then it is left to the various states to partition their territory into geographically-delimited districts that will elect a single member each. In most cases, it is the state legislature that controls the process of drawing not only Congressional boundaries, but also the boundaries for their own legislative districts. In the U.S. system, which is dominated by two major parties, this opens the door to *partisan gerrymandering*, where the lines are carefully arranged to maximize the seats secured by their side, given the anticipated pattern of voting. For instance, the last few years have seen successful legal challenges in both Pennsylvania and North Carolina, where enacted plans produced by Republican-controlled legislatures were found to be Republican-favoring "gerrymanders"—in other words, to be so favorable to the party in charge that they impermissibly subordinate neutral criteria to a partisan agenda. There are many other kinds of gerrymandering, including agendas with racial components, or to favor or target particular incumbents, but the present paper focuses on developing mathematical tools to understand the interplay of district lines and party outcomes. For more background on redistricting, see [29, 43] and the forthcoming book [31].

1.2. **Persistent homology.** Persistent homology is the most widely recognized tool used in the burgeoning field of topological data analysis (TDA), where concepts from algebraic topology are used to simplify, summarize and compare complex datasets. Methods from topological data analysis have proven successful in neuroscience [25, 6], medical diagnostics [53, 23], and machine learning [2, 14, 7, 41], among many other applications—these references are just a selection of many dozens of examples in each of these areas. However, applications to demographic or electoral geospatial data seem to be comparatively limited; we only know of a few examples. In [4], Bajardi et al. use mobile phone data to study patterns of immigration. In [57], Stoltz et al. use TDA to study Brexit voting patterns, finding for instance that Switzerland is surrounded by countries with a different voting tendency. In [5], Banman and Ziegelmeier apply persistent homology to global health and wealth statistics to uncover subtle developmental disparities between geographically close countries. Finally, in the example closest to our own approach, Feng and Porter introduce in [34] a filtration for producing a topological signature from districting and vote data which is similar to filtration we employ below. (See also the broader paper [35] for spatial applications.) The applications considered in [34] are quite different from those considered here: Feng and Porter primarily used topological

2020 Mathematics Subject Classification. Primary: 62R40, 55N31; Secondary: 68T09.

Key words and phrases. Topological data analysis, Fréchet means, Markov chain ensembles, Political redistricting.

signatures to study the vote patterns at the precinct level for fixed regions, whereas our goal is to understand the variation among districting plans, so that a single map does not suffice. We will do that by studying topological features across collections or "ensembles" containing thousands of alternative plans.



Congressional plan (enacted 2018) State Senate plan (enacted 2011)

FIGURE 1. This paper studies the aggregation of votes from geographic units to districts, which is intractable to study by complete enumeration of possibilities. We develop persistent homology techniques to summarize the districting problem. This figure shows a precinct map of Pennsylvania with key cities identified. The precincts are colored by voting results from the 2016 Presidential election (blue for Democratic, red for Republican). Current districts in Pennsylvania are shown with the same coloring. Note that the Congressional districts are balanced to within one-person population deviation (by 2010 Census count).

1.3. **Studying districting ensembles with persistent homology.** Suppose we want to evaluate a districting plan, either proposed or enacted, to label it as either a partisan gerrymander or not. Some authors have proposed universal metrics for this purpose, which do not take the specific state, time period, or level of redistricting into account: the same score, denominated in the same units, is supposed to flag gerrymanders for Iowa's Congressional seats in 2020 or for Mississippi state House in 1960. Examples which are widely cited include the efficiency gap [56], and the

mean-median score [48], which belongs to a family of "partisan symmetry" metrics [37].

On the other hand, there has been growing consensus around the idea that evaluating a map requires comparing it to a range of alternatives that hold the geography, demography, and redistricting rules constant. In the last several decades, the use of "demonstration plans" has been rising in court cases: in addition to the map under consideration, a court also views plans submitted by various advocacy groups and political actors. While traditionally all of these benchmark plans are drawn "by hand," the last ten years has seen the adoption of computer algorithms in court cases, where large generated ensembles of possible districting plans are used as a backdrop of comparison for a proposed or enacted plan. The use of algorithmic *ensemble methods* has already had a major impact on a number of recent court cases, notably Pennsylvania and North Carolina, where ensemble methods were cited by courts in the successful invalidation of the state's enacted maps as partisan gerrymanders [47, 54, 30].

Algorithmic ensembles can easily boost the number of alternatives to consider into the hundreds of thousands, far beyond the point where each map can be carefully analyzed with the human eye. For this reason, *summary statistics* are computed for the maps in the ensemble, then compared to the plan under evaluation. For example, one might choose one or several recent elections in order to fix a background pattern of voting, then count the number of districts won by each party as the district lines vary. If the map under evaluation gives the controlling party the majority in an outlying number of districts—in other words, if the plan secures more seats for Party A than the vast majority of alternatives—this could reveal that the map was made with impermissible partisan intent and/or has unreasonably skewed effects.

We can therefore say that modern research into redistricting forces us to be able to understand the essential qualities of large collections of territorial partitions, preferably in a way which captures both geography and relevant summary statistics, such as those connected to party or race. The method we propose in this paper is illustrated with respect to partisan statistics, but can be applied more generally. We combine and summarize the adjacency and vote share information from a districting plan into a *dual graph* (see §2.1), decorated by vote shares under some given electoral pattern. We can now ask questions like: how many connected components of support for Party A are there? A more sophisticated version of this question, which considers multiple thresholds at once, is answered by persistent homology, the mathematical theory underlying our method.

In this paper we apply methods from TDA to summarize ensembles of districting plans generated via Markov chain Monte Carlo (MCMC) sampling. To be precise, we study the ensembles by transforming them into persistence diagrams, which admit manageable summaries. This is accomplished by computing degree-0 persistent homology of the filtered dual graphs of the plans and passing to a barycenter construction through the use of Fréchet means. Similar methods could be attempted for higher-dimensional homology, which is left for future work. There has been some prior work using TDA to study ensembles of generated data, such as simulated polymer conformations [33] or simulated neuronal spike trains [25, 21]. To our knowledge, this is the first geospatial application.

It is important to note that the TDA methods defined here are agnostic to the Markov chain (or other) methods used to generate an ensemble of districting plans.

We use a method called Recombination [28] for most of the illustrative analysis, but this can be treated as a black box for the current paper.

1.4. **Broad goals and contributions.** Before we proceed, we highlight an important guiding principle in this paper: namely, we will measure the success of our topological data analysis techniques by their ability to uncover information which cannot be easily observed from maps "with one's eyes." Consider Figure 1, for example, which shows the distribution of votes in the 2016 Presidential election in Pennsylvania (following the contemporary U.S. color convention of red for Republican, blue for Democratic). No advanced data analysis techniques are required to observe that Philadelphia and Pittsburgh are far bluer than the rural north of the state. What is not clear, however, is where the blue *districts* will fall in a typical redistricting plan. This is partly because the blue clusters in the map have widely different population sizes. A small Democratic-leaning city like State College (≈50,000 people) will have little impact on a 700,000-person Congressional district, but may dominate the voting in a smaller 63,000-person state House district if the boundaries fall just so. To sharpen this point, consider the single precinct on the north border of Pennsylvania that is blue in a sea of red. In 2018, this precinct had 66 votes for Clinton and 18 votes for Trump, for a Democratic share of over 78%. This is manifestly irrelevant for redistricting, but would produce a very persistent feature in a precinct-level analysis. This shows that we need to contend with aggregation to distinguish signal from noise in the districting application. Though this is phrased in the language of redistricting, a similar challenge of summarizing the vast possibilities for aggregation and partition occurs in many application domains.[1]

We develop the following techniques, stated here for the redistricting application but with an eye to generalization.

- **Scale and zoning.** We leverage TDA to study nuanced scale effects in partition problems. In particular, persistence diagrams let us read off a *zoning* of a state into redistricting-relevant regions at each districting scale.
- **Comparing elections.** Fixing a scale of aggregation, we can study the difference between two filtrations in natural units. Here, that lets us compare different vote patterns on the same precincts.
- **Signals of gerrymandering.** Finally, we consider the use of TDA to study differences between ensemble methods. For elections, we can compare sets of districting plans gathered by partisan-biased procedures, looking for systematic differences at the level of geography.

Using the techniques above, we have several concrete findings in our case study states of Pennsylvania and North Carolina. Details are found below, but we highlight the findings here.

- **Successful localization and summarization.** Figures 8 and 12 show a remarkable degree of consistency between persistence data and geography with respect to the presidential contest in 2016, and we observed the same for each election in our dataset. This is a fundamental finding that enables the rest of the analysis. Furthermore, the iterative Fréchet mean algorithm succeeds at producing ensemble summaries on these datasets at a reasonable computational cost, which is not always the case in data-intensive applications.

[1] For just one example, minimizing communications latency in the assignment of subtasks and data to processors in massively parallel high-performance computing would benefit from partition sampling to provide an overview of clustering.

- **(Non-)uniformity of partisan swing.** A longstanding debate in political science concerns the use of a *uniform swing* hypothesis for shifts between voting patterns; that is, if one election has a Republican share that is 5 points higher than another statewide, can we reasonably assume that preferences shifted 5 points in every district or even every precinct? Our work produces detailed findings on the relative swing in regional clusters from one election to the next. For instance, we find that swing is fairly uniform in the Pittsburgh cluster of districts across five elections held in 2010 and 2012. However, a fundamental shift can be detected between 2012 and 2016, making the Pittsburgh cluster more persistent. In other words, we observe a marked mid-decade increase in polarization, seen in the difference between Pittsburgh districts and their surroundings, at all three levels of redistricting ($k = 18, 50, 203$ districts). We get similarly granular insights in North Carolina: Charlotte swings fairly uniformly across the Census cycle; Asheville does not.
- **Correlates of partisan tilt.** The partisan-biased ensembles admit several nontrivial observations correlating regional persistence to overall party advantage. For instance, State Senate plans that have a higher Democratic share in the Pittsburgh districts are associated with Republican advantage overall—this is an instance of the well-known phenomenon of "packing" one kind of voter, which "wastes" their votes. More surprisingly, though the Scranton cluster always stays competitive, surrounding Scranton with redder districts (i.e., increased persistence) is correlated with Democratic advantage in the ultimate Senate outcome; Scranton is the only Pennsylvania zone with that property. Greensboro, NC exhibits this property in reverse: making the Greensboro districts more like their periphery (i.e., decreased persistence) is associated with greater Democratic success overall.

These observations can be corroborated by secondary analysis. Relative to the precinct-level data from which they were built, these findings pass the "can't-be-seen-with-your-eyes" test with flying colors, and we hope that they will generate interest in the wider TDA community as well as among political scientists and redistricting practitioners.

1.5. **Outline of the paper.** In Section 2, we define districting plans in the language of graph theory and briefly describe the methods that generate ensembles of plans. The persistent homology framework for studying districting plans is laid out in detail; to keep the exposition self-contained, we describe the relevant tools from TDA with a narrow focus on the application at hand. Section 3 presents the three principal applications described above. We use the persistent homology formalism to compare scales of redistricting, to compare data from different elections on the same districting plan ensembles, and to determine distinguishing features from ensembles of partisan-tilted districting plans. In Sections 4-5, we apply our methods to ensembles of plans in Pennsylvania and North Carolina. Finally, the stability and robustness of the TDA signatures is explored in Section 6. Drawing on a standard result from [22] on the stability of persistence diagrams to perturbations in their input data, we derive a guaranteed upper bound on our districting plan representations when perturbing electoral and geographical data. These theoretical results are accompanied by data that raises our confidence in the robustness of our analysis.

2. **Background and mathematical framework.** In this section we describe the details of our topology-based approach to studying districting plans. We begin by precisely defining the notion of a districting plan and briefly describing our method for generating ensembles of plans. Next we apply the *persistent homology* pipeline to certain filtered graphs associated to the plans. To keep the description accessible, we opt to restrict our exposition of persistent homology to our specific application rather than appealing to any general theory. We use [32, 11] as general references for persistent homology.

2.1. **Redistricting as a graph partition problem.** A *districting plan* is a partition of a state into geographical regions meeting certain criteria, which vary by state. The building blocks are geographic units such as census blocks or election precincts. A very brief summary of key redistricting rules and principles is needed to set up the problem. The number of Congressional districts in a state is deterministically related to the Census population by a formula that has been in place for nearly sixty years; the new apportionments will be announced in 2021. The number of districts in the state legislative houses is typically fixed in state constitutions and is rarely changed. A criterion for a districting plan which is common to most states (and essentially active in all) is that each district must be "contiguous," i.e., the induced subgraphs on the pieces of the partition should be connected. And it is a universal requirement that the districts be population-balanced: each district must have equal population, up to a small tolerance for deviation. For our purposes we will fix a deviation limit ϵ, so that each of the districts should be no less than $(1-\epsilon)$ and at most $(1+\epsilon)$ times ideal size, which is total Census count divided by k.

After fixing the choice of units, we are able to define a districting plan more formally as a graph partition, with notation following [29, 1, 28]. Given a tiling of a state by geographic units that overlap only along boundaries, let $G = (V, E)$ be the *unit dual graph*—for instance, we can form a precinct dual graph or a census block dual graph. This is defined by having each vertex in V correspond to a geographic unit of the state, then connecting two vertices by an edge if and only if their units are geographically adjacent (i.e., share a boundary of positive length). We denote vertices in V by v or w and an edge in E with endpoints v and w is denoted $\overline{vw}$. For a subset $V' \subseteq V$, recall that the *induced subgraph* on V' is the graph $G' = (V', E')$ containing all the edges of E that have both endpoints in V'. We adopt the slight abuse of notation to refer by G' to either the induced subgraph or its vertex set V', as needed. We will frequently make use of functions $p : V \to \mathbb{R}$, such as a population count for every unit. We will write $p(G')$ to refer to the sum $\sum_{v \in V'} p(v)$.

Definition 2.1 (Districting plans as balanced connected partitions)**.** Given a graph $G = (V, E)$, a number of districts $k \geq 2$, a weighting function $p : V \to \mathbb{R}$, and a balance threshold ϵ, let $\mathcal{P}_{k,\epsilon}(G)$ be the set of partitions $P = \{P_1, \ldots, P_k\}$ where $P_i = (V_i, E_i)$ are induced graphs on a vertex partition $V = V_1 \sqcup \cdots \sqcup V_k$ with each P_i connected and satisfying the balance condition

$$(1-\epsilon) \cdot \frac{p(G)}{k} \leq p(P_i) \leq (1+\epsilon) \cdot \frac{p(G)}{k}.$$

Each $P \in \mathcal{P}_{k,\epsilon}(G)$ is called a *districting plan*, and each P_i is called a *district*.

As alluded to above, districting plans are typically subjected to other tests of quality. One prominent example is "compactness," which requires that the shapes of districts look reasonable on a map. Other criteria prefer plans that hew to county

and municipal boundaries, that respect "communities of interest," and so on. And finally, but by no means insignificantly, the whole country is subject to the Voting Rights Act of 1965 (VRA), a rule that requires adequate access to representation for certain kinds of demographic minorities. Generating a districting plan meeting all relevant validity criteria is a complicated and delicate process, and in this paper we will only operationalize population balance, contiguity, and compactness.[2]

Perhaps the most concise summary of the geographic layout of a redistricting plan is the data of which districts are next to which. This information is naturally captured in a graph structure that is formed by aggregation from the dual graph of the state's geographic units: each vertex of the graph now corresponds to a district, and edges connect nodes from adjacent districts (see Figures 3 and 16). This construction is referred to as a *district dual graph* for the plan.

One might hope that even within a large ensemble of plans, the types of graphs one obtains are relatively restricted, thereby uncovering a common feature of the plans in the ensemble. It turns out in practice that there is a huge variety in the types of graphs in even a modest ensemble of 1000 plans. In a generated ensemble of 18-district plans for Pennsylvania (details of ensemble generation are provided in Section 2), 991 distinct graph isomorphism classes are represented among the 1000 dual graphs. Examples of some of these graphs are shown in Figure 2 and some basic graph statistics of the ensemble are shown in Figure 19. As the number of districts in the plan increases, so does the variety of graphs: the dual graphs constructed from our generated ensembles of 1000 50-district plans are typically all unique up to isomorphism. This suggests that passing from districting plans to dual graphs is not enough of a reduction in representation complexity to make direct analysis of the space of a plans a tractable option.

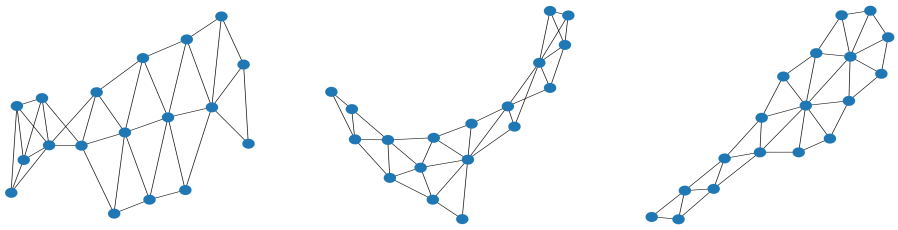


FIGURE 2. Abstract graph representations of three 18-district plans for Pennsylvania.

2.2. **Markov chains and districting ensembles.** In this paper we will generate plans via a Markov chain algorithm called ReCom (short for Recombination), but our TDA methods can be used to study any ensemble of districting plans regardless of how it was generated. We briefly describe the ReCom algorithm; more details can be found in [28] (see also [27] for an early application of this algorithm to redistricting in Virginia). The GerryChain package, which implements the ReCom algorithm, is publicly available at: https://github.com/mggg/GerryChain.

Let $G = (V, E)$ denote a unit dual graph and suppose that $P = \{P_1, \ldots, P_k\}$ is a valid districting plan (meaning that it satisfies all of the geographic and demographic criteria required by the particular state). ReCom is a Markov chain whose state

[2]For work that layers in more complicated districting criteria, see for instance [26].

space is $\mathcal{P}_{k,\epsilon}(G)$. A step in the chain is taken as follows. First, two distinct adjacent districts P_i and P_j are chosen uniformly at random. Next, a random spanning tree of the connected subgraph $P_i \cup P_j \subset G$ is constructed. Deleting an edge from this spanning tree repartitions $P_i \cup P_j$ into $P_i' \cup P_j'$, which together with the other districts defines a new partition of G. After a validity check, the chain can move there as its next state. If no edge is found whose deletion produces a valid partition, then a new spanning tree is constructed, and the process iterates.

Details of the ReCom algorithm, as well as theoretical and experimental evidence of its performance, are provided in [28]. The main takeaways are that ReCom runs quickly enough to generate large ensembles, with convergence heuristics indicating approximate independence of starting point. The ReCom stationary distribution is well-approximated by a distribution on $\mathcal{P}_{k,\epsilon}(G)$ that weights a plan P in proportion to its compactness, given by an appropriate isoperimetric score on the districts [10].

2.3. **Filtrations on dual graphs.** Let $P = \{P_1, \dots, P_k\} \in \mathcal{P}_{k,\epsilon}(G)$ be a districting plan for a state with unit dual graph G. As described in Section 2.1, we form the *district dual graph* associated to P, denoted $G_P = (V_P, E_P)$, as the quotient graph of G associated to the partition P, with vertices w_i corresponding to districts P_i.

Since the vertices of the dual graph G represent geographic units, numerical data assigned to the units (such as population, demographic counts, or vote counts) induce functions $f : V \to \mathbb{R}$. The main data of interest for the applications in this paper will be the number of votes for a Republican candidate $r : V \to \mathbb{R}$ and the number of votes for a Democratic candidate $d : V \to \mathbb{R}$ in a particular election. Given this data, we can compute the Republican vote share by district $R : \mathcal{P}_{k,\epsilon}(G) \times \{1, \dots, k\} \to \mathbb{R}$ as follows:

$$R_P(i) = R(P, i) = \frac{r(P_i)}{r(P_i) + d(P_i)}.$$

Note that this can be thought of as an average of a binary Republican voting function over the major-party voters of the district.

Although the analysis in this paper will focus on the Republican vote share function R, we remark that our method can be used to analyze any function of interest $f : V \to \mathbb{R}$. From such a function, one can always induce a function F summing or averaging f over the units that make up each district. We regard such an F as a *filtration function* on the vertices of G_P. For the sake of simplicity, we sometimes make the (mild) assumption that the values of any filtration function F are distinct over the districts of a plan, and we will also assume that the range of F is in the interval $[0, 1]$. Then the output of any filtration function F on a plan with k districts is k values, which can be organized in increasing order, $0 \leq t_1 < t_2 < \cdots < t_k \leq 1$.

Definition 2.2 (Filtration)**.** Let G_P be a district dual graph for a plan with k districts and a fixed filtration function F, and suppose that the vertices $\{w_1, \dots, w_k\}$ are indexed in increasing order $F(w_1) < \cdots < F(w_k)$. Let $t_i = F(w_i)$. Then we define the *filtered dual graph* associated to (G, P, F) to be the sequence of graphs

$$G_P^{t_1} \subset G_P^{t_2} \subset \cdots \subset G_P^{t_k},$$

where each $G_P^{t_j}$ is the induced graph on the vertices $\{w_1, \dots, w_j\}$.

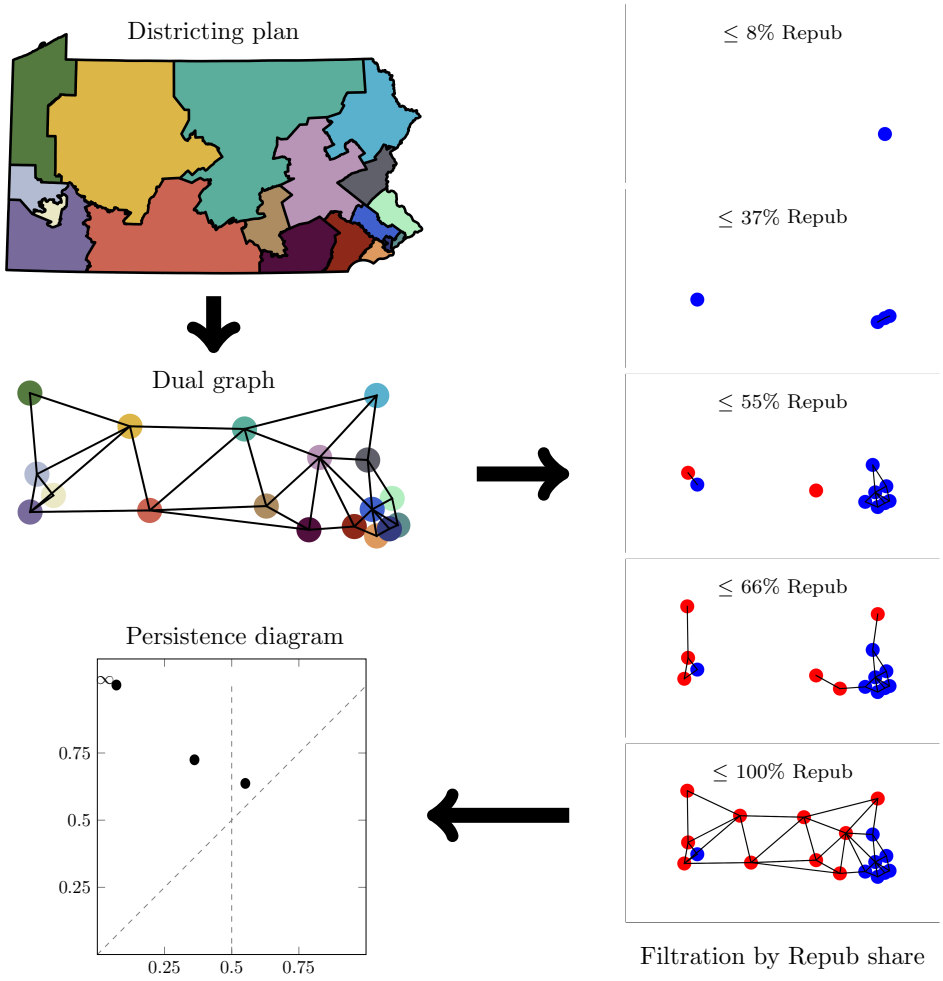


FIGURE 3. The main method: turning a districting plan into a persistence diagram. The persistence diagram captures both the adjacency information and the vote shares for the districts in a plan.

This is illustrated in Figure 3. The definition of $E_P^{t_j}$ simply means that edges from G_P are added to the filtered dual graph as soon as possible (i.e., as soon as both of their endpoint vertices appear). In the standard language of persistent homology, the filtered dual graph is a *one-dimensional filtered simplicial complex* obtained via a *sublevel set filtration* [22].

2.4. **Persistence diagrams.** We would like to track the evolution of the shape of a filtered dual graph as we run through the values t_j. We focus on the evolution of the simplest topological features of the graphs: their number of connected components. This evolution is encoded in a persistence diagram, as defined below.

Definition 2.3. The *persistence diagram* of a filtered dual graph $G_P^{t_1} \subset \cdots \subset G_P^{t_k}$ is a set of points

$$\mathcal{D} = \{(b_1, d_1), \ldots, (b_k, d_k)\},$$

thought of as recording "birth" and "death" times of each connected component in the filtration. For each i, set $b_i = t_i$ and set $d_i = t_{i'}$, where $i' \geq i$ is the minimum index such that the component of w_i in $G_P^{t_i}$ contains a node whose index is less than i. (Since no such index exists for $i = 1$, we put $d_1 = \infty$.)

The persistence diagram is a set of points (b, d) with $b \in \mathbb{R}$, $d \in \mathbb{R} \cup \{\infty\}$ and $0 \leq b \leq d$, so can be drawn in a unit square above the diagonal, with ∞ represented at $d = 1.1$ (say), as in Figure 3. In general, a persistence diagram could be a *multiset*, containing repeated points. In our application, we have assumed that the t_i values are distinct, so the points in $\mathcal{D}$ have multiplicity one.

Persistence diagrams are a ubiquitous summary statistic used in TDA, with existence and uniqueness results in more general settings going back to [60]. The precise, general definition of a persistence diagram relies on a decomposition theorem from representation theory known as Gabriel's theorem, which was proved in [36], with the connection to topological data analysis first observed in [12]. When representing a persistence diagram one typically omits points along the diagonal, as these represent "trivial" topological features.

Remark 1.

1. The persistence diagram defined here describes the evolution of the zeroth-homology of the filtered dual graph; this is the so-called *degree*-0 *persistent homology* of the filtered dual graph. Likewise, persistent homology in higher degrees can be computed, tracking appearances and disappearances of higher-dimensional topological features such as loops and voids. We hope to explore first homology in future work.
2. The evolution of connected components in a filtered graph (or in more general filtered spaces) can be summarized by a richer invariant called a *merge tree*, which is closely related to the persistence diagram (see [52, 24]). The standard metrics on the space of merge trees (Gromov-Hausdorff distance or merge tree interleaving distance) are not computationally tractable [3], so we conduct our analysis at the level of persistence diagrams. We remark that in the standard TDA setting of Vietoris-Rips filtrations of point clouds, merge trees are equivalent to dendrograms obtained through single linkage hierarchical clustering [13]. Since our filtration is not Vietoris-Rips, our method can not be simply rephrased in the SLHC framework.
3. The definition above entails a convention sometimes referred to as the *Elder Rule* in the persistent homology literature: when two components merge, the component with the more recent birth time is annihilated. See [32] for a discussion of this convention, as well as the more recent theoretical justification provided in [24].

The entire pipeline described so far is summarized in Figure 3. In the end, this pipeline reduces a very complex object (a geographical partition of a state together with a function describing, e.g., election data) to a simple, if somewhat abstract, summary statistic (a persistence diagram encoding the evolution of connectivity of districts according to the function). It is the thesis of this paper that this large reduction in complexity preserves, and in fact illuminates, interesting geographical

and political properties of the plans. Our study of persistence diagrams for redistricting ensembles will require some quantitative tools described in the following subsections.

2.5. **Wasserstein distances.** Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be a pair of persistence diagrams. We introduce a standard family of distance metrics used to compare persistence diagrams.

Definition 2.4 (Wasserstein p-distance)**.** Let $p \in [1, \infty]$ and let (b, d) be a point in a persistence diagram. Define the *diagonal p-distance* $\Delta_p(b, d)$ to be the ℓ_p distance from (b, d) to the nearest point on the *diagonal* $\{(x, x) \mid x \geq 0\}$. A *partial bijection* between $\mathcal{D}_1$ and $\mathcal{D}_2$ is a bijection ϕ from a subset of $\mathcal{D}_1$ onto a subset of $\mathcal{D}_2$. We use the notation $\phi : \mathcal{D}_1 \nrightarrow \mathcal{D}_2$ for a partial bijection, $\mathrm{dom}(\phi)$ for the domain of ϕ and $\mathrm{im}(\phi)$ for its image. For $p < \infty$, define the *p-cost* of a partial bijection ϕ by

$$(\mathrm{cost}_p(\phi))^p = \sum_{(b,d)\in \mathrm{dom}(\phi)} \|(b, d) - \phi(b, d)\|_p{}^p + \sum_{(b,d)\notin \mathrm{dom}(\phi)} \Delta_p(b, d)^p + \sum_{(b',d')\notin \mathrm{im}(\phi)} \Delta_p(b', d')^p.$$

This extends to a $p = \infty$ cost in the usual way, as a supremum:

$$\mathrm{cost}_\infty(\phi) = \max\left\{ \max_{(b,d)\in \mathrm{dom}(\phi)} \|(b, d) - \phi(b, d)\|_\infty, \max_{(b,d)\notin \mathrm{dom}(\phi)} \Delta_\infty(b, d), \max_{(b',d')\notin \mathrm{im}(\phi)} \Delta_\infty(b', d') \right\}.$$

Finally, define the *Wasserstein p-distance* between $\mathcal{D}_1$ and $\mathcal{D}_2$ by

$$d_p(\mathcal{D}_1, \mathcal{D}_2) = \inf_{\phi:\mathcal{D}_1 \nrightarrow \mathcal{D}_2} \mathrm{cost}_p(\phi). \tag{1}$$

The Wasserstein p-distances are naturally described in the language of matching features. The goal is to match points in the diagram, which we now recall represent persistent topological features. A partial bijection ϕ defines such a matching of features; those points not in the domain or range of ϕ are considered to be matched with a "trivial feature" on the diagonal (i.e., a feature with zero lifespan). The p-cost $\mathrm{cost}_p(\phi)$ of matching ϕ measures how far points have to be moved in L^p distance, and the goal is to find a matching which minimizes this cost. For a given pair of persistence diagrams, we will refer to the partial bijection ϕ that achieves the infimum in (1) above as an *optimal L^p matching*. This will be referenced later when comparing Fréchet means for different ensembles or vote data.

Traditionally, the $p = \infty$ Wasserstein distance, also known as *bottleneck distance*, has been the standard choice for comparing persistence diagrams in topological data analysis. It was one of the first metrics of this type to be introduced [22] and its matching cost is the simplest to interpret: the cost of matching (b, d) to (b', d') is whichever is larger, the difference in birth times or the difference in death times. It was moreover shown in that paper that the bottleneck distance enjoys stability under perturbations of filtrations generating persistence diagrams (this result will be stated and used in Section 6 below). The use of bottleneck distance was later justified at a more theoretical level by the famous *Isometry Theorem*. This theorem, established in [15] and [42], states that the bottleneck distance between persistence diagrams coincides with the so-called *interleaving distance*, a distance which is defined purely in terms of morphisms between persistence modules using the language

of category theory. This category-theoretic interpretation of bottleneck distance was extended to the other Wasserstein distances in the recent paper [9]. On the computational side, the optimal partial bijections required by the Wasserstein distances can be computed using the Hungarian algorithm (see, e.g., [16]), a polynomial-time algorithm for solving the assignment problem required to find an optimal matching.

2.6. **Fréchet means.** Let (X, d) denote an arbitrary metric space. For a finite sample $x_1, \ldots, x_n$ from X, define a discrete probability measure $\frac{1}{n}\sum_{j=1}^{n} \delta_{x_j}$, where δ_{x_j} the Dirac measure at x_j. The associated *Fréchet functional* $F : X \to \mathbb{R}$ is defined by

$$F(x) = \frac{1}{n}\sum_{j=1}^{n} d(x, x_j)^2.$$

A *Fréchet mean* of the sample is a minimizer of this Fréchet functional, i.e., a *barycenter* for the collection of points. Below, we will make heavy use of barycenters to summarize ensembles of diagrams.

Metric barycenters are guaranteed to exist uniquely in spaces of nonpositive curvature. We are not in that setting, but are nonetheless able to leverage nice metric features to get a well-behaved barycenter construction. With the $p = 2$ Wasserstein distance, Turner et al. show that the space of all persistence diagrams is an Alexandrov space of nonnegative curvature: a complete, separable geodesic space with curvature ≥ 0 in the sense of comparison triangles [58]. Despite the positive curvature, these metric spaces have helpful structure: they admit a notion of a tangent space at each point, which can be endowed with an inner product (see [44] for general theory). Moreover, semiconcave functions have well-defined gradients. It is shown in [58] that the Fréchet functional is semiconcave, motivating its optimization via gradient descent.

The [58] authors then propose the following algorithm. For persistence diagrams $\{\mathcal{D}_1, \ldots, \mathcal{D}_m\}$, suppose that $\mathcal{D}^{(i)}$ is the current Fréchet mean candidate. An update is performed as follows:

1. Compute optimal partial bijections $\phi_j^{(i)} : \mathcal{D}^{(i)} \nrightarrow \mathcal{D}_j$;
2. Extend $\phi_j^{(i)}$ to a function $\phi_j^{(i)}\mathcal{D}^{(i)} \to \mathcal{D}_j \cup \{(x, y) \in \mathbb{R}^2 \mid x = y\}$ by declaring $\phi_j^{(i)}(b, d)$ to be the point on the diagonal nearest to (b, d) for each $(b, d) \notin \mathrm{dom}(\phi_j^{(i)})$;
3. For each $(b, d) \in \mathcal{D}^{(i)}$, update to the arithmetic mean of the set of images $\phi_j^{(i)}(b, d)$;
4. Let the new Fréchet mean candidate $\mathcal{D}^{(i+1)}$ be the updated point collection.

This process is shown to converge, at least to a local minimum of the Fréchet functional. As this is an iterative algorithm, the local minimum found depends on the seed $\mathcal{D}^{(1)}$, that is, the initial Fréchet mean candidate. To mitigate this dependence in our analysis, we run the algorithm multiple times, using each persistence diagram in our collection as a seed, and take the local minimum with the lowest Fréchet functional value as our final Fréchet mean.

Remark 2. For our analysis, we seeded the iterative Fréchet mean algorithm at each of the 1000 districting plans in the ensemble; from across these runs, we took the local minimum with the lowest Fréchet functional value. Computing a Fréchet mean for each seed took about 10 minutes for the Congressional ensembles, while the Pennsylvania state House analysis ($k = 203$) took the longest, at 23 hours (on

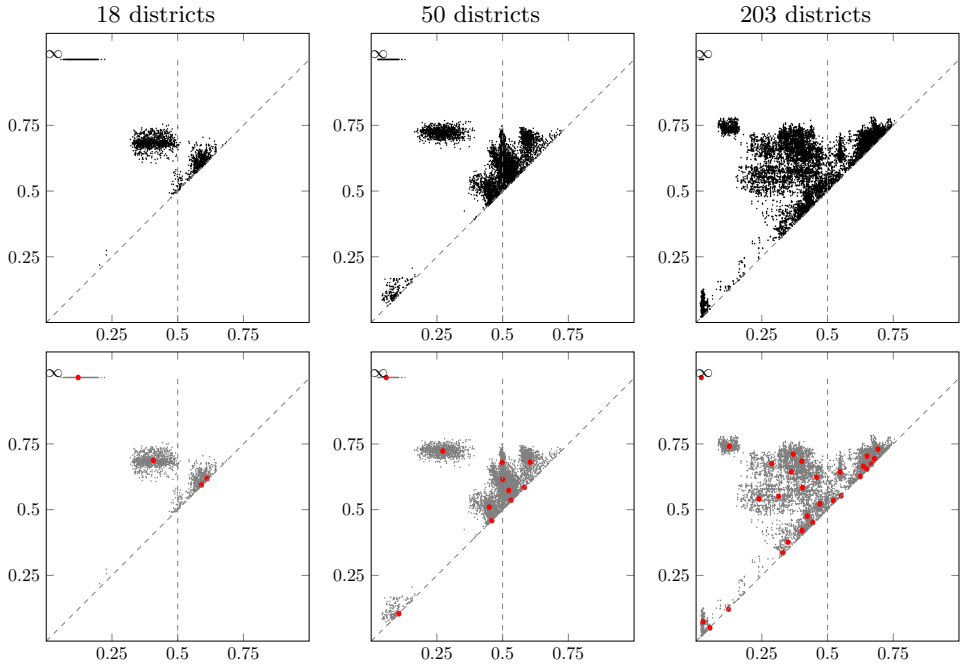


FIGURE 4. Pairing each plan in a districting ensemble with PRES16 vote data gives an ensemble of persistence diagrams. These plots show overlays of the diagram ensembles for $k = 18, 50, 203$ respectively. In the second row, we add the Fréchet mean $\mathcal{F}$ for each ensemble, shown in red, representing an "average" persistence diagram for the ensemble.

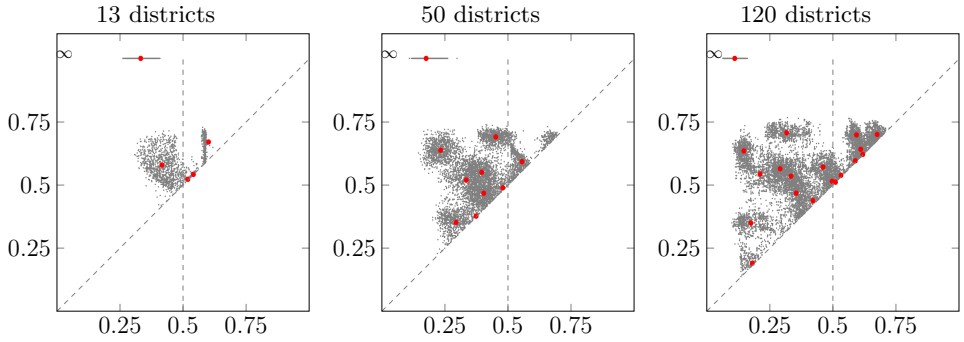


FIGURE 5. Overlaid persistence diagrams with Fréchet mean (in red) for each of the NC ensembles (with PRES16 vote data).

a single node of a high performance cluster). We regard the use of so many runs as valuable despite the significant increases in computational cost. Ultimately, a successful seed will produce a local minimum that is near the global minimum, so that the mean is as good a fit as possible to the ensemble; importantly, a failed

run may not only have too high a Fréchet functional value but may have the wrong number of clusters to enable a good summary. In our experiments, we found that a good heuristic for choosing a starting point for the Fréchet mean is to choose a plan with high *total persistence* (defined as the sum of the persistence over the features).[3]

2.7. **Computational details.** The main code base used for this paper is available on GitHub [4]. The ensembles used in this paper were generated using the open-source package GerryChain developed by the Metric Geometry and Gerrymandering Group [59]. GerryChain implements numerous Markov chains on $\mathcal{P}_{k,\epsilon}(G)$ for generating ensembles of districting plans, including the recombination and flip chains discussed in this paper. For the main ensembles used in Sections 3.1 and 3.2, we sampled every 50th plan from an ensemble of 50,000 plans generated using a ReCom chain with $\epsilon = .02$ as the maximum population deviation. The Congressional and state Senate ensembles were generated from dual graphs constructed from the precinct shapefiles available from MGGG [50]. For the state House ensembles, more granularity is required in order to get satisfactory population balance, so we used census block dual graphs. In the case of North Carolina, the state releases electoral data at the block level for redistricting purposes. For Pennsylvania, election results were disaggregated from the precinct level down to blocks using the geospatial data package called MAUP [49]. The persistence diagrams were generated using the library GUDHI (https://gudhi.inria.fr) [45].

3. **Methods.** We propose three types of analysis enabled by persistent homology on ensembles of districting plans.

3.1. **Scale and zoning.** The winner-take-all plurality structure of the American districting system introduces significant non-linearity to the relationship between votes and seats, whose patterns we seek to capture. The key question here asks how the fine-grained vote data by precinct (see Figures 1 and 11) translates into an outcome at the district level. We would like to isolate the effects of district *size* on the geography and partisan properties of redistricting plans. A scale study was carried out in [55] by a direct comparison of the district size to the range of observed partisan outcomes in an ensemble, using recent Pennsylvania voting patterns. In this section, we use topological signatures to probe scale effects further. The main finding is that at each scale, the state decomposes into zones of partisan support whose anchor points can be detected by persistent homology techniques.

Before we discuss the application of our method to ensembles of maps, we first illustrate the information captured on individual plans. Figure 1 shows the currently enacted Congressional and Senate districts for Pennsylvania, colored by Trump vs. Clinton share (Pres 2016). Comparing to the choropleth at the top of Figure 1, we can see that at the Congressional scale, the Democratic voting in Erie and State College is outweighed by the surrounding Republican votes, while the Senate scale allows those cities to pull their districts more strongly. Thinking in terms of peaks of Democratic vote share, Pittsburgh, Harrisburg, and Philadelphia are all pronounced

[3] Note that for the scale and zoning analyses, the Fréchet mean is used only to cluster features. As a result, the exact value of the Fréchet mean is not as important as the number and the geographic signature of clusters. When comparing elections, on the other hand, the value of the Fréchet mean matters, which puts more of a premium on running the minimizing algorithm from many seeds.

[4] https://github.com/mggg/TDA-redistricting

peaks on the precinct level, with the corresponding Congressional districts (18, 10, and 3, respectively) also showing up as peaks on the district level.

Homological persistence will capture a *relative height* of these peaks in the following way: Each district which is more Democratic than all its neighbors shows up as a point (b, d) in the persistence diagram, where b is precisely its Republican vote share. A value of d on the 'death' axis means that any path from that district to a more Democratic one must pass through a district whose Republican share is at least d. Thus the persistence $p = d - b$ is a kind of relative intensity of Democratic voting. For example, District 10 is only 9 percentage points more Democratic than neighboring District 11, which in turn is adjacent to District 6, a more Democratic district than District 10. So the persistence of its feature is 9. Contrast this with District 18: any path from District 18 to a more Democratic district must pass through either District 13 or 15, each of which is more than 36 points less Democratic than District 18. This gives Pittsburgh a persistence of 36. It is clear that a Senate plan, with a larger number of smaller districts, will tend to have more peaks, such as Erie in the northwest corner; the greater number of features shows up in the Senate plan's persistence diagram, together with the solutions to the minimax problems that describe their relative height.

So far we have discussed the interpretation of persistence for specific plans, but of course different plans may have different structure. The use of Fréchet means will allow us to examine how much common structure there is across the plans of a districting ensemble.

Step 1: Overlay. Start with a state, a number of districts k, and some fixed vote data such as from a recent election. We generate an ensemble of one thousand plans (see Section 2.7) and form diagrams $\mathcal{D}_1, \ldots, \mathcal{D}_{1000}$. As a first step, we can overlay the thousand diagrams (taking the unions of the sets $\{(b, d)\}$) as the top row of Figure 4 does for Pennsylvania ensembles. To tease apart the structure of these plots, we compute the Fréchet mean $\mathcal{F}$ of the ensemble as in the next row of Figure 4. This gives us one summarized view of the ensemble as an "average" persistence diagram.

Step 2: Marking. Next, we want to extract a *marking* of the individual persistence diagrams. Where k is the number of districts, we can choose $\ell \leq k$ points in $\mathcal{F}$ that are deemed sufficiently far from the diagonal to be significant. Then for every diagram $\mathcal{D}_i$ in our ensemble, we match the points in $\mathcal{D}_i$ to the chosen points in $\mathcal{F}$ using an optimal L^2 matching (see Section 2.5), inducing a partial labeling of the individual persistence diagrams by elements of the Fréchet mean. Returning to the overlaid diagram, we can now create individual Fréchet point plots corresponding to each significant feature of $\mathcal{F}$, with the property that each diagram $\mathcal{D}_i$ contributes at most one point to this plot.

Step 3: Localization. For a significant feature of $\mathcal{F}$, many of the $\mathcal{D}_i$ have a corresponding feature, picking out a district that initiated the connected component in the filtration of the district dual graph. We can then measure how often in the ensemble each geographical unit belonged to that district. This gives us a *heat map* that shows us where the feature is located in the state. Figure 8 shows Fréchet point plots and corresponding precinct heat maps for PA Congressional and Senate districts. (Supplemental Figure 20 uses a census block heat map for the House ensemble.)

Step 4: Zoning. If the filtration is in terms of one party's share (here, Republican), we can then zone the state into clusters of districts won by the other party (here,

Democratic), as follows. There is a one-to-one correspondence between persistence features in the northwest quadrant of the diagram $\mathcal{D}$ (i.e., $b < .5, d > .5$) and clusters of Democratic-won districts. The birth time ensures that the cluster is anchored by a Democratic-won district and the death time ensures that it is separated from other clusters by Republican-won districts. For each localized point in $\mathcal{F}$, we consider northwest-quadrant points marked by it in the $\mathcal{D}_i$, representing Democartic clusters of districts. We can map these clusters and report their average number of districts.

When studying scale effects, we are interested in which areas represent peaks at each scale (usually cities from the Democratic point of view). Crucially, cities only show up as individuated peaks in Democratic vote strength when they are separated from other cities by redder *districts*—a property that it would be hard to see with one's eyes.

Note that there is no a priori reason that the localization step should produce a geographically specific position for each Fréchet feature. Picking out an identifiable location at this step is a signal that there is consistency across the ensemble in the correspondence of geography and aggregated voting. The method has succeeded in zoning the state at a given scale if the heat maps are geographically coherent. If localization succeeds, the workflow lets us refer to the Fréchet features by geographical names and expected number of Democratic-won districts for a given election: the Philadelphia zone (5.6 Congressional districts), the Pittsburgh zone (1.2 Congressional disricts), and so on.

3.2. **Comparing elections.** In the previous section, we studied ensembles of plans relative to some fixed vote pattern. Our second type of analysis will involve varying the vote data. We ask several questions. Do the Fréchet means themselves tell us something interesting about the relationship in voter preferences between elections? How robust are the patterns learned above to a shift in the chosen electoral baseline?

For a given state and district size, we again use an ensemble of one thousand districting plans. We then vary the vote data (but not the ensemble of plans) to produce multiple ensembles of diagrams, with an associated Fréchet mean plot for each vote pattern. Suppose that $\mathcal{F}$ and $\mathcal{F}'$ are Fréchet means for the same plans and two different elections. We can pair points from $\mathcal{F}$ to $\mathcal{F}'$ (by either a geographic pairing or an optimal L^2 matching) to provide a summary of electoral difference.

For a pair of features that is matched, the direction of displacement carries information. Suppose the mean point is (b, d) in $\mathcal{F}$ and (b', d') in $\mathcal{F}'$. If $b > b'$, this zone was anchored by a district that is more Republican in the first election than the second. If $d > d'$, then the area surrounding the peak was more Republican in the first election than the second. Putting these together can give a summary of the voting performance of a city relative to its exurban ring at district scale.

If a feature in $\mathcal{F}$ is not matched (i.e., is matched to the diagonal) in $\mathcal{F}'$, this gives us a different kind of information. This represents a substantial difference in the basic geography of representation under these vote data, picking out a district cluster that represents a peak in one election but not the other.

3.3. **Signals of gerrymandering.** The examination of zoning and electoral shifts described above is not targeted at identifying gerrymandering, but at understanding the tendencies of partisan-neutral districting. Since the detection of gerrymandering is a difficult and high-profile problem, we now turn attention to whether persistent homology can flag properties of plans that are designed to extract advantage for one

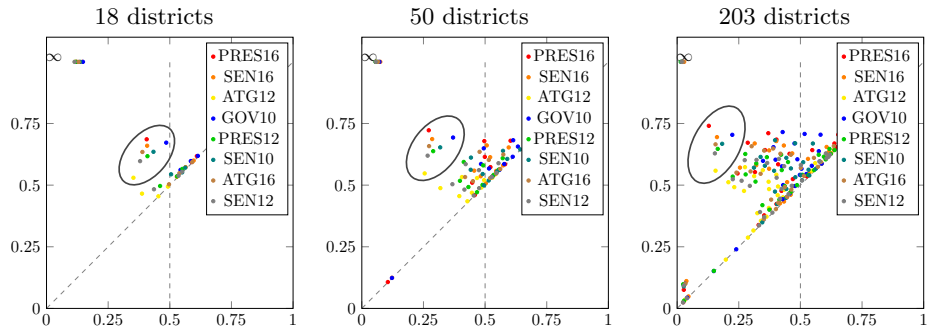


FIGURE 6. Overlaid Fréchet means for various elections in PA. In all three levels of redistricting, the geographic localization is strong enough to label the Pittsburgh feature (circled).

major party over the other. We are able to use this in our case studies to find several instances of notable relationships between persistence data and partisan advantage.

Given a state, number of districts k, and vote pattern, our first step is to generate two ensembles: one creating an unusually large number of Republican-majority districts and a second favoring Democrats. In order to avoid flagging extremely competitive plans, we will identify a *safe seat* for Party A when the party's vote share exceeds .53. (Noting, of course, that for practical gerrymandering purposes, there may be a more complicated calculus for trading off safety with win maximizaiton.)

To generate an ensemble of plans biased towards Party A, we use the ReCom algorithm with a Metropolis-style weighting variant. When a new plan P is proposed by the algorithm, it is immediately accepted if the number of safe seats does not decrease. A standard Metropolis weighting with respect to a score σ is to choose a coefficient β and accept a plan scoring worse with probability $e^{-\beta\Delta\sigma}$. We will do that here: if the safe seat count decreases by Δs, then we accept P with probability $e^{-2\Delta s}$.[5] The purpose of the Metropolis rule is to avoid getting stuck at local optima, while favoring better plans long-term. We use this algorithm to generate two ensembles of one thousand plans each: a Democratic-favoring ensemble and a Republican-favoring one. To illustrate the effectiveness of the weighted Markov chains, see Figure 7, showing a difference in expectation of 3-4 seats out of 50.

From the two biased ensembles we then generate Fréchet point plots and associated heat maps, then compare the Fréchet means with either a geographic or optimal L^2 matching (see Section 2.5 for the definition of an optimal L^2 matching). As before, we have no guarantee that these two matching methods agree, but in our case studies below we find that they do.

The aim of this analysis is to find and interpret geographical correlates of partisan advantage—not to classify entire plans as gerrymandered or not. For that, we are unlikely to do better than simply comparing the number of seats won under various electoral assumptions to a neutral ensemble.

[5]We note that for standard recombination moves, this can only occur with $\Delta s = 1$, corresponding to a roughly 13.5% probability of acceptance. The standard protocol is to adjust β until you have a run with a decent acceptance rate and an ensemble that passes heuristic tests of convergence and quality.

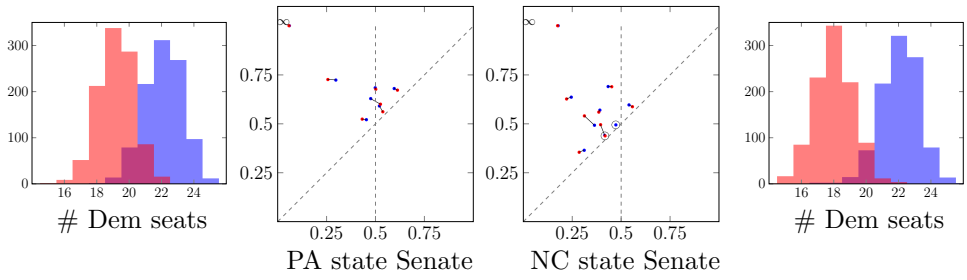


FIGURE 7. Comparison for the weighted ensembles of Senate plans, with Democratic-favoring in blue and Republican-favoring in red. The simple weighting adjustment to the Markov chain has successfully created a difference of 3-4 seats between the ensemble means. Despite the significant change in the partisan outcome, we see only small movement in the Fréchet means.

In our case studies below, we can identify a few Fréchet features which show a significant difference between the two ensembles, and we can make inferences about the reasons for these differences. However, on the whole there is surprisingly little difference between the persistence data for the two ensembles despite their very different partisan seat shares. This further emphasizes that persistent homology is not primarily useful for classifying plans, while reinforcing the narrative of robustly identifying relationships between geography and aggregated vote data.

4. **Case study: Pennsylvania.** Pennsylvania has 18 Congressional districts, 50 state Senate districts, and 203 state House districts. We have cleaned electoral data from eight recent statewide elections: two Presidential elections (2012, 2016), three U.S. Senate elections (2010, 2012, 2016), one Gubernatorial election (2010), and two Attorney General races (2012, 2016). The statewide party performance in these elections is shown here in Table 1.

	PRES12	PRES16	SEN10	SEN12	SEN16	GOV10	ATG12	ATG16
R %	47.29	50.35	51.05	45.44	50.72	54.52	42.58	48.57

TABLE 1. Overall Republican vote shares (with respect to two-party vote) for a range of recent statewide elections in Pennsylvania.

4.1. **Scale and zoning in PA.** We begin with partisan-neutral ensembles of maps for each choice of $k = 18, 50, 203$. We carry out the four steps of the scale analysis for each: overlay, marking, localization, and zoning. We find four zones of relative Democratic strength at the Congressional level (Figure 8, top), corresponding to Philadelphia, Pittsburgh, Scranton, and Harrisburg, in that order. Philadelphia and Pittsburgh are highly persistent while other two zones are weaker; only those two correspond to Democratic district clusters, comprising 5.6 and 1.2 districts, respectively (Supplemental Figure 24).

In the state Senate ensemble (Figure 8, bottom), we find seven zones: Philadelphia, Pittsburgh, Erie, Harrisburg, State College, Lancaster, and Scranton, in sequence. It is quite valuable to note that the order as well as the membership is

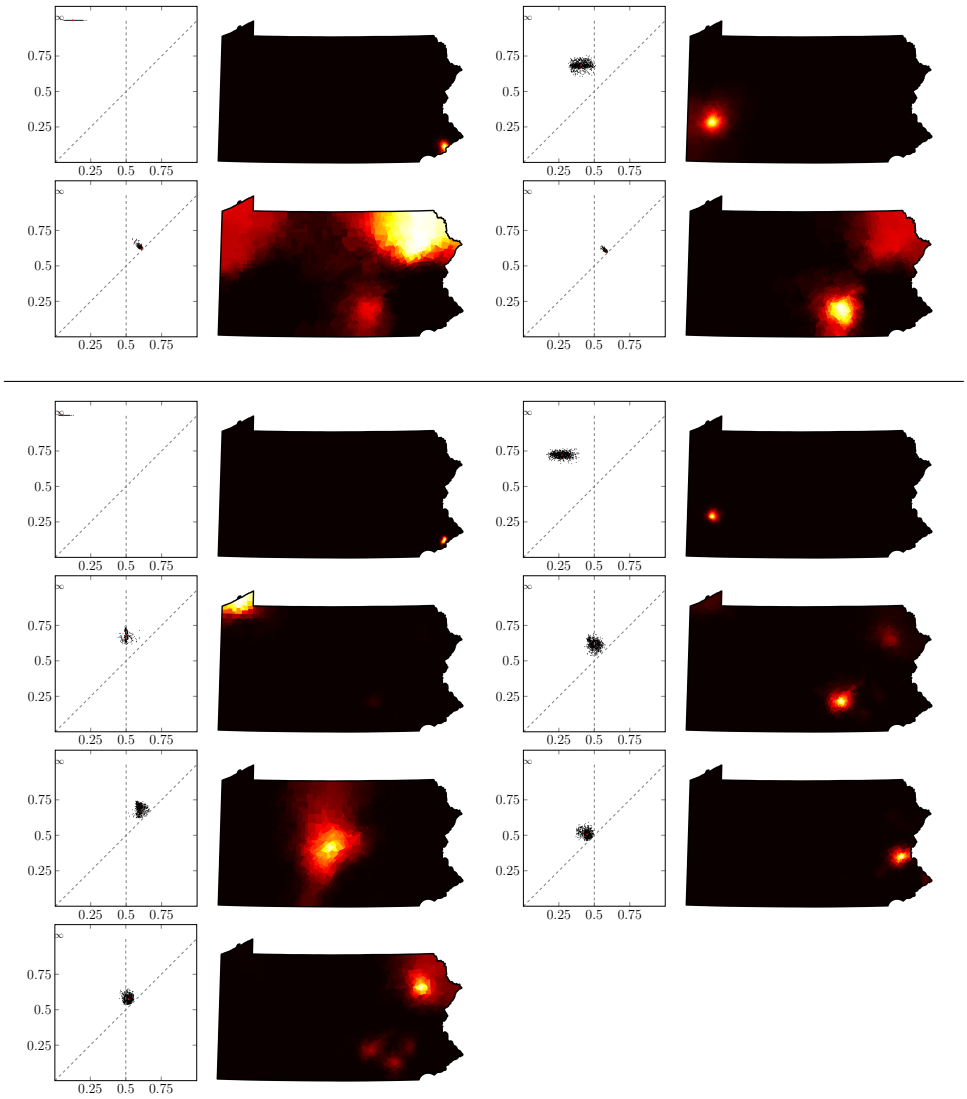


FIGURE 8. Geographical localization of Fréchet features in Pennsylvania Congressional and Senate plans ($k = 18, k = 50$) with respect to PRES16 voting. The Senate features fairly clearly capture the medium-sized cities of Pennsylvania.

different between Congressional and Senate zoning. Erie, Harrisburg, and Scranton anchor zones that will be competitive overall, while the State College zone favors Republicans and the Lancaster zone favors Democrats. The zoning analysis indicates 14.7 Senate districts in the Philadelphia zone and 3.4 in Pittsburgh; even at this scale, the others each contribute less than one district on average.

For the state House ensemble (Supplemental Figure 20), the degree of geographic coherence (localization) is weaker, but we can read off up to eleven zones in persistence order as follows: Philadelphia, Pittsburgh, Harrisburg, State College, Reading, Lancaster, Erie, Allentown, Scranton, York, Hermitage. And now all but Hermitage should expect to anchor clusters of Democratic-won districts (not just Democratic relative to their neighbors). We find 55.8 House districts in the Philadelphia zone—a massive bloc of the 203 seats.

4.2. **Comparing elections in PA.** The Fréchet means for the eight elections under consideration were plotted for each scale of redistricting in Figure 6. For the Congressional ensemble, every election has two highly persistent features in its Fréchet mean $\mathcal{F}$ (corresponding to Philadelphia and Pittsburgh). Five out of eight elections (SEN12, PRES12, SEN10, and most prominently ATG12) have a third feature away from the diagonal. By looking at the geographic data (not shown), we can tag all of these third features as belonging to a Scranton zone. The presence of a Scranton feature means that Scranton is separated from Philadelphia by a district redder than itself. Since the Scranton mean points always lie below $d = .5$, the intervening districts between Scranton and Philaldelphia are still Democratic-favoring.

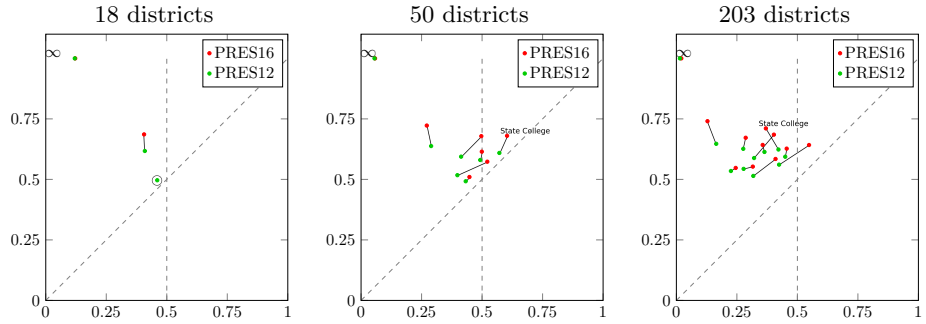


FIGURE 9. We isolate the Fréchet means for two successive Presidential elections in PA. The mean for PRES12 has a third off-diagonal feature at Scranton, while PRES16 does not. A uniform partisan swing would show each red point displaced by $(.03, .03)$ relative to its paired green point, so the diagram summarizes the non-uniformity of the electoral shift.

We isolate two elections of interest, PRES12 and PRES16, for a head-to-head comparison in Figure 9. We draw lines to indicate the geographic pairing between the top eleven most persistent features (this pairing differs from the optimal L^2 matching between the Fréchet means). Recall that the Republican statewide share is .4729 in PRES12 versus .5035 in PRES16, a difference of about 3 percentage points. If the elections were related by a *uniform partisan swing* (in which each geographic unit has its vote share additively perturbed by the same amount), then each b_i and each d_i would increase by the same amount, shifting the whole diagram to the northeast with a $(.03, .03)$ translation. This is not at all what we observe. We note that the Pittsburgh feature actually shifts in a Democratic direction at all three scales, most notably at the smallest (House) scale. The difference is in the middle of

the state, which separates Pittsburgh from Philadelphia. There, the Trump election has a significantly higher d value, indicating the reddening of the mid-state districts at all three scales. The State College feature is especially interesting, and is marked in the state Senate and state House plots. In the state Senate ensemble, the State College feature more or less adheres to the uniform partisan swing direction: b and d are both lower for PRES16 than PRES12. However, for the state House election, the trend reverses. This is because state House districts are small enough to separate the city of State College (which was more Democratic in 2016 than 2012) from the surrounding area (which was less Democratic in 2016 than 2012). Indeed, this effect can be seen in the geographic heat maps for this feature in Figures 8 and 20—the State College feature is much more diffuse in the state Senate ensemble than in the state House ensemble. In the Congressional ensemble, the Scranton feature is present in PRES12, but not PRES16, telling us that the level of Trump preference in the Scranton district is not significantly distinct from its surroundings (in contrast with Romney support).

4.3. **Signals of gerrymandering in PA.** We now turn to the party-biased ensembles. We focus on the state Senate and the PRES16 vote for an illustrative discussion.

The matching between the two biased ensembles is shown with line segments in Figure 7, helping organize the overlaid plots in Figure 21. All the off-diagonal points were successfully paired, and the geographic pairing and optimal L^2 matchings were identical, indicating that the persistence structure of the two ensembles was very similar despite the significant difference in the overall number of Republican districts.

It is instructive to compare the Pittsburgh and the Harrisburg features in the two ensembles, which show opposite effects. The Harrisburg district is quite competitive, and the Republican ensemble shows the effect you would expect: the district is several percentage points more Republican than in the Democratic ensemble—enough to secure a Republican win in most cases. On the other hand, an overall Republican advantage is correlated with a more *Democratic* Pittsburgh district; this is the well-known phenomenon called "packing," where wastefully high vote totals in some districts lead to overall disadvantage. In interesting contrast to Harrisburg, the Erie cluster was very rarely anchored by a "safe" district for *either* party, despite the weighting of the algorithm towards safe seats. Finally, the Republican-favoring ensemble clearly "cracks" Democratic support near Scranton (6th feature), while the Democratic-favoring plot resembles the neutral ensemble, indicating that in 1000 largely independent attempts, careful line-drawing near Scranton was not a frequent property of maps selected for local improvements in Democratic seats overall.

5. **Case study: North Carolina.** We turn to North Carolina, which has 13 Congressional districts, 50 state Senate districts, and 120 state House districts. We have cleaned data for seven statewide elections in North Carolina: Presidential elections from 2012 and 2016, the US Senate elections from 2010, 2014 and 2016, and the Gubernatorial elections from 2012 and 2016. The statewide two-way vote shares for these elections can be found in Table 2.

5.1. **Scale and zoning in NC.** The Fréchet point plots and accompanying heat maps using vote data from the 2016 Presidential election are shown in Figure 12 and

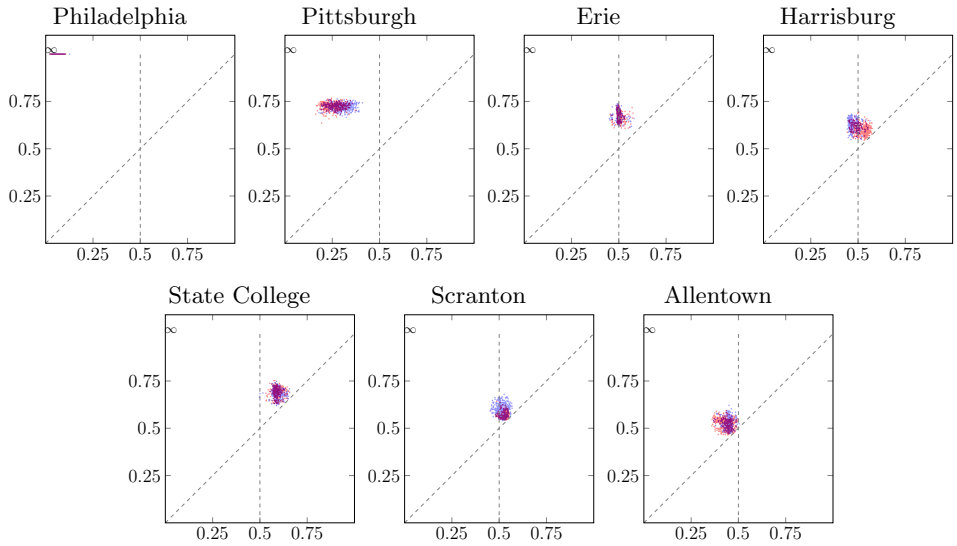


FIGURE 10. Comparison of point plots for the successive Fréchet features in the PA Senate ensembles that are biased for Democratic and Republican safe seats, respectively. Note the contrast between the rigidity of the Erie feature (similar placement for both ensembles) and the manipulability of the Harrisburg feature, though both anchor competitive districts.

	PRES12	PRES16	SEN10	SEN14	SEN16	GOV12	GOV16
R %	51.08	51.98	56.02	49.17	53.02	55.87	49.95

TABLE 2. Overall Republican vote shares (with respect to two-party vote) for a range of recent statewide elections in North Carolina.

Supplemental Figure 22. Geographic localization shows clear success in all three cases.

At the Congressional level, North Carolina admits three clear zones for this vote data: Durham, Charlotte, and Asheville—interestingly, Congressional districts are large enough that the Asheville district contains everything in the western tip of the state. A fourth zone anchored in Winston-Salem is sometimes present, and there are indications of a possible fifth zone stretching south from Fayetteville. These geographically more diffuse zones are often quite a bit more persistent than the corresponding zones in Pennsylvania.

In the state Senate ensemble, as with the Congressional ensemble, two highly persistent features appear near Durham and Charlotte. In contrast with the Congressional ensemble, the Asheville zone is now more geographically localized and is anchored by a Democratic-won district ($b < .5$) in all maps. It is notable that the fourth feature is localized near the fourth from the Congressional ensemble, but a close inspection shows that it has shifted from Winston-Salem to the nearby city of Greensboro, which is bluer but smaller. For the first time, a rural area appears

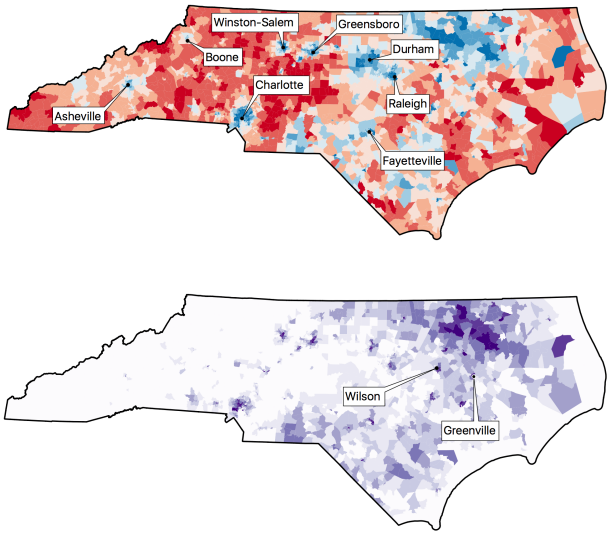


FIGURE 11. Maps of North Carolina. Top: results from the 2016 Presidential election (blue for Democratic, red for Republican). Bottom: North Carolina has a significant rural Black population, unlike Pennsylvania. Greenville and Wilson are the largest towns in the Northeast, but they are not the hubs of Black population.

as a feature: the heavily African-American northeast of the state anchors the sixth Fréchet feature, and not the medium-sized cities of Rocky Mount, Greenville, and Wilson that are nearby.

Raleigh is represented as a seventh Senate zone: a feature with $b < .5$ by a substantial amount, but low persistence. This represents a cluster anchored by very Democratic Senate districts, but that quickly merges with the Durham cluster. Its low persistence is an indicator that it is is not a completely independent zone from Durham.

Moving from the Senate to the House scale, we get new (low-persistence) zones in the 9th and 10th positions corresponding to Wilmington and Boone. This stands in contrast to the PA case, where the House added four medium-persistence features compared to the Senate. In particular, none of the small towns in the Northeast of North Carolina such as Greenville or Wilson are picked out in these plots. Rather, they are embedded in the large Northeast zone. This is not solely attributable to size: each of Greenville, NC and Reading, PA has the population of 1–1.5 House districts in its respective state. Rather, it reflects a difference in the degree of urban concentration in the Democratic vote.

Since TDA is comparing vote totals aggregated to districts, it is not automatic that it would detect this rural Democratic strength in North Carolina, but its signature is unmistakable at the Senate and House scale, and already visible in Congressional ensembles.

5.2. **Comparing elections in NC.** This time, we will compare a larger set of elections: all seven statewide elections in our dataset, ranging from 2010 to 2016.

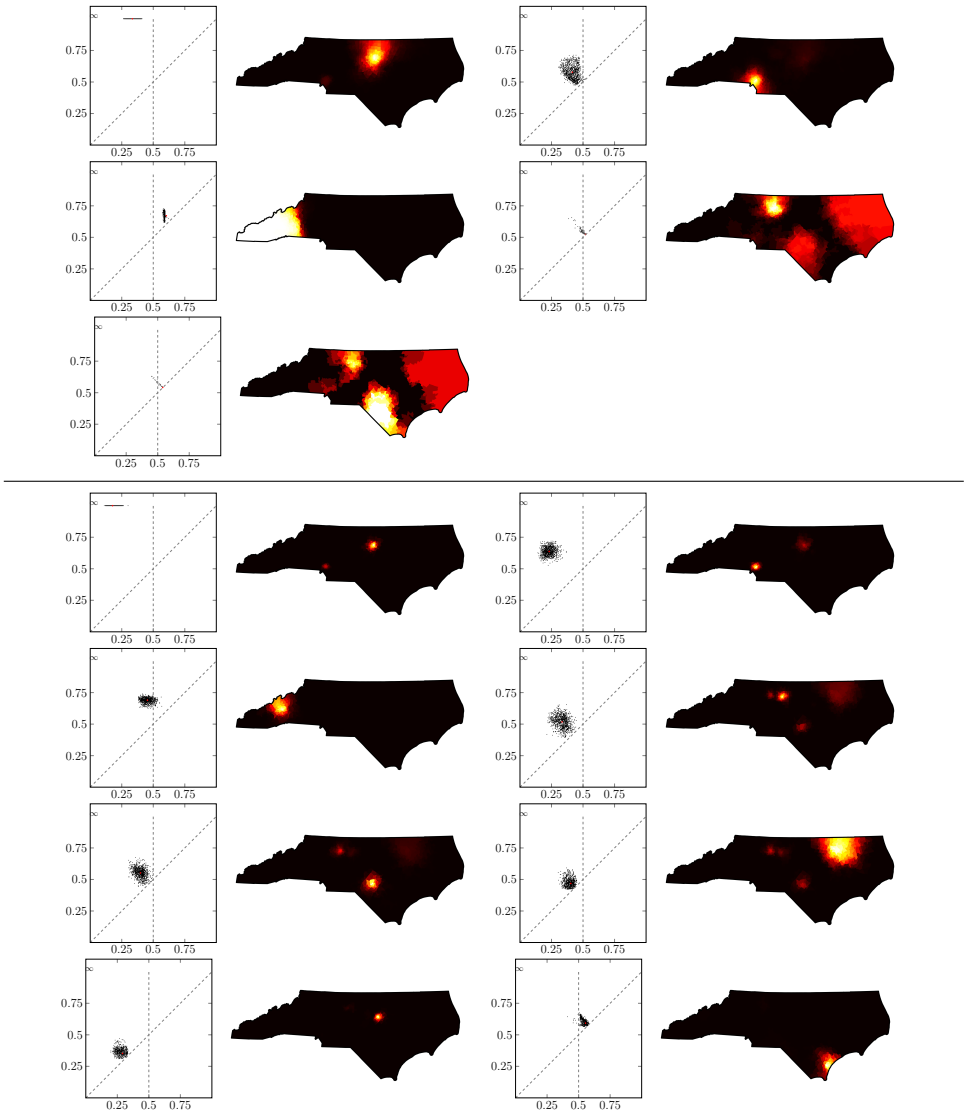


FIGURE 12. Geographical localization of Fréchet features in North Carolina Congressional and Senate plans ($k = 13, k = 50$) with respect to PRES16 voting.

Since the association of regions with Fréchet mean points was successful in the last section, we can again use the plots to summarize whether peaks of Democratic support swing in sync with their periphery between elections.

For the most persistent feature (the Durham zone, at $d = \infty$), the spacing shows you the difference in Democratic support between elections. If voting preferences swung linearly and uniformly around the state, the displacement of the colored dots would be diagonal, with both coordinates displacing by the same amount as the first feature. So we learn something nontrivial from this summary diagram: vote patterns in Charlotte and surroundings swing much more linearly than in Asheville.

In particular, Donald Trump overperformed both in heavily Democratic Asheville and (even more) in its Republican periphery, relative to a uniform swing prediction. More generally, Asheville's partisan preference is unpredictably related to the rest of Western North Carolina.

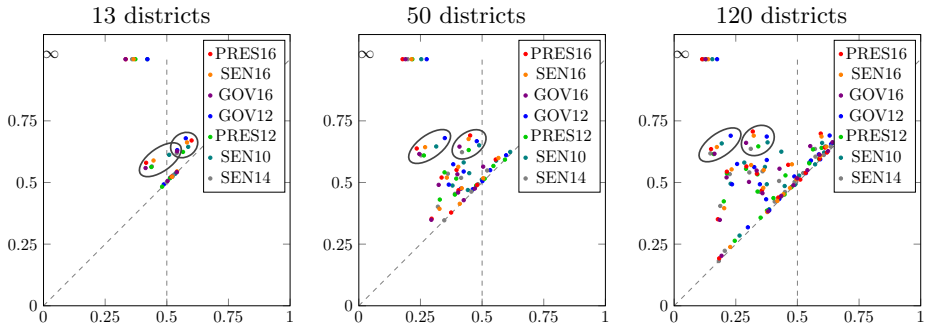
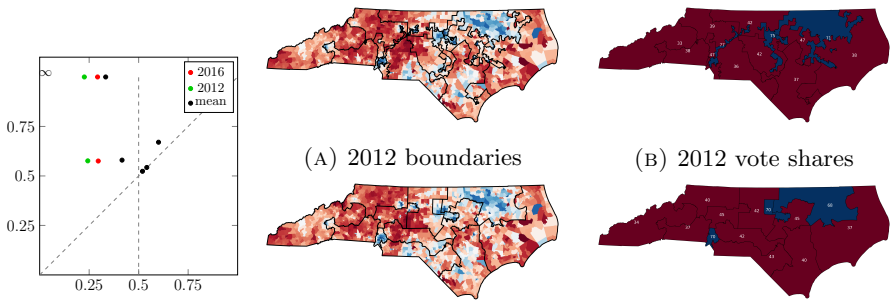


FIGURE 13. Overlaid Fréchet means for various elections in NC. In all three levels of redistricting, the geographic localization is strong enough to label the second and third features, shown circled, as Charlotte (left) and Asheville (right).

5.3. **Signals of gerrymandering in NC.** We begin by evaluating the two enacted Congressional maps for North Carolina (2012 , 2016) against the PRES16 vote pattern (Figure 14). Since both plans were identified by courts as Republican gerrymanders, we treat that as a reliable label. We overlay the persistence diagrams for the two enacted plans with the Fréchet means for the partisan-neutral ensemble of Congressional plans.



(A) 2012 boundaries (B) 2012 vote shares

(C) 2016 boundaries (D) 2016 vote shares

FIGURE 14. Comparing two enacted Congressional plans for North Carolina with the ensemble mean, against the PRES16 vote pattern in all three cases.

The features for the enacted plans have significantly lower b values than the ensemble, indicating especially Democratic districts anchoring two distinct clusters. The enacted plans each have only these two features, while the mean—and over 80% of the maps in the neutral ensemble—includes a third feature corresponding

to Asheville. This is because both enacted plans slice a large part of Asheville into a district that avoids the moderate partisan areas nearby; this district touches the Charlotte district and so becomes part of the Charlotte zone. Note the suggestively carved boundaries in Figure 14A,C. This finding squares with public perception: Asheville has been at the center of gerrymandering protests and activism for some time.

As in Pennsylvania, we close by examining the party-biased ensembles with respect to state Senate districts and PRES16 voting. Figure 7D shows the matching between the Fréchet means used to organize the Fréchet point plots in Figure 23 (again the geographic and optimal L^2 matchings agreed).

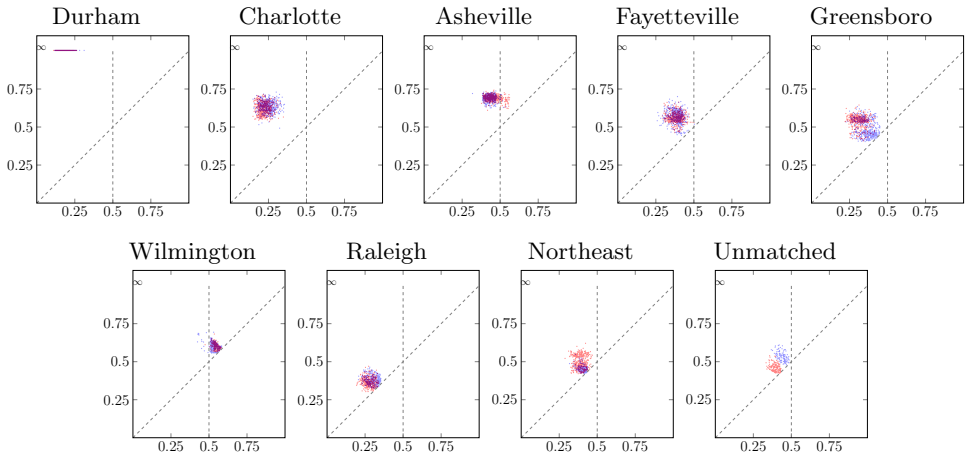


FIGURE 15. Comparison of point plots for the successive Fréchet features in the NC Senate ensembles that are biased for Democratic and Republican safe seats, respectively.

6. **Stability.** We now address the question of stability of our persistent homology signatures of vote and districting data. In order to do so, we appeal to a stability result of [22] which is well known in the field. In order to keep our exposition self-contained, we state the result in a special case which is relevant to our application.

Theorem 6.1 ([22])**.** *Let $G = (V, E)$ be a graph with f and g real-valued functions on V and corresponding persistence diagrams $\mathcal{D}_f$ and $\mathcal{D}_g$. Then*

$$d_\infty(\mathcal{D}_f, \mathcal{D}_g) \leq \|f - g\|_\infty ,$$

where $\|\cdot\|_\infty$ is the ℓ_∞ norm on functions $V \to \mathbb{R}$.

We then immediately obtain a stability result in our application, for the district dual graphs and Republican share filtration used above.

Corollary 6.2 (Stability under perturbations in vote data)**.** *The bottleneck distance between persistence diagrams associated to two elections and the same districting plan is bounded above by the largest change in Republican vote share at any district.*

6.1. **Theoretical stability under geographic perturbations.** For a fixed plan, we have a strong and easily interpretable stability result for bottleneck distance as elections shift. One could similarly ask if the persistence diagram summaries are robust under small changes in the plan itself. A truly general result in this direction is not possible, as is illustrated by the example in Figure 16.

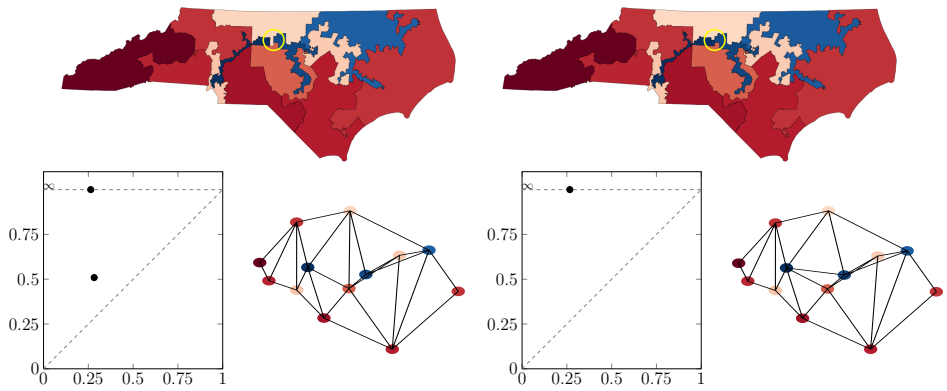


FIGURE 16. A small change to districts can cause a large change in persistence. In this example (closely modeled on the enacted NC Congressional plan from 2012) we have changed the assignment of only one geographic unit (a precinct of 3,300 people), and it removes a highly persistent off-diagonal feature.

In light of this fragility, some care must be taken to formulate a correct result on stability of bottleneck distance under changes in districting plan. We formulate such a result in this subsection, and relate it to experimental analysis in the following section using random perturbations.

Fix a state's unit dual graph $G = (V, E)$ and a districting plan $P = \{P_1, \dots, P_k\} \in \mathcal{P}_{k,\epsilon}(G)$. Let p be the population function and $T = p(V)/k$ be the ideal district size.

Definition 6.3. For districting plans $P, P' \in \mathcal{P}_{k,\epsilon}(G)$, P' is called a *(one-way) perturbation* of P if, after reindexing the districts as necessary, the plans agree in all but two districts $P_i \cup P_j = P'_i \cup P'_j$, and they differ only by the move of vertices $V_{ij} = \{v_1, \dots, v_\ell\}$, so that $V'_j = V_j \cup V_{ij}$ and $V'_i = V_i \setminus V_{ij}$. A *general perturbation* has vertices V_{ij} moved in one direction and V_{ji} in the other direction between the two altered districts.

A perturbation P, P' is called *graph preserving* if the labeled correspondence of districts induces an isomorphism of the district dual graphs.

For instance, the two North Carolina plans in Figure 16 are perturbations of each other by a single unit, which means that $|V_{ij}| = 1$. However, it is not a graph-preserving perturbation; the dual graphs G_P and $G_{P'}$ are not even abstractly isomorphic.

A ReCom step (as in our districting ensembles) can potentially produce a one-way perturbation of a plan, but this is extremely unlikely because population is swapped in both directions with very high probability. Moreover, ReCom steps are not always graph-preserving: in re-splitting the combined population of two

districts in an arbitrary connected fashion, they can alter district adjacency with the surrounding nodes.[6] A simpler Markov chain method for generating ensembles of plans by single-unit flips is studied here.

A *flip move* on a plan P is performed by reassigning a single vertex from one district to another in a manner that preserves connectedness for all districts. It therefore necessarily induces a one-way perturbation, and frequently a graph-preserving perturbation. A Markov chain can be defined by randomly selecting a vertex to reassign at each step. Taking sequences of flip steps (obtained by rejection sampling flip step proposals) defines a simple and commonly used Markov chain on the space of districting plans [39, 20].

Recall that for a function $f : V \to \mathbb{R}$ and a subset $W \subseteq V$ we write $f(W)$ for the sum $\sum_{v\in W} f(v)$, and that p is a weight function thought of as population on the nodes. We use the following notation below. Let $a, b : V \to \mathbb{R}$ be functions satisfying $0 \leq a(v) \leq b(v) \leq p(v)$ for all $v \in V$, and for simplicity assume $b > 0$. Let f be given by $f = a/b$, and we will abuse notation by using the same notation for a function on V or on the district nodes V_P. This induces district-level persistence diagrams $\mathcal{D}_P$ and $\mathcal{D}_{P'}$. Fix $\alpha = \alpha_{P,P'} > 0$ to be a lower bound on $b(W)/p(W)$ for $W = V_i, V_i', V_j, V_j'$.

We will apply this to $a = r$, the Republican vote totals per node, letting $b = r + d$ be the total major-party votes per node, so that $f = R$, the Republican voting share, and α is a lower bound on the voting turnout in districts affected by the perturbation. The more general notation is introduced here because there are several other functions a, b that could induce interesting filtration functions on the dual graphs, for instance to take third-party voting or other demographics into account.

We now state the stability result. Its proof is a straightforward application of Theorem 6.1, but we are able to formulate the estimate in terms of quantities which are directly interpretable in districting terms.

Theorem 6.4 (Stability under geographic perturbations)**.** *For a graph-preserving one-way perturbation P, P' where vertices V_{ij} are flipped from P_i to P_j,*

$$d_\infty(\mathcal{D}_P, \mathcal{D}_{P'}) \leq \frac{2\epsilon}{\alpha(1-\epsilon)} \max\left(|f(V_{ij}) - f(P_i)|, |f(V_{ij}) - f(P_j)|\right).$$

This gives us a bound on bottleneck distance for a perturbation in terms of how much the party vote share in the flipped units compares to the districts that they flip between. The proportionality constant depends on the population tolerance ϵ and the voter turnout constant α, and is small under the realistic assumption that $\epsilon \ll \alpha$. (For instance, with population deviation tolerance of 2% and an election with district-level turnouts of at least 25%, we obtain a coefficient of roughly .163.)

Proof. Relabel if necessary so that the affected districts of P are P_1, P_2, and P' replaces those by P_1', P_2', with districts $3, \ldots, k$ unchanged. Since the district dual graphs agree by assumption, the filtration function by f is defined on that graph. Theorem 6.1 says that

$$d_\infty(\mathcal{D}_P, \mathcal{D}_{P'}) \leq \|f_P - f_{P'}\|_\infty = \max\left(|f(P_1) - f(P_1')|,\ |f(P_2) - f(P_2')|\right).$$

Recall that $f(P_1) = a(P_1)/b(P_1)$, which in our case is Republican vote share, and likewise for the other districts. To simplify the exposition, we write $a_i = a(P_i)$

[6] Several trials show that a ReCom step (for Pennsylvania, $k = 18$) preserves the district dual graph with probability roughly .65. See Supplement for more discussion.

for $i = 1, 2$ and similarly define a_i', b_i, b_i'; we also write $\bar{a} = a(V_{12})$ and similarly for $\bar{b}$. Note that $a_1' = a_1 - \bar{a}$ and $b_1 = b' - \bar{b}$. It follows that

$$f(P_1) - f(P_1') = \frac{a_1}{b_1} - \frac{a_1'}{b_1'} = \frac{a_1}{b_1} - \frac{a_1 - \bar{a}}{b_1 - \bar{b}} = \frac{b_1 \cdot \bar{a} - a_1 \cdot \bar{b}}{b_1(b_1 - \bar{b})} = \frac{\bar{b}}{b_1'} \cdot (f(V_{12}) - f(P_1)) \,.$$

This works exactly similarly for P_2.[7]

Since P and P' are districting plans, the population balance condition implies that $p(P_1)$ and $p(P_1')$ both lie in the interval $[(1-\epsilon)T, (1+\epsilon)T]$. Since the vertices of V_{12} were moved without violating the balance condition, we have $p(V_{12}) \leq 2\epsilon T$.

Since $b(W) \leq p(W)$ for any vertex set W, we have

$$\alpha(1-\epsilon)\bar{b} \leq 2\epsilon\alpha(1-\epsilon)T \leq 2\epsilon\alpha \cdot p(V_1) \leq 2\epsilon \cdot b(V_1') = 2\epsilon b_1'.$$

To complete the proof, we rearrange to obtain $\frac{\bar{b}}{b_1'} \leq \frac{2\epsilon}{\alpha(1-\epsilon)}$. □

6.2. **Experimental stability under geographic perturbations.** While Theorem 6.4 gives a theoretical stability guarantee for geographic perturbations, its applicability is somewhat limited in scope, since the graph-preserving property is not automatic. In this section we present experimental evidence of stability.

We begin with Congressional and Senate ensembles for Pennsylvania. ReCom steps make large changes to a map, and the ensembles here were sub-sampled so that each successive plan collected in a ReCom ensemble is essentially independent of the last. So to illustrate stability we first turn to the much smaller perturbations made by flipping the assignment of a single unit at a time. Flip steps make small changes to the appearance of a plan—though most every pairwise district boundary has been changed after 1000 flip steps, the configuration of districts is clearly recognizable. In the examples shown in Figure 17, 1000 flip steps moves the persistence diagram by $d_\infty = .009$ and $d_\infty = .006$ for Congress and Senate, respectively. To contextualize distance magnitudes, note that maps in our ReCom ensembles (approximately independent maps) often have pairwise bottleneck distances between .05 and .1 (see Supplemental Figure 30).

We repeat the 1000-flip experiment on a larger set of initial maps to conclude the investigation of stability. We take one hundred Congressional and one hundred state Senate maps from the ReCom ensembles and make a sequence of 1000 flip steps starting from each of them. After each flip step, we compute the bottleneck distance between the initial plan and the current one. The results are shown in Figure 18. Looking at these bottleneck distance trace plots, we see that for the majority of maps, randomly flipping 1000 geographic units had a modest effect on the persistence diagram, typically in the order of 0.01 in bottleneck distance. This raises our confidence that the methods outlined here are reasonably robust to small geographic perturbation.

7. **Conclusion.** We have examined three approaches to the use of persistent homology in analyzing the complicated geography of voting at different scales of aggregation: zoning, election comparison, and signatures of gerrymandering. In all three cases, our methods use persistence diagrams to summarize a large ensemble

[7] Note that the proof carries through almost unchanged for a two-way perturbation, where $\bar{a}$ and $\bar{b}$ now represent net change in a and b. It produces a bound that is not very constraining for recombination moves, where population change $\bar{b}$ is very small but vote shifts $\bar{a}$ can be substantial. Nonetheless, we examine typical bottleneck displacement in recombination runs in Supplemental Figure 31.

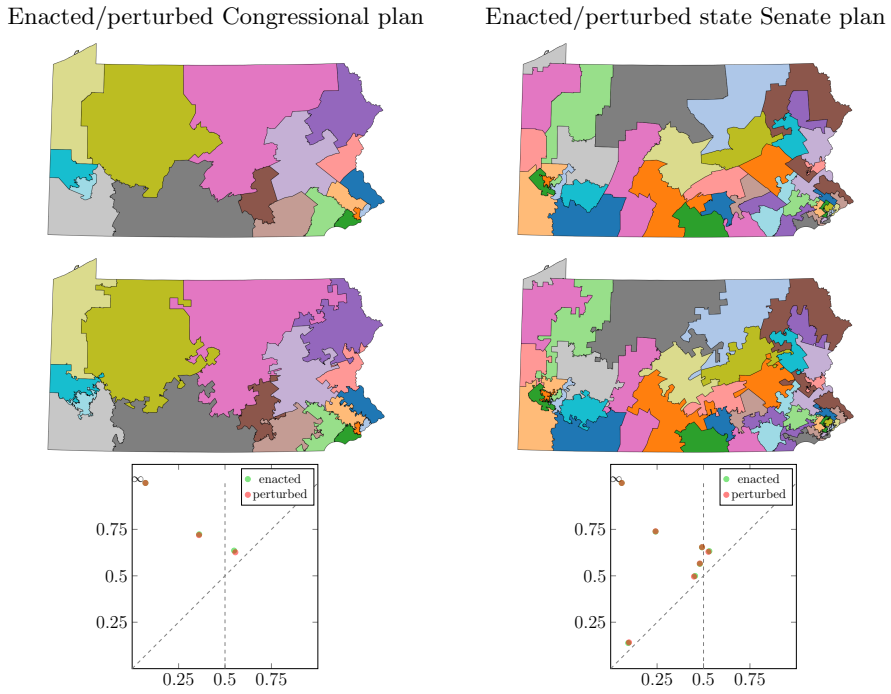


FIGURE 17. Maps of enacted plans before and after perturbation by 1000 flip steps; district adjacencies changed in both cases (e.g., blue/lavender for Congress, and purple/purple for state Senate districts in the Southeast). Nonetheless we see only a small change in persistence, which is typical for repetitions of this experiment.

of alternative districting plans. While the space of *all* valid redistricting plans for a state is intractably vast by almost any definition, tools like these allow us to cut through that complexity to uncover key information about the relationship of district lines to representation.

Vote-filtered dual graphs have been shown to lend themselves to interpretable (and nontrivial) observations about redistricting, opening up several directions for future research.

Classifying salient geoclusters. A long-standing debate in the political science literature centers on the question of whether racial gerrymandering protections like the Voting Rights Act actually produce counterintuitive partisan effects by disallowing the splitting of communities of color. The ideas introduced here can reframe the question in a far more sophisticated and general way, locating areas in a state that are split differently in party-biased ensembles or whose splitting correlates with party advantage.

Metrics on the space of plans. Converting districting plans to persistence diagrams allows us to compute a bottleneck distance between districting plans given some fixed vote data. It will be interesting to compare the bottleneck distance to other notions of distance between districting plans, such as the optimal transport-based distance studied in [1] and the information-theoretic distance in [38].

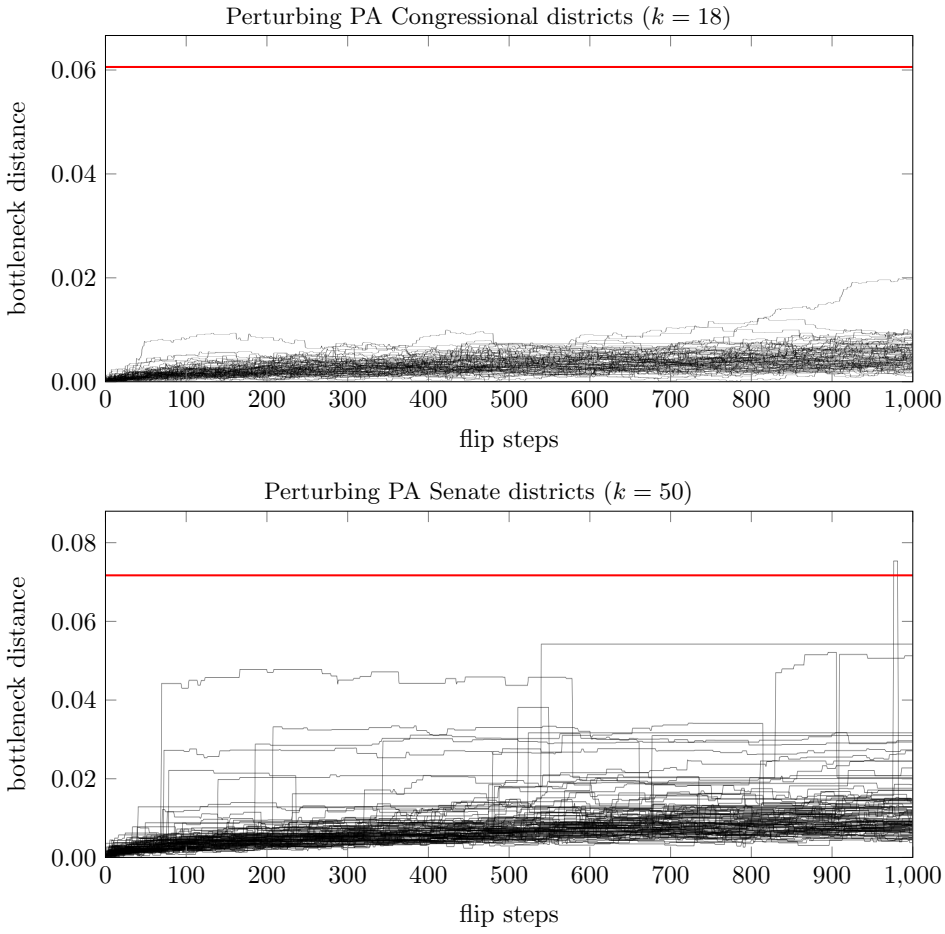


FIGURE 18. Effect of iterating small geographic perturbations (single-unit flip steps) on persistence. The red line indicates the average bottleneck distance over the $\binom{100}{2}$ plans in a ReCom ensemble $\mathcal{E} = \{P^{(1)}, \ldots, P^{(100)}\}$, shown as a baseline for the distance between approximately independent plans. For each $i = 1, \ldots, 100$, we run 1000 flip steps from $P^{(i)}$ and track the bottleneck distance to $P^{(i)}$. For $k = 18$, the persistence diagrams are seen to be stable. The same is true for *most* 50-district plans, but not all.

Embeddings. There are several methods in the literature for vectorizing persistence diagrams, such as *persistence landscapes* [8] and *persistence images* [2]—that is, each such method provides an embedding of a set of persistence diagrams into a Hilbert space as a set of vectors. We can use these methods to embed (ensembles of) plans in Euclidean space either for visualization or for clustering.

Distributions. Finally, the use of summary diagrams enables an approach to the comparison of different sampling techniques, such as by the use of different Markov chains. We can fit distributions to ensembles of persistence diagrams, allowing a summarized comparison of sampling distributions, which is a long-standing goal

in this research area. There is a large literature on distributions on the space of diagrams, obtained either by first vectorizing them and fitting classical distributions, or by applying more recent techniques directly to persistence diagrams themselves—see, e.g., [8, 17, 18, 19, 40, 51, 46].

Acknowledgments. The authors thank the TDA breakout group at the Voting Rights Data Institute—especially Colby Brown, Michelle Feng, Michelle Jones, Brandon Kolstoe, Brian Morris, Jasmine Noory, and Cameron Thieme—for inspiring conversations and explorations. We also thank the staff and students of the MGGG Redistricting Lab, especially Ruth Buck, for the data curation that makes projects like this possible. Finally, we are grateful to the anonymous referee(s), whose suggestions greatly improved the paper.

Appendix A. **Supplemental figures.**

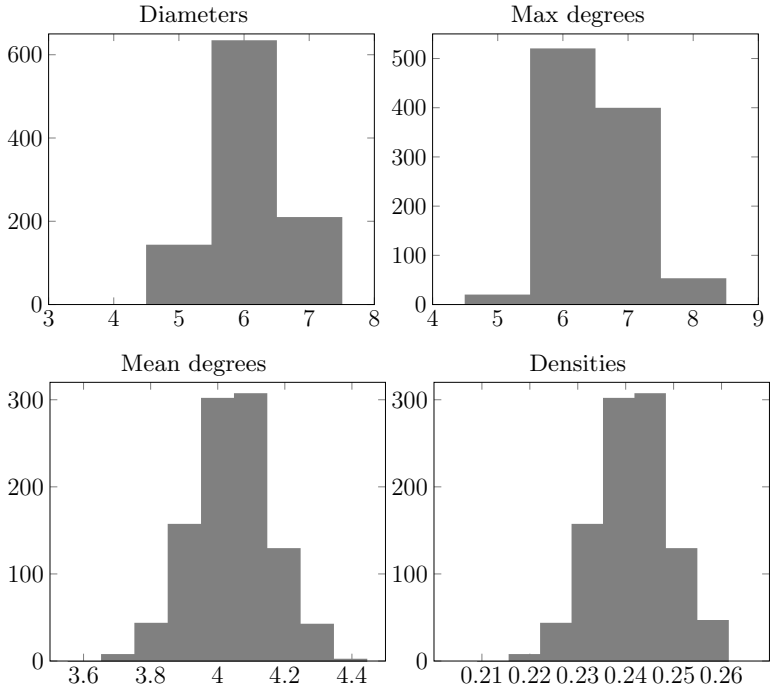


FIGURE 19. Histograms of statistics of the dual graphs for the ensemble of 1000 18-district plans for Pennsylvania. (*Density* is the ratio of the number of edges to the number of possible edges, $\binom{18}{2} = 153$.)

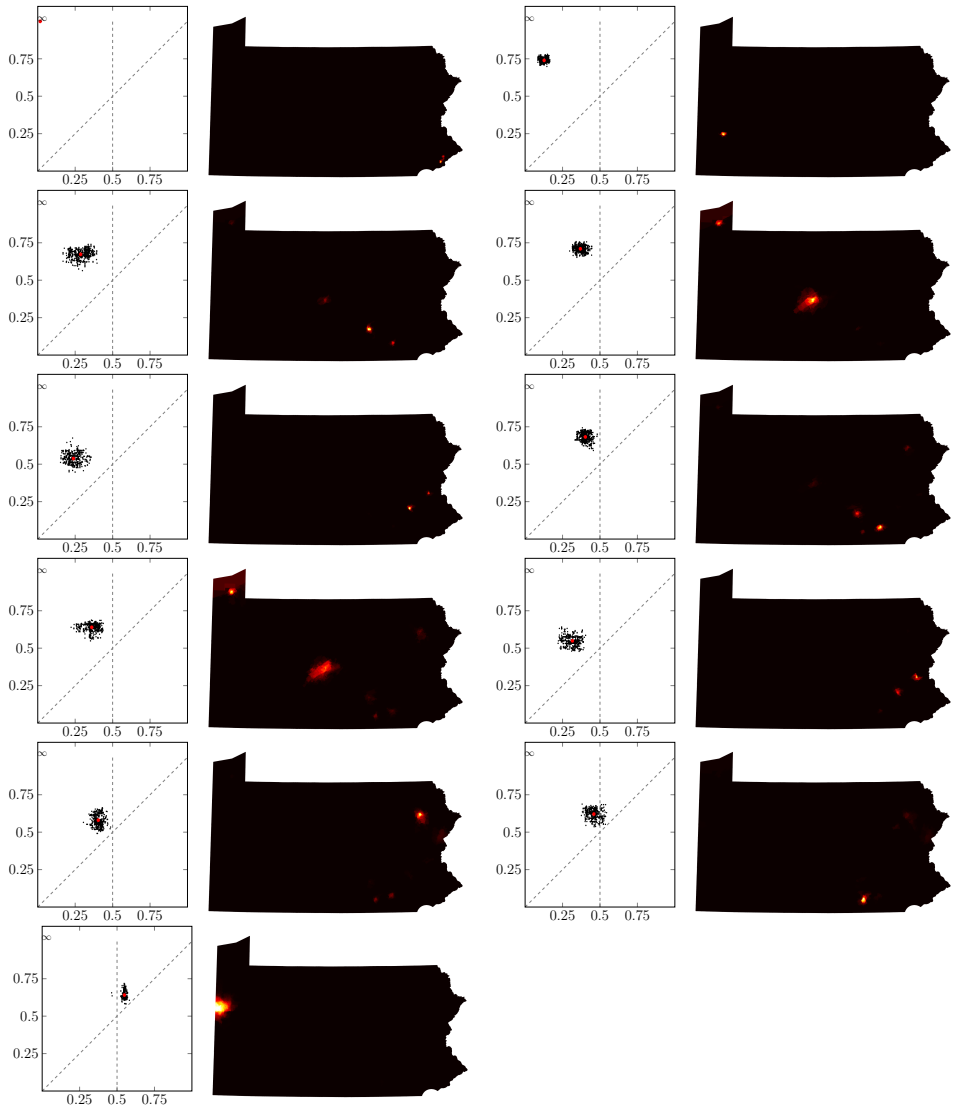


FIGURE 20. Geographical localization of eleven Fréchet features in Pennsylvania House plans ($k = 203$) with respect to PRES16 voting. These now pinpoint small cities, all the way down to Hermitage (population 16,220).

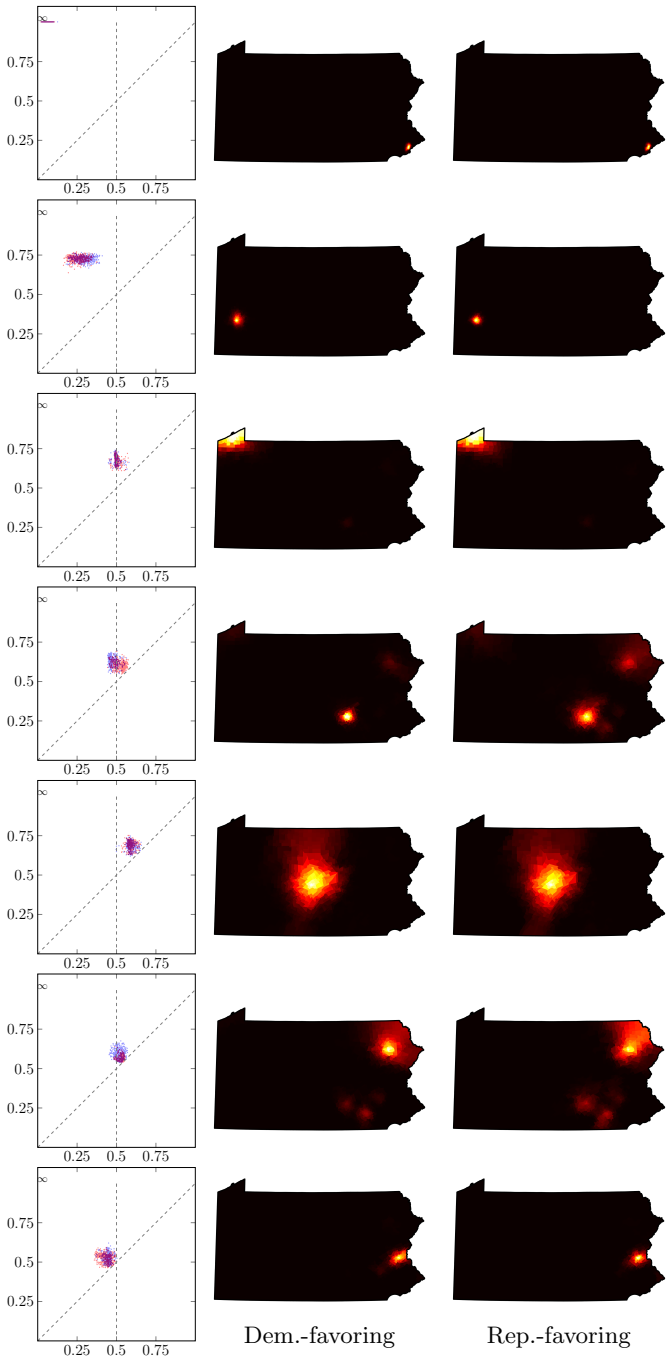


FIGURE 21. Expanding on Figure 10. Point plots and heat maps for the successive Fréchet features in the PA Senate ensembles that are biased for Democratic and Republican safe seats, respectively. Note the contrast between the rigidity of the Erie feature (similar placement for both ensembles) and the manipulability of the Harrisburg feature, though both anchor competitive districts.

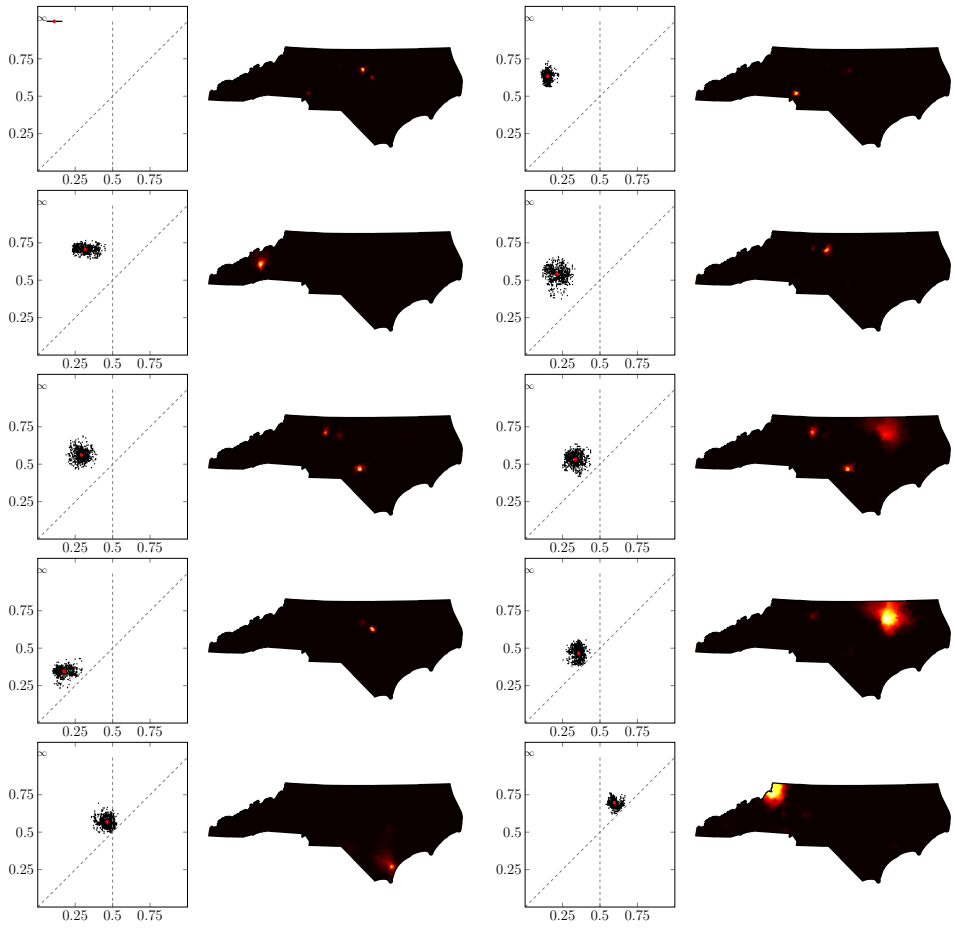


FIGURE 22. Geographical localization of ten Fréchet features in North Carolina Senate plans ($k = 50$) with respect to PRES16 voting.

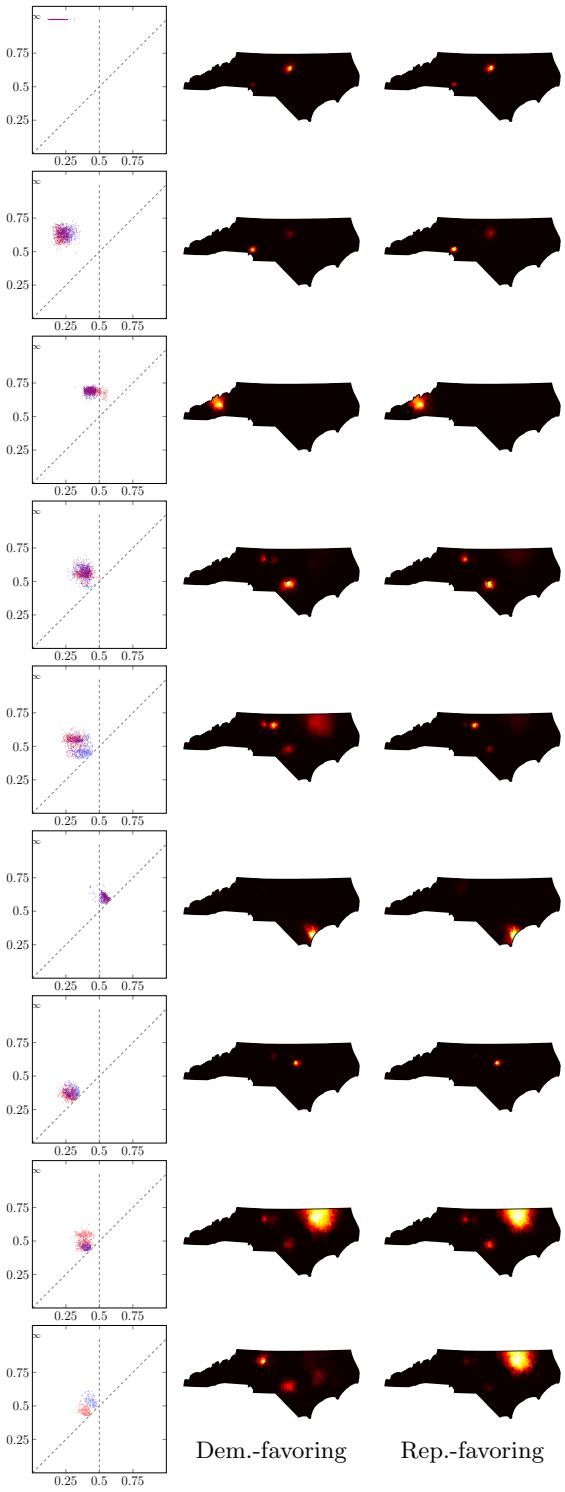


FIGURE 23. Comparison of point plots and heat maps for the successive Fréchet features in the NC Senate ensembles that are biased for Democratic and Republican safe seats, respectively.

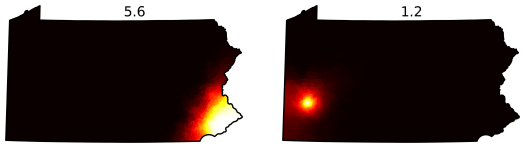


FIGURE 24. Location and average number of districts in connected components of Democratic-won districts for Pennsylvania Congressional plans ($k = 18$) with respect to PRES16 voting.

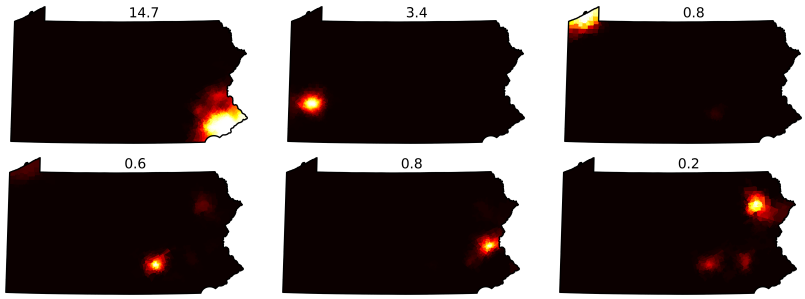


FIGURE 25. Location and average number of districts in connected components of Democratic-won districts for Pennsylvania state Senate plans ($k = 50$) with respect to PRES16 voting.

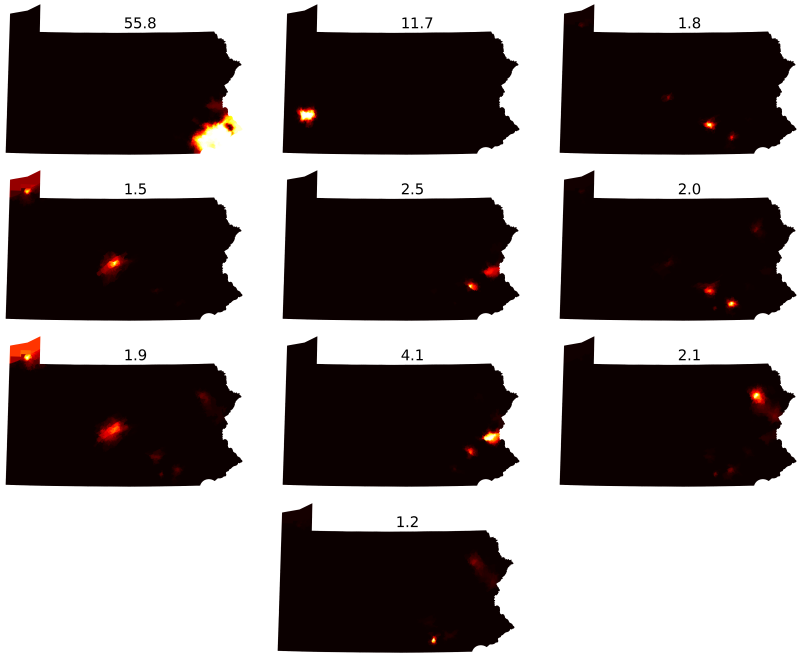


FIGURE 26. Location and average number of districts in connected components of Democratic-won districts for Pennsylvania state House plans ($k = 203$) with respect to PRES16 voting.

FIGURE 27. Location and average number of districts in connected components of Democratic-won districts for North Carolina Congressional plans ($k = 13$) with respect to PRES16 voting.

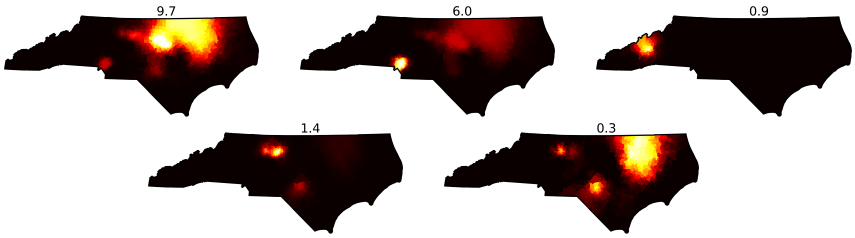


FIGURE 28. Location and average number of districts in connected components of Democratic-won districts for North Carolina state Senate plans ($k = 50$) with respect to PRES16 voting.

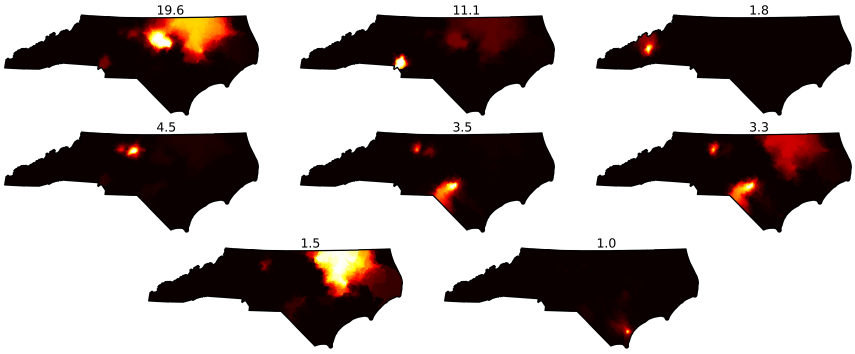


FIGURE 29. Location and average number of districts in connected components of Democratic-won districts for North Carolina state House plans ($k = 120$) with respect to PRES16 voting.

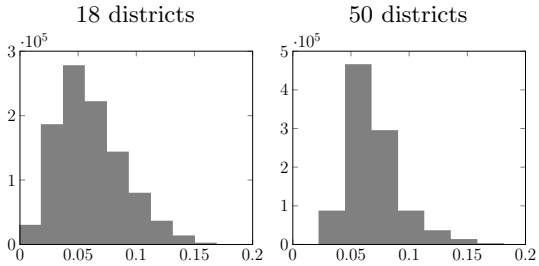


FIGURE 30. Distribution of pairwise bottleneck distances in the Pennsylvania ReCom ensembles. The distances help contextualize the size of the displacement by Flip steps.

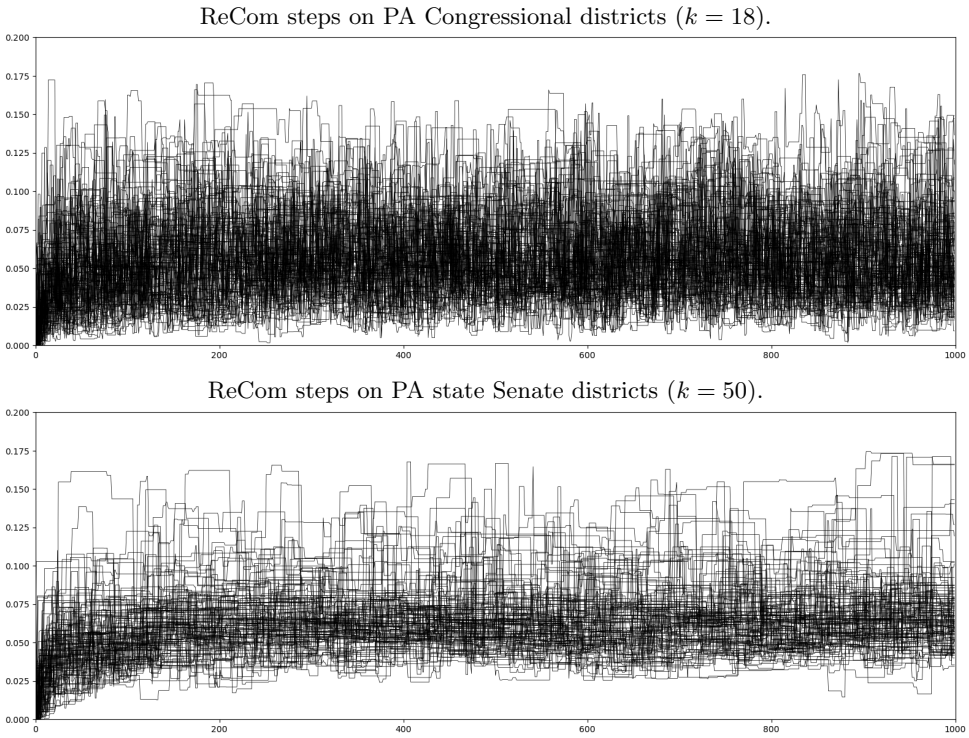


FIGURE 31. Effect of iterating ReCom steps (merging and resplitting pairs of adjacent districts) on bottleneck distance. We choose 100 starting plans $P^{(i)}$ from our Pennsylvania ensembles. For each of $i = 1, \ldots, 100$, we run 1000 ReCom steps (with no subsampling) from $P^{(i)}$ and track the bottleneck distance to $P^{(i)}$ on the y-axis.

REFERENCES

[1] T. Abrishami, N. Guillen, P. Rule, Z. Schutzman, J. Solomon, T. Weighill and S. Wu, Geometry of graph partitions via optimal transport, *SIAM J. Sci. Comput.*, **42** (2020), A3340–A3366.

[2] H. Adams, T. Emerson, M. Kirby, R. Neville and C. Peterson, et al., Persistence images: A stable vector representation of persistent homology, *J. Mach. Learn. Res.*, **18** (2017), 35pp.

[3] P. K. Agarwal, K. Fox, A. Nath, A. Sidiropoulos and Y. Wang, Computing the Gromov-Hausdorff distance for metric trees, *ACM Trans. Algorithms*, **14** (2018), 20pp.

[4] P. Bajardi, M. Delfino, A. Panisson, G. Petri and M. Tizzoni, Unveiling patterns of international communities in a global city using mobile phone data, *EPJ Data Science*, **4** (2015).

[5] A. Banman and L. Ziegelmeier, Mind the gap: A study in global development through persistent homology, in *Research in Computational Topology*, Assoc. Women Math. Ser., 13, Springer, Cham, 2018, 125–144.

[6] P. Bendich, J. S. Marron, E. Miller, A. Pieloch and S. Skwerer, Persistent homology analysis of brain artery trees, *Ann. Appl. Stat.*, **10** (2016), 198–218.

[7] R. Brüel-Gabrielsson, B. J. Nelson, A. Dwaraknath, P. Skraba, L. J. Guibas and G. Carlsson, A topology layer for machine learning, preprint, arXiv:1905.12200.

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

[9] P. Bubenik, J. Scott and D. Stanley, An algebraic Wasserstein distance for generalized persistence modules, preprint, arXiv:1809.09654.

[10] S. Cannon, M. Duchin, D. Randall and P. Rule, A reversible recombination chain for graph partitions, work in progress.
[11] G. Carlsson, Topological pattern recognition for point cloud data, *Acta Numer.*, **23** (2014), 289–368.
[12] G. Carlsson, V. de Silva and D. Morozov, Zigzag persistent homology and real-valued functions, in *Proceedings of the Twenty-Fifth Annual Symposium on Computational Geometry*, ACM, 2009, 247–256.
[13] G. Carlsson and F. Mémoli, Characterization, stability and convergence of hierarchical clustering methods, *J. Mach. Learn. Res.*, **11** (2010), 1425–1470.
[14] M. Carrière, F. Chazal, Y. Ike, T. Lacombe, M. Royer and Y. Umeda, PersLay: A neural network layer for persistence diagrams and new graph topological signatures, preprint, arXiv:1904.09378.
[15] F. Chazal, D. Cohen-Steiner, M. Glisse, L. J. Guibas and S. Y. Oudot, Proximity of persistence modules and their diagrams, in *Proceedings of the Twenty-Fifth Annual Symposium on Computational Geometry*, ACM, 2009, 237–246.
[16] F. Chazal, D. Cohen-Steiner, L. J. Guibas, F. Mémoli and S. Y. Oudot, Gromov-Hausdorff stable signatures for shapes using persistence, in *Computer Graphics Forum*, 28, Wiley Online Library, 2009, 1393–1403.
[17] F. Chazal, B. T. Fasy, F. Lecci, A. Rinaldo and L. Wasserman, Stochastic convergence of persistence landscapes and silhouettes, in *Computational Geometry (SoCG'14)*, ACM, New York, 2014, 474–483.
[18] F. Chazal, M. Glisse, C. Labruère and B. Michel, Convergence rates for persistence diagram estimation in topological data analysis, *J. Mach. Learn. Res.*, **16** (2015), 3603–3635.
[19] F. Chazal and B. Michel, An introduction to topological data analysis: Fundamental and practical aspects for data scientists, preprint, arXiv:1710.04019.
[20] M. Chikina, A. Frieze and W. Pegden, Assessing significance in a Markov chain without mixing, *Proc. Natl. Acad. Sci. USA*, **114** (2017), 2860–2864.
[21] S. Chowdhury, B. Dai and F. Mémoli, The importance of forgetting: Limiting memory improves recovery of topological characteristics from neural data, *PLoS One*, **13** (2018).
[22] D. Cohen-Steiner, H. Edelsbrunner and J. Harer, Stability of persistence diagrams, *Discrete Comput. Geom.*, **37** (2007), 103–120.
[23] L. Crawford, A. Monod, A. X. Chen, S. Mukherjee and R. Rabadán, Predicting clinical outcomes in glioblastoma: An application of topological and functional data analysis, *J. Amer. Statist. Assoc.*, **115** (2020), 1139–1150.
[24] J. Curry, The fiber of the persistence map for functions on the interval, *J. Appl. Comput. Topol.*, **2** (2018), 301–321.
[25] Y. Dabaghian, F. Mémoli, L. Frank and G. Carlsson, A topological paradigm for hippocampal spatial map formation using persistent homology, *PLoS Computational Biology*, **8** (2012).
[26] D. DeFord and M. Duchin, Redistricting reform in Virginia: Districting criteria in context, MGGG, 2019. Available from: https://mggg.org/va-criteria.pdf.
[27] D. DeFord, M. Duchin and J. Solomon, Comparison of districting plans for the Virginia house of delegates, MGGG, 2018. Available from: https://mggg.org/VA-report.pdf.
[28] D. DeFord, M. Duchin and J. Solomon, Recombination: A family of Markov chains for redistricting. preprint, arXiv:1911.05725.
[29] M. Duchin, Gerrymandering metrics: How to measure? What's the baseline? *Bull. Amer. Acad. Arts Sci.*, **71** (2018), 54–58.
[30] M. Duchin, Outlier analysis for Pennsylvania congressional redistricting, report, 2018. Available from: https://mggg.org/uploads/md-report.pdf.
[31] M. Duchin and O. Walch, *Political Geometry*, in press, Birkhäuser.
[32] H. Edelsbrunner and J. L. Harer, *Computational Topology. An Introduction*, American Mathematical Society, Providence, RI, 2010.
[33] K. Emmett, B. Schweinhart and P. Rabadan, Multiscale topology of chromatin folding, in *Proceedings of the 9th EAI International Conference on Bio-Inspired Information and Communications Technologies*, ACM, 2016, 177–180.
[34] M. Feng and M. A. Porter, Persistent homology of geospatial data: A case study with voting, *SIAM Rev.*, **63** (2021), 67—99.
[35] M. Feng and M. A. Porter, Spatial applications of topological data analysis: Cities, snowflakes, random structures, and spiders spinning under the influence, *Phys. Rev. Research*, **2** (2020).
[36] P. Gabriel, Unzerlegbare Darstellungen. I, *Manuscripta Math.*, **6** (1972), 71–103.

[37] B. Grofman and G. King, The future of partisan symmetry as a judicial test for partisan gerrymandering after LULAC v. Perry, *Election Law J.*, **6** (2007), 2–35.
[38] L. Guth, A. Nieh and T. Weighill, Three applications of entropy to gerrymandering, preprint, arXiv:2010.14972.
[39] G. Herschlag, H. S. Kang, J. Luo, C. Vaughn Graves, S. Bangia, R. Ravier and J. C. Mattingly, Quantifying gerrymandering in North Carolina, *Stat. Public Policy*, **7** (2020), 30–38.
[40] Y. Hiraoka, T. Shirai and K. D. Trinh, Limit theorems for persistence diagrams, *Ann. Appl. Probab.*, **28** (2018), 2740–2780.
[41] C. D. Hofer, R. Kwitt and M. Niethammer, Learning representations of persistence barcodes, *J. Mach. Learn. Res.*, **20** (2019), 45pp.
[42] M. Lesnick, The theory of the interleaving distance on multidimensional persistence modules, *Found. Comput. Math.*, **15** (2015), 613–650.
[43] J. Levitt, A citizen's guide to redistricting, *SSRN*, 2008, 139pp.
[44] A. Lytchak, Open map theorem for metric spaces, *St. Petersburg Math. J.*, **17** (2006), 477–491.
[45] C. Maria, J.-D. Boissonnat, M. Glisse and M. Yvinec, The Gudhi library: Simplicial complexes and persistent homology, in *International Congress on Mathematical Software*, Lecture Notes in Computer Science, 8592, Springer, Berlin, Heidelberg, 2014, 167–174.
[46] V. Maroulas, F. Nasrin and C. Oballe, A Bayesian framework for persistent homology, *SIAM J. Math. Data Sci.*, **2** (2020), 48–74.
[47] J. Mattingly, *Report on Redistricting: Drawing the Line.* Common Cause v. Rucho, 318 F.Supp.3d 777, Ex.3002, **871** (2018).
[48] M. D. McDonald and R. E. Best, Unfair partisan gerrymanders in politics and law: A diagnostic applied to six cases, *Election Law J.*, **14** (2015), 312–330.
[49] MGGG, MAUP: The geospatial toolkit for redistricting data. Available from: https://github.com/mggg/maup.
[50] MGGG, MGGG-states: Open collection of precincts shapefiles for U.S. states. Available from: https://github.com/mggg-states.
[51] Y. Mileyko, S. Mukherjee and J. Harer, Probability measures on the space of persistence diagrams, *Inverse Problems*, **27** (2011), 22pp.
[52] D. Morozov, K. Beketayev and G. Weber, interleaving distance between merge trees. *Discrete Comput. Geom.*, **49** (2013), 22–45.
[53] 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, *PNAS*, **108** (2011), 7265–7270.
[54] W. Pegden, *Pennsylvania's Congressional Districting is an Outlier: Expert Report*, League of Women Voters, et. al v. the Commonwealth of Pennsylvania, et. al. 159 MM, 2017.
[55] J. Rodden and T. Weighill, Political geography: A case study of districting in Pennsylvania. preprint, arXiv:2010.14608.
[56] N. O. Stephanopoulos and E. M. McGhee, Partisan gerrymandering and the efficiency gap. *The University of Chicago Law Review*, (2015), 831–900.
[57] B. J. Stolz, H. A. Harrington and M. A. Porter, The topological "shape" of Brexit, *SSRN*, (2016), 9pp.
[58] K. Turner, Y. Mileyko, S. Mukherjee and J. Harer, Fréchet means for distributions of persistence diagrams, *Discrete Comput. Geom.*, **52** (2014), 44–70.
[59] Voting Rights Data Institute, GerryChain, GitHub Repository, 2018. Available from: https://github.com/mggg/GerryChain.
[60] A. Zomorodian and G. Carlsson, Computing persistent homology, *Discrete Comput. Geom.*, **33** (2005), 249–274.

Received August 2020; revised February 2021.

E-mail address: moon.duchin@tufts.edu
E-mail address: tneedham@fsu.edu
E-mail address: t_weighill@uncg.edu