A PERSISTENT HOMOLOGY-BASED GERRYMANDERING METRIC

KENZ KALLAL

Abstract. We present an extension of Feng and Porter's 2019 paper on their level-set method for the construction of a filtered simplicial complex from geospatial election data. The main additional feature of our method is that it takes into account the local density of one party's advantage (that is, the size of that party's margin in the smallest region in which the data is known divided by the area of that region). Such a construction is useful to us because it is the input to persistent homology, the output of which — a persistence barcode — is the summary we hope explains some useful topological features of the election data. In particular, we produce a concrete application of our method to measuring the representation that areas containing heavy majorities for one party recieve in the election results for the congressional districts containing them (the main idea being that persistent H_1 homology classes in the barcode corresponds to areas heavily favoring one party). The purpose of this is to detect and measure the effects of gerrymandering on the representation that connected regions containing large margins for one party (e.g. cities) end up recieving in the district-wide results; such representation, which might normally appear in a fair districting plan, may not appear when the *cracking* technique of gerrymandering is applied. To demonstrate what kind of information the barcodes produced by our algorithm contain, we apply it to recent election and districting data from eleven states obtained from the collection of GIS shapefiles pooled by the Metric Geometry and Gerrymandering Group, and analyze its output in some interesting cases.

Contents

Date: January 22, 2019.

1. Introduction

Suppose that there are two political parties, A and B who gain representation in congress in proportion to the number of districts in which they obtain a majority of the votes. The party in charge of drawing the districts, say party B, can manipulate the districts to gain advantage in what is called *partisan gerrymandering*. The work in this paper is focused on the following question:

Question 1.1. *Given a district plan and precinct-level vote totals for parties A and B, is there a way to measure the extent to which the district plan is gerrymandered?*

Of course, Question 1.1 is really asking for a normative standard for what constitutes "fairness" of a district plan, so there has naturarally existed many distinct answers to it. Some of the previous work on Question 1.1 has focused on the geometric notion of *compactness* of a district, in which the normative standard is that districts with boundaries which look unreasonable due to being twisted and elongated could be the result of gerrymandering. In particular, the boundaries of gerrymandered districts are manipulated in order to do *cracking* (diluting the votes from an area with a majority of A voters by manipulating district boundaries in areas with a majority of B voters so that they include parts of that area) and *packing* (manipulating district boundaries so that that one district contains several non-contiguous blocks of voters for party A that might normally cause more than one district to have majorities for A). The most common example of blocks of A voters whose representation is affected by gerrymandering is the case of a city with high margins for party A. A gerrymandered district might be drawn to pack two cities together and waste votes for party A, or to crack a city into many districts, each containing large rural or suburban areas which force many of those districts to end up being won by party B. A canonical example of this from Wikipedia[1] is the 2004 redistricting of Texas (see Figure 1). In the 2004 district plan, all three districts with nontrivial intersection with Travis county (where the heavily-democratic city of Austin is located) show evidence of gerrymandering because of their noncompactness: one district is elongated in the eastern direction to dilute democratic votes with republican ones in less urban areas; the other two are elongated towards the south and southwest to contain large swathes of rural areas, as well as portions of other cities, namely San Antonio and McAllen. This could be

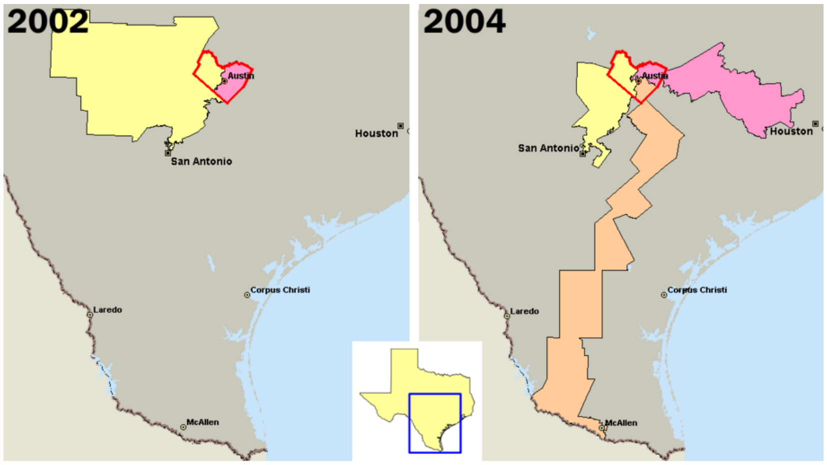


Figure 1. The 2004 Texas redistricting. Note the cracking of Austin into three districts.

[1] https://en.wikipedia.org/wiki/Gerrymandering, accessed January 22, 2019.

an example of either packing or cracking, depending on whether the intention was to win or lose those districts. In any case, the districts visibly fail the compactness test. However, compactness alone as a measure of partisan gerrymandering is not sufficient for many reasons, not least because of disagreements about what the formal definition of district compactness should be. To summarize the more thorough description of Duchin in [7], the main problem with using purely geometric criterion on the shapes of districts is that it doesn't take into account the actual distribution of voters, and is too sensitive to the geography of the region (in concrete terms, districts might have non-compact-looking boundaries because of mountain ranges or bodies of water, or because they are drawn to give representation to a specific contiguous group of people who happen to be distributed geographically in a certain way).

Some other methods ignore the geometry of the districts completely, and just focus on vote totals and representation. For example, the *efficiency gap*, originally due to Stephanopoulos–McGhee [11], is based on the number of votes wasted by both parties (the number of wasted votes for party A in a district is the number of votes over 50 percent received by A if the district is won by A, or the total number of votes recieved by A if the district is lost). It is defined simply as

$$EG = \frac{W_A - W_B}{T},$$

where W_P is the total number of wasted votes by party P, and T is the total turnout. The efficiency gap also has flaws, such as the fact that it penalizes proportionality and rewards non-competitive districts (see [3]). Also, like the compactness criterion, it doesn't take into account the distribution of voters amongst the districts. The canonical example of this, which we take from [6], is of a state where every household has 2 voters for B and 1 for A. The efficiency gap of any districting plan in such a state will be 1/6, much higher than the limit of 0.08 proposed by [11].

An influential and currently-developing program of Duchin improves on this issue via the method of sampling (see [6, §4] for a brief overview). The basic normative principle is that a districting plan with an extreme outcome (that is, one in which the number of districts won by party A is uncommon with respect to a large sampling of random district plans). This method applies much more generally than any of the previous ones mentioned so far.

The main idea of our contribution to Question 1.1 is to again incorporate the shape of the districts, this time only focusing on their topology, and also taking into account the distribution of the voters within those districts. The main technical input is *persistent homology*, a tool from algebraic topology which is the central feature of topological data analysis (see e.g. [4]).

1.1. **Persistent homology.** Much of this section comes directly from Jun Hou Fung's exposition for his summer 2019 tutorial. Let $X \subset \mathbf{R}^n$ be a finite point cloud. The main task of topological data analysis is to cook up an appropriate *filtered simplicial complex* on X, which must be carefully chosen to suit the type of data.

Definition 1.2. A *filtered simplicial complex* on X is a functor $C_{-}(X) : \mathbf{R}_{\geq 0} \to \mathsf{SimpCplx}$, where $\mathbf{R}_{\geq 0}$ denotes the poset category on the nonnegative real numbers with the usual total ordering, and for all $\epsilon \in \mathbf{R}_{\geq 0}$, $C_\epsilon(X)$ is a simplicial complex on a set of vertices coming from X (In particular the number of vertices in $C_\epsilon(X)$ is at most $|X|$). In this context, we restrict maps between simplicial complexes to inclusions, so that for $\epsilon_1 \leq \epsilon_2$, the induced map $C_{\epsilon_1}(X) \to C_{\epsilon_2}(X)$ is an inclusion of simplicial complexes.

The purpose of coming up with a filtered simplicial complex for a point cloud is usually as follows:

Slogan 1.3. At a given $\epsilon \geq 0$, the simplicial complex $C_\epsilon(X)$ is a topological model for X, viewed at a scale proportional to ϵ. In particular, the model gets "fuzzier" as ϵ increases. Instead of tuning ϵ as its own parameter to the appropriate level to detect topological features at any particular scale, persistent homology allows one to observe topological features at every possible scale, and to see which scales any particular feature is present at.

This brings us to the method of detecting topological features, which for our purposes will just just be the simplicial homology[2] of a given simplical complex on X. Since simplicial homology is functorial, for each $i \geq 0$ and field $\mathbf{F}$, we have an i-th homology functor $H_i(C_-(X); \mathbf{F}) : \mathbf{R}_{\geq 0} \to \mathsf{Vect}_\mathbf{F}$, which is a *filtered* $\mathbf{F}$*-vector space.* This particular example of a filtered $\mathbf{F}$-vector space is the *i-th persistent homology group* of X. Since X is finite, there are only finitely many simplicial complexes on it, and since the filtered simplicial complex $C_-(X)$ is strictly increasing in size, it can only change at finitely many values, where the intervals in which it is constant are represented by $\epsilon_j \in \mathbf{R}_{\geq 0}$. Plus, the finiteness of X also guarantees that the homology groups of a simplicial complex on it are finite-dimensional $\mathbf{F}$-vector spaces. So this filtered $\mathbf{F}$-vector space is really a finitely-generated graded $\mathbf{F}[x]$-module, where the j-th graded part is $H_i(C_{\epsilon_j}(X))$ and the action of x is the induced map on H_i of the inclusion of simplicial complexes $C_{\epsilon_j}(X) \to C_{\epsilon_{j+1}}(X)$. As a consequence of this and the structure theorem for finitely-generated graded modules over a PID, we have

Theorem 1.4. *Let X be a finite point cloud. There exist finite lists of integers $\{r_j\}, \{s_j\}, \{t_j\}$ such that as a graded $\mathbf{F}[x]$-module,*

$$H_i(C_-(X); \mathbf{F}) \cong \bigoplus_j x^{t_j} \cdot \mathbf{F}[x] \oplus \bigoplus_j x^{r_j} \cdot \mathbf{F}[x]/(x^{s_j})$$

In more concrete terms, there is a choice of bases for the vector spaces $H_i(C_{\epsilon_j}(X))$ such that one basis always maps injectively to the next, not counting the basis elements that map to zero[3]. In particular, once a filtered simplicial complex is chosen, a *persistence barcode* consisting of bars representing the lifespans of these basis vectors can be drawn, and by the above theorem, the barcode is a complete invariant for the i-th persistent homology group of $C_-(X)$. In terms of Theorem 1.4, the H_i homology classes represented by the bars in the barcode are born at times r_j and die at times s_j; the ones that do not die are born at times t_j.

1.2. **Filtered simplicial complexes for geospatial data.** When it comes to analyzing the topological features of a specific type of data, the key construction is that of the filtered simplicial complex $C_-(X)$. The main ones found in the literature are completely distance-based: in the Čech complex, for example, a simplex $[v_0, \ldots, v_n]$ is included if and only if $\bigcap_{i=0}^n B_\epsilon(v_i) \neq \emptyset$.

This type of method doesn't adapt well to geospatial voting data, because of the variance in density of the data points. The features that really matter are the adjacencies between districts or precincts, since they are what determine the relative locations of the voters. This is why Feng and Porter [8] considered the *adjacency complex* construction, where each voting district is represented by a single data point, and two districts are connected if and only if they are adjacent, but only included if one party wins by a certain percentage decreasing in ϵ (then as in the Vietoris–Rips complex, we take the maximal simplicial complex containing this 1-skeleton). Feng–Porter [8, §3.4.1] explains, "more persistent features represent holes with voting results that are very different

[2] The definition of simplicial homology can be found in [10, §2.1], for example.

[3] NB there was no reason to invoke the structure theorem for PID's to arrive at this conclusion.

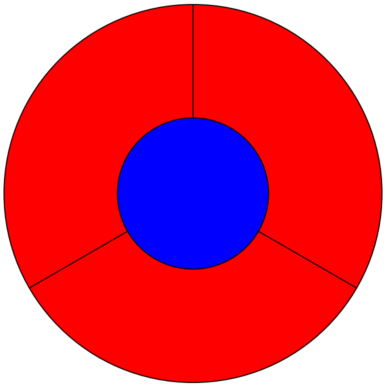


FIGURE 2. An example of a hole undetected by the adjacency complex

from their neighbors. Consequently, the most persistent features are exactly the most meaningful ones, as they indicate which regions have the strongest outlying signals."

This is a perfectly reasonable way of carrying out a topological analysis of voting data, except for the fact that it doesn't take into account the actual relative positions of the districts since it represents each district with just a single point. Consider the configuration in which three mutually adjacent districts surround a smaller one (see Figure 2). Then the adjacency misses the fact that if these three districts vote a different way then the one in the middle, there should be a hole present in the middle. Instead, there is a triangle in the 1-skeleton which automatically gets completed to a 2-simplex and fills in the hole that should be present. Feng–Porter's solution to this problem is to treat the districts (without loss of generality those in which a majority voted for party A) as a 2-dimensional submanifold with boundary of $\mathbf{R}^2$, from which a simplicial complex can be generated just by taking a triangulization of this. In practice, this is done by taking every pixel in this set to be a different point, and connecting them if they are adjacent as pixels. As ϵ increases, this submanifold evolves by expanding its boundary as the level-set of some function.

1.3. **Organization.** This paper is organized as follows. In Section 2, we explain Feng–Porter's levelset construction that forms the basis for our work on Question 1.1, as well as the reason why our improvement is necessary. We will see that this fits well into the context of previous criticisms of other attempts to measure gerrymandering such as compactness scores and the efficiency gap. In Section 3, we describe our modification to their algorithm and the idea for how it can be used to measure Gerrymandering. In Section 4, we demonstrate the application of our method to recent election data from eleven states. Finally, in Section 5 we suggest further ways in which our approach may be enhanced to provide more useful information.

2. THE LEVELSET CONSTRUCTION

2.1. **The construction.** In practice, voting data is given as a GIS SHAPEFILE, which can be rasterized onto an $N \times M$ grid of pixels. In particular, we can choose to color the regions with a majority voting for a certain party with a corresponding value. In theory, the map of the state is a compact submanifold with boundary of $\mathbf{R}^2$. We can view a region (i.e. the union of some collection of districts or precincts or whatnot) as also being a submanifold $M \subset \mathbf{R}^2$, with boundary Γ. We can take the simplicial homology of such a manifold, just by taking an arbitrary triangulation. In practice this is done as described in the previous section, by considering the simplicial complex which is maximal for the 1-skeleton given by all the pixels in M where they are connected if they are queen-adjacent in the $M \times N$ grid. The construction of a filtered simplicial complex, even in practice, just depends on something which in our language would be a "filtered submanifold" of $\mathbf{R}^2$, namely a functor $\mathbf{R}_{\geq 0} \to \mathsf{Man}$ (where our version of Man is much more restrictive than the usual smooth category: we are just talking about the compact submanifolds with boundary of $\mathbf{R}^2$ where the morphisms must be inclusions). As usual there is no reason to use this language to define a filtered manifold, because in practice it only changes at finitely many points (the process of expanding the boundary will have to be finite since there are finitely many pixels on the grid), and this filtered manifold is really just a sequence of regions

$$M \subset M_{t_1} \subset M_{t_2} \subset M_{t_3} \subset \cdots \subset M_{t_n}.$$

In the simple version of the levelset construction from [8, §3.3], the way to construct these is just to expand the boundary Γ with constant velocity in the normal direction. In practice this is done by viewing $\Gamma_t = \partial M_t$ as the levelset (i.e. zero set) of a function $f_t : \mathbf{R}^2 \to \mathbf{R}$ in a 1-parameter family. The functions f_t are constructed by setting f_0 to be given by

$$f_0(p) := d(p, \Gamma),$$

where the distance $d(p, \Gamma)$ is signed, taking negative values when $p \notin M$, so that $\Gamma = \Gamma_0$ is clearly the levelset of f_0. The rest of the 1-parameter family is defined to evolve via the differential equation (easily simulated on a computer)

$$\frac{\partial}{\partial t} f_t(x) = \alpha |\nabla|_x f_t|$$

for an appropriate constant α. This clearly causes the levelset to evolve at a constant velocity (controlled by α) in the normal direction.

After constructing this filtered manifold as an increasing sequence of subsets of the $N \times M$ grid, the above remarks explain how to get a filtered simplicial complex and therefore a persistence barcode. We repeat the description again now. The sequence of manifolds $\{M_t\}$, represented as subsets of an $N \times M$ grid of pixels, induce a filtered simplicial complex just by taking the 1-skeleton given by the pixels in M_t plus their adjacencies as pixels. Then you can take the persistent homology (see Section 1.1) of this filtered simplicial complex to get the desired barcode. The point now is that persistent H_1 homology classes correspond to "holes" in the areas voting for A. In practice, these come from cities with strong majorities for B.

2.2. **The main disadvantage.**

The main problem with tring to apply the levelset method of Feng–Porter to Question 1.1 is that the construction of the simplicial complex is not sensitive to variations inthe absolute voter margin per square meter (which is usually a proxy for population density). In fact, this is the same type of issue that causes the Cech, Vietoris–Rips, etc. complexes to be rather weak when applied to this kind of data (see [8, §3.4]). It's also behind what Feng–Porter [8, §3.4.1] refer to as the "scaling" problem. Specifically, a region with only a small margin one way or the other gets filled in by the expanding levelset at the same rate as a region of the same size (e.g. a city) with much larger contribution to the margin. This means that the persistence barcode coming from setting M to the region where party A received a majority of votes is not actually sensitive to the difference between a hole that comes from a small region where B received a slight majority, and a small region where B received a much large majority[4]. In fact, this is the main advantage of the adjacency complex construction – the persistence is in terms of a parameter that controls

[4] or alternatively, it treats large regions without many votes as being more persistent than small regions with many marginal votes. This is a problem because the most persistent features ought to be those that come from exactly the opposite: small, densely-populated areas with high margins for one party.

exactly this. It might be worthwhile to consider using multidimensional persistence to control this as well, but our solution is much simpler. We explain it in the next section.

3. The modified levelset construction

The problem we experience with Feng–Porter's levelset construction is the fact that it is insensitive to differences in voting margin. Using a similar idea as the original one behind *weighted persistent homology* (see [2]), we fix this by making the rate α at which the level set expands proportional to the local density of the margin for party A. This is measured by a function $m : \mathbf{R}^2 \to \mathbf{R}$ which at each point returns the margin by which party B loses. In practice, we can only measure the margin on certain small regions (since voting data can only go down to the precinct level). Luckily, it's easy using the standard GIS tools to determine the area of a region in a SHAPEFILE, so we can just take $m(x)$ to be the absolute margin of victory for party A in the precinct containing x divided by the area of that precinct. It measures the number of votes (per unit of area) that A wins by close to x. We can convince ourselves that this is the relevant quantity: the gerrymanderer only cares about the total number of votes they can get by manipulating the boundaries of the districts.

The appropriate modification to the evolution of the boundaries for the construction of the filtered simplicial complex is very simple. This time, we still expand the boundaries in the normal direction, but we do so at a velocity which is inversely proportional to m (so that holes with lower margins per unit of area are filled in faster than others; this is with the hope that cities with high margins for party B will show up as very persistent homology classes):

$$\frac{\partial f_t}{\partial t} = \frac{1}{m(x)}\alpha|\nabla f_t|.$$

Similarly to what is observed in [8], this method will cause M_t to be more and more connected, until it fills up the entire state with a hole for each city voting for the opposing party. This means that homology classes which are born near $t = 0$ are typically still the only relevant ones. Homology classes born after $t = 0$ represent areas with high densities of voters for party B surrounded by areas which party B also wins, but with a smaller majority.

3.1. Application to gerrymandering.

The key idea of our modified levelset construction was as follows:

Slogan 3.1. Suppose the modified levelset construction is used to produce a barcode for a set of voting data, starting with $M = M_0$ equal to the region which voted for party A. A very persistent homology class born near $t = 0$ represents a region where B won by a large margin and which is surrounded or mostly surrounded by regions where A won a majority.

To use this idea to detect gerrymandering, we can't just simply apply persistent homology. This will just give you an idea of which cities exist. Instead, we need to detect whether the votes coming from a city are represented when it comes to district-level votes. Under the assumption that precincts are small enough to not allow much gerrymandering (probably true enough in large cities where precincts are very small and close together), all the cities with the property described in Slogan 3.1 should show up in the barcode obtained from the precinct-level voting data.

On the other hand, we can measure how much these votes are represented in the district-level counts by repeating the same thing, except only considering the district-level voting data (i.e. only rasterizing points based on who won the district they are in, instead of the precinct) when constructing M. If the state is not gerrymandered, we expect the district-level barcode to be similar to the precinct-level barcode. In the language of Slogan 3.1, cities full of voters for party B should create districts which also form holes in the outside regions voting for A. In the gerrymandered situation, a hole representing a city in the precinct-level data dissapears or becomes smaller (i.e. less persistent) once we move to the district-level barcode. So this method should be very effective at measuring the *cracking* form of partisan gerrymandering, since it can measure the representation that areas containing many votes for party A receive in the district results. It isn't necessarily limited to cracking, though: for instance, if two cities with majorities for party N are packed into the same district, two H_1 homology classes in the precinct-level barcode both born at $t = 0$ become one born at $t = 0$ plus another one which is born at $t > 0$. The lateness of the birth of the second homology class reflects how much space is in between them[5]. On the other hand, the barcode doesn't necessarily indicate whether this is the result of A packing B districts together, or of B diluting its own votes in the cities in order to win many districts. To fix this issue, one would need to add additional information to the barcode that includes how many districts are included in each hole.

4. Experimental Results

In this section, we demonstrate the application of the ideas in Section 3 to recent election data in the following eleven states: Colorado, North Carolina, Pennslyvania, Texas, Maryland, Michigan, Minnesota, New Mexico, Ohio, Utah, and Virginia. Hopefully these examples will make those ideas clearer. We used the Phat python package [1] to compute persistent homology when needed, and applied the appropriate modifications to the code from Feng–Porter [8]. The necessary GIS shapefiles for the election data were obtained from the publically-available github repository provided by the Metric Geometry and Gerrymandering Group [9]. In fact, our choice of states is exclusively due to which states in that repository had usable state-wide data. In all of our examples, we let B be the Democratic party and A the Republican. The goal is to understand how cities with high blue margins are represented in district-level results.

Let us begin with a more detailed explanation for the Texas data. First off, in the precinct-level data, just by filtering based on which party won a majority in each precinct (see Figure 3), we can observe four holes coming from the four major cities in Texas, namely Houston, Dallas, Austin, and San Antonio. We see a similar trend in the calculation of the vote margins per unit of area (Figure 4). Next, in Figure 6 we show a few steps in the evolution of the modified levelset (we ran it for 25 time steps). It doesn't come into play in Texas, but it's important to note that many homology classes won't die until after $t = 25$, in which case there full lifespan is not completely represented by the barcode (so it's impossible to tell whether a homology class lives until 25 or much longer just from the barcode). Note, too, that in practice it's best to include the empty space around the state in the rasterized levelset. This is so that cities sitting on the edge of a state do not get ignored by the barcode. As expected, most of the features we aren't interested in (e.g. small blocks containing only a small blue majority) quickly dissapear, and the most persistent ones come from three major cities Austin, Houston and Dallas. Slightly less persistent are those from the San Antonio, El Paso, and McAllen areas. This clearly comes through in the precinct-level

[5] In practice, we cap the value of $m(x)$ at some upper bound, so that the constant of proportionality $1/m$ doesn't blow up. This means that oftentimes the rate of expansion of the levelset in rural areas between two cities voting for B is still constant as in the method of Porter–Feng. From this lucky coincidence, we are able to interpret the lateness of the birth of the second homology class as a measure of the space between them, and not as a measure of the number of voters. This is exactly what we want to measure when we deal with packing.

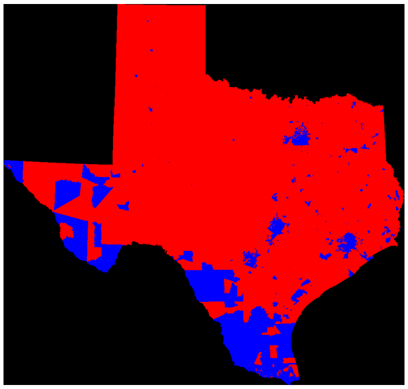
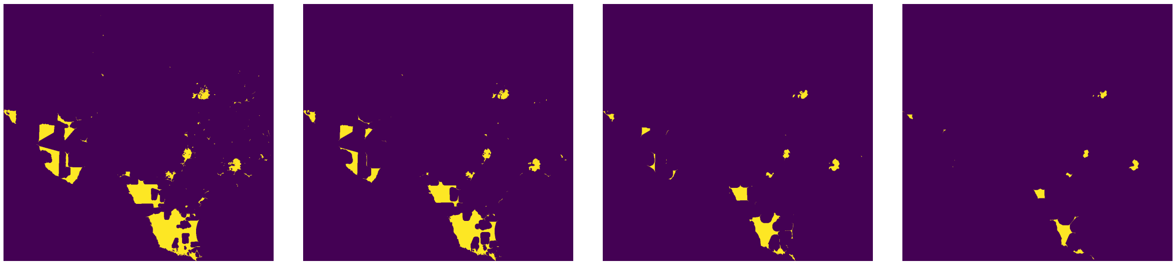


FIGURE 3. The 2016 presidental election, precinct-level results. Blue areas represent precincts won by Clinton, and red areas represent precincts won by Trump

FIGURE 4. Darker-colored precincts are those with more marginal votes per unit of area



barcode Figure 7 (together they are the six longest bars that are born at $t = 0$; there is another bar which dies at the same time as one of these but is born later — it comes from the large blue area to the north of McAllen). It makes sense to think of H_1 classes, even those born at zero, which are not very persistent, as noise. This is because they represent areas which really are holes in the red-voting area, but the blue majority in those areas is so weak that there really isn't any reason the party should get representation just because they win that area.

Now that we have the expected result for the precinct-level barcode, we can repeat the same process at the district level[6]. Figure 8 is a diagram analogous to Figure 3, except this time the districts are colored instead of the precincts. Notice that in Figure 8, the intersection of the blue

[6] NB: since this uses presidential election data for the sake of consistency, the district-level diagrams do not necessarily match with the actual results of the House of Representatives elections.

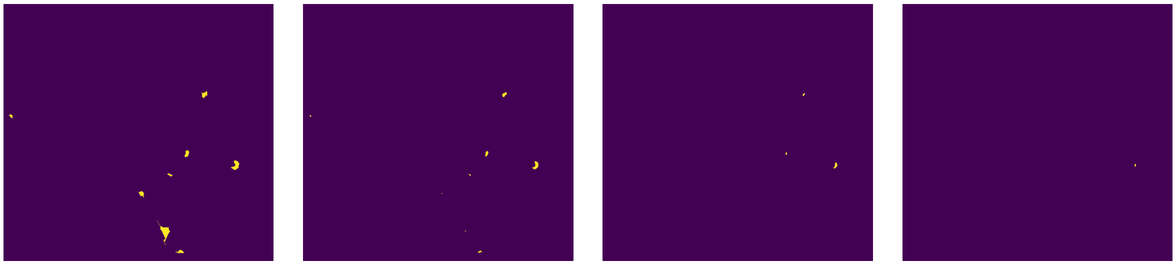


Figure 6. Steps $t = 1, 2, 4, 6, 8, 11, 15$, and 19 of the district-level levelset construction for the Texas 2016 presidential election

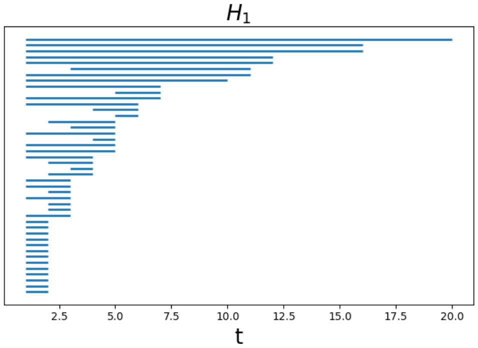


Figure 7. H_1 barcode for Texas 2016 presidential election precinct-level data

area with the actual Austin-area blue votes observed in Figure 3 and Figure 4 is pretty small. This is due to the same cracking of Travis county we saw in Figure 1 (which showed a different districting plan with the same intentions for Travis county). The point of the topological approach is that it will hopefully detect precisely this phenomenon. Indeed, this is what happens: in the barcode for the levelset filtered simplicial complex applied to the district-level initial configuration (Figure 9) we see only four very persistent bars: the two born at $t = 0$ correspond to Houston and Dallas (nothing we've done here, other than possibly the fact that the bars are somewhat less persistent than they were in the precinct-level barcode Figure 7, suggests that these areas are affected by partisan gerrymandering); the two which are born later correspond also to bars that existed in

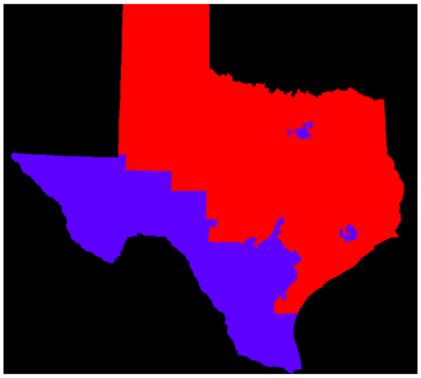


Figure 8. 2016 presidential election results by congressional district

Figure 7 since they come from the large blue area to the southeast being split into two halves (but NB since the precinct-level blue votes in that area are more fragmented, in Figure 7 those bars really correspond to McAllen and El Paso, while in Figure 9 they correspond to the large blue districts containing them; this distinction is not important, since district-level data is bound to exhibit this type of behaviour)

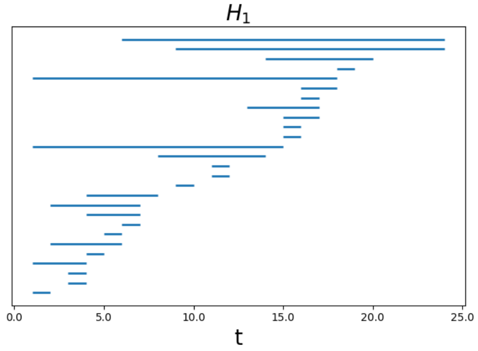


FIGURE 9. H_1 barcode for Texas 2016 presidential election district-level data

As we've already mentioned, it isn't completely clear what exactly the barcode says about the data. In particular, this doesn't guarantee that the cause of the dissapearance of persistent homology classes is gerrymandering in favor of any particular party: for all the barcode knows, it could have been gerrymandered so that a massive area connecting two cities became completely blue at the district level. But, it does say something which is perhaps not completely trivial, which is that the red districts of Texas are topologically not representative of the actual shape of the red region of Texas. This could suggest that the districts are not drawn in a representative way, despite not offering any indication of whose favor it is gerrymandered in. Note, too, that if these results were the result of a massive blue area connecting all the cities, then the levelset would eventually converge on these cities (since the local margin function $m(x)$ used in the district-level data is still taken from the precinct level) and cause many rather persistent bars, born somewhat later. This is exactly what happened to the bars on top of Figure 9, but is missing for the rest of the bars, which suggests that the topological differences observed by differences in persistent homology really are indicative of representation lost via partisan gerrymandering. Now that we've analyzed the output in this case, we go ahead and present the exact output in all 11 states considered. Those results are collected in Figure 10. We should still make some remarks on the interpretation of the barcodes. First, it is useful to completely ignore the homology classes which are of very short lifespan, since these represent areas which do not contribute many votes at all. We see then just by visually inspecting the barcodes that there is clearly some dropping of very persistent bars in Texas, North Carolina, Utah, and Ohio. These are all states in which large cities full of blue voters are cracked in the districting plan (e.g. Austin, Texas; Greensborough ad Fayetteville, North Carolina; Cincinatti, Ohio; and Salt Lake City, Utah). Even though the main strength of our method is its ability to detect cracking used against cities, we also see in the case of Maryland that we can observe the symptoms of a pro-Democrat partisan gerrymander there, due to the preservation of three persistent bars which are born later in the district-level barcode (in particular, it takes longer for the levelset to reach the urban areas represented by the bars; so this suggests that the Democratic district-level holdings are larger than they ought to be given what they are at the precinct level).

5. Further Directions

The examples in Section 4 show that our method has some limitations. We now go through three possible improvements to the ideas in this paper:

5.1. **Bottleneck distance.** Most of the time, answers to Question 1.1 require a 1-number summary of the data, while our method produces two whole barcodes as a summary. One natural way to produce a single number while following the ideas in this paper is to come up with a natural metric on the set of barcodes. The *bottleneck distance* lends itself well to this, since it depends on a matching between persistence barcodes (which is essentially what we do when we visually inspect the two barcodes and fiure out which bars correspond to each other). In particular, the bottleneck distance between two barcodes B_1 and B_2 is defined to be

$$\min_{f:A_1\to A_2} \max\left\{\max_{b\in A_1}\|b-f(b)\|_\infty, \max_{b\in B_1\sqcup B_2\setminus(A_1\sqcup A_2)} \frac{b_1-b_2}{2}\right\}$$

where we treat each bar b in a barcode as an ordered pair (b_1, b_2) of birth and death time.

5.2. **Multidimensional persistence.** One way to improve the statistical robustness of our method would be to protect against the case in which the voting outcomes are different than what the gerrymandering party estimated. In particular, despite a district plan being gerrymandered, it might still barely be won by party B. So we might want to very a parameter ϵ that we could add to the number of votes for party A in each precinct. Our construction is clearly also functorial in this ϵ, so we could use multidimensional persistence (a theory which is still developing and less clean than the one we reviewed in Section 1.1) [5] to try and include this extra information.

5.3. **Comparison to other measures of gerrymandering.** It would be instructive to see this output alongside other gerrymandering metrics like the efficiency gap or a compactness score, and to find some real-world examples in which one detects what the other cannot.

References

[1] Ulrich Bauer, Michael Kerber, Jan Reininghaus, and Hubert Wagner. Phat - persistent homology algorithms toolbox. *J. Symb. Comput.*, 78(C):76–90, January 2017. URL: http://dx.doi.org/10.1016/j.jsc.2016.03.008, doi:10.1016/j.jsc.2016.03.008.

[2] Gregory Bell, Austin Lawson, Joshua Martin, James Rudzinski, and Clifford Smyth. Weighted persistent homology. *Involve, a Journal of Mathematics*, 12(5):823–837, 2019.

[3] Mira Bernstein and Moon Duchin. A formula goes to court: Partisan gerrymandering and the efficiency gap. *Notices of the AMS*, 64(9):1020–1024, 2017.

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

[5] Gunnar Carlsson and Afra Zomorodian. The theory of multidimensional persistence. *Discrete & Computational Geometry*, 42(1):71–93, 2009.

[6] Moon Duchin. Gerrymandering metrics: How to measure? what's the baseline? *arXiv preprint arXiv:1801.02064*, 2018.

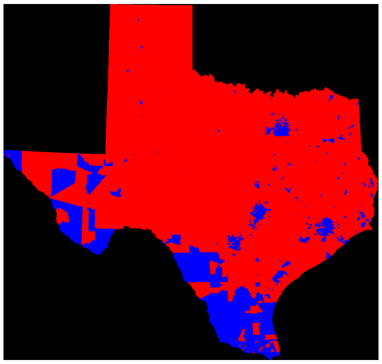
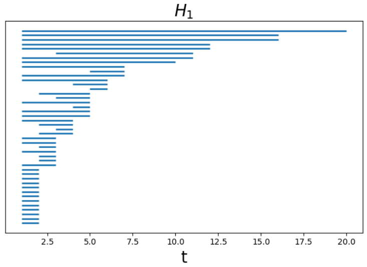
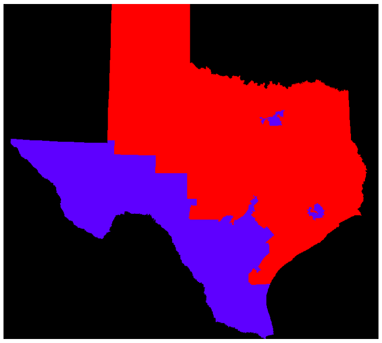
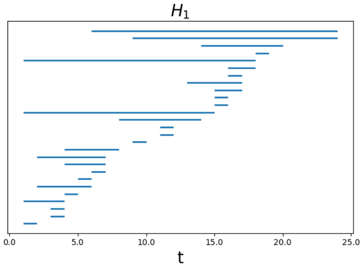
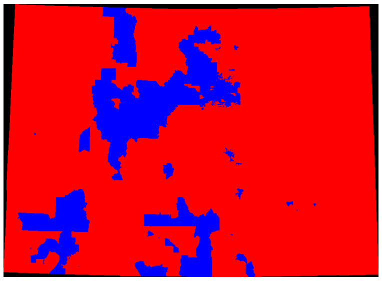
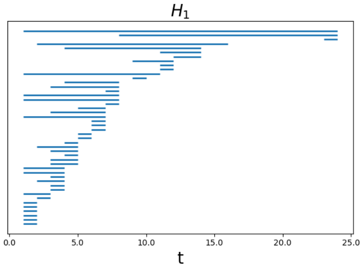
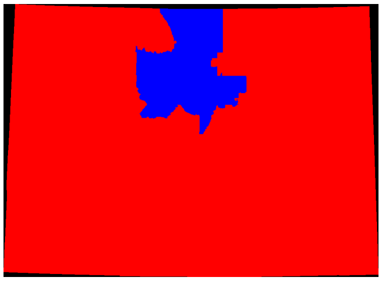
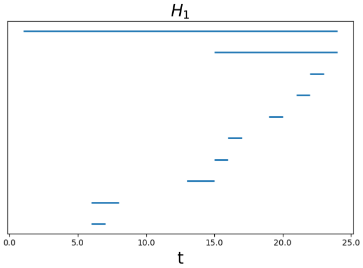
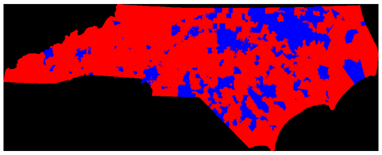
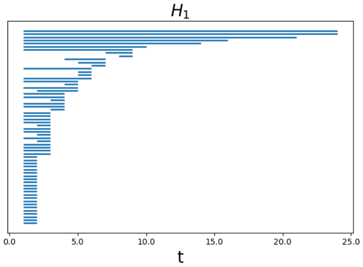
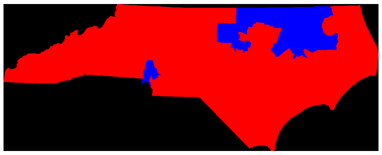
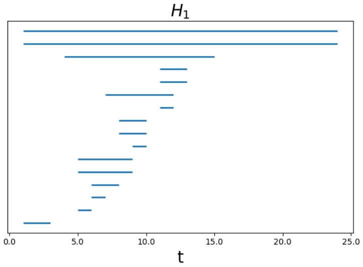
[7] Moon Duchin and Bridget Eileen Tenner. Discrete geometry for electoral geography. *arXiv preprint arXiv:1808.05860*, 2018.

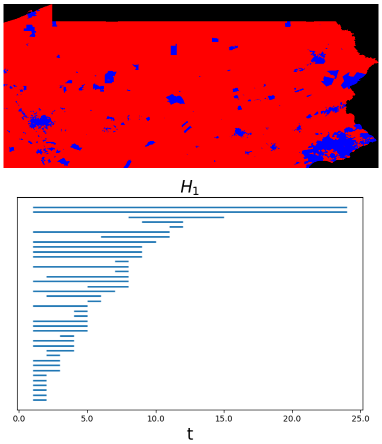
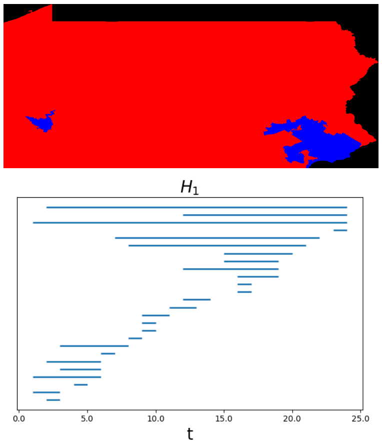
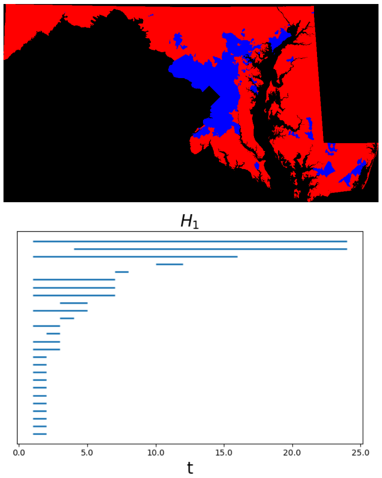
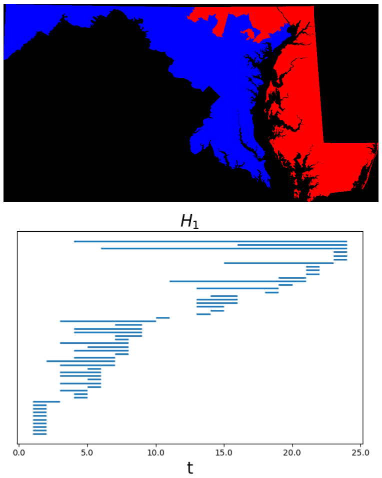
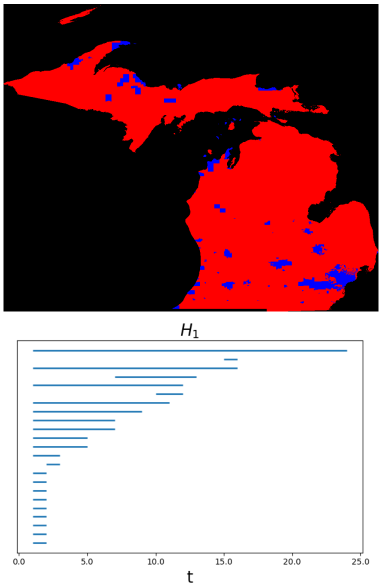
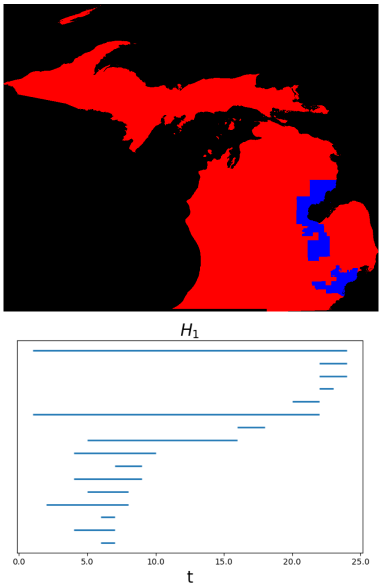
[8] Michelle Feng and Mason A Porter. Persistent homology of geospatial data: A case study with voting. *arXiv preprint arXiv:1902.05911*, 2019.

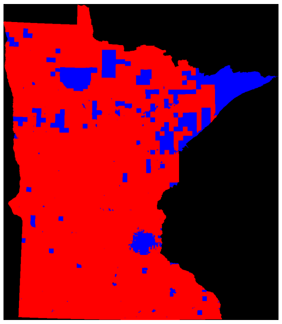
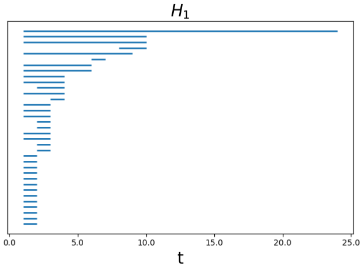
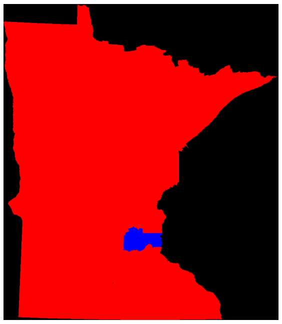
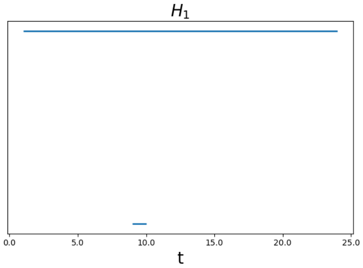
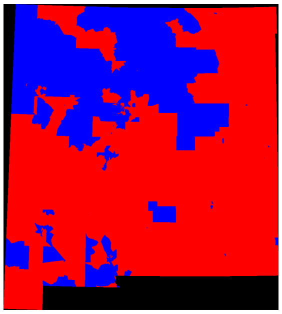
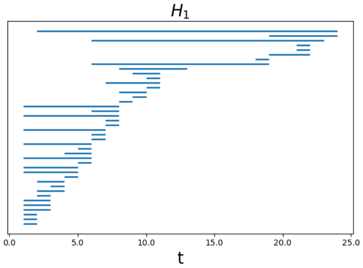
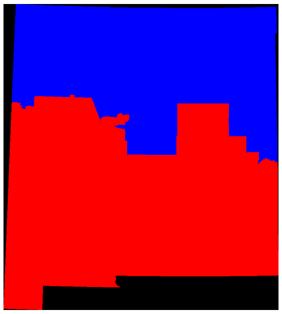
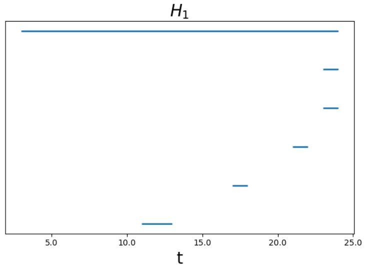
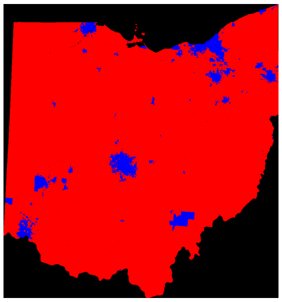
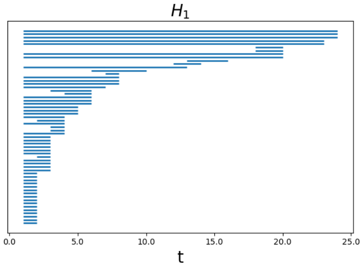
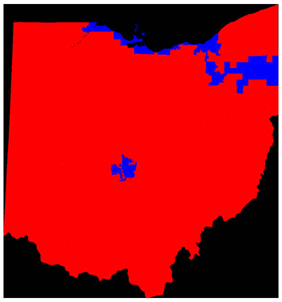
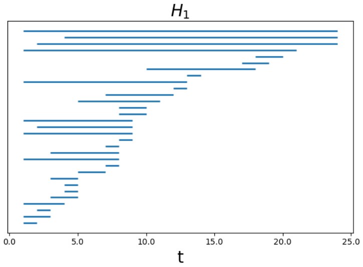
[9] Metric Geometry and Gerrymandering Group. Election geodata. https://github.com/mggg-states, 2013.

[10] Allen Hatcher. *Algebraic topology*. Qinghua University Publishing Co., Ltd., 2005.

[11] Nicholas O Stephanopoulos and Eric M McGhee. Partisan gerrymandering and the efficiency gap. *U. Chi. L. Rev.*, 82:831, 2015.

state	precinct-level	district-level
Texas	  	 
Colorado	  	 
North Carolina	  	 

state	precinct-level	district-level
Pennsylvania		
Maryland		
Michigan		

state	precinct-level	district-level
Minnesota	  	 
New Mexico	  	 
Ohio	  	 

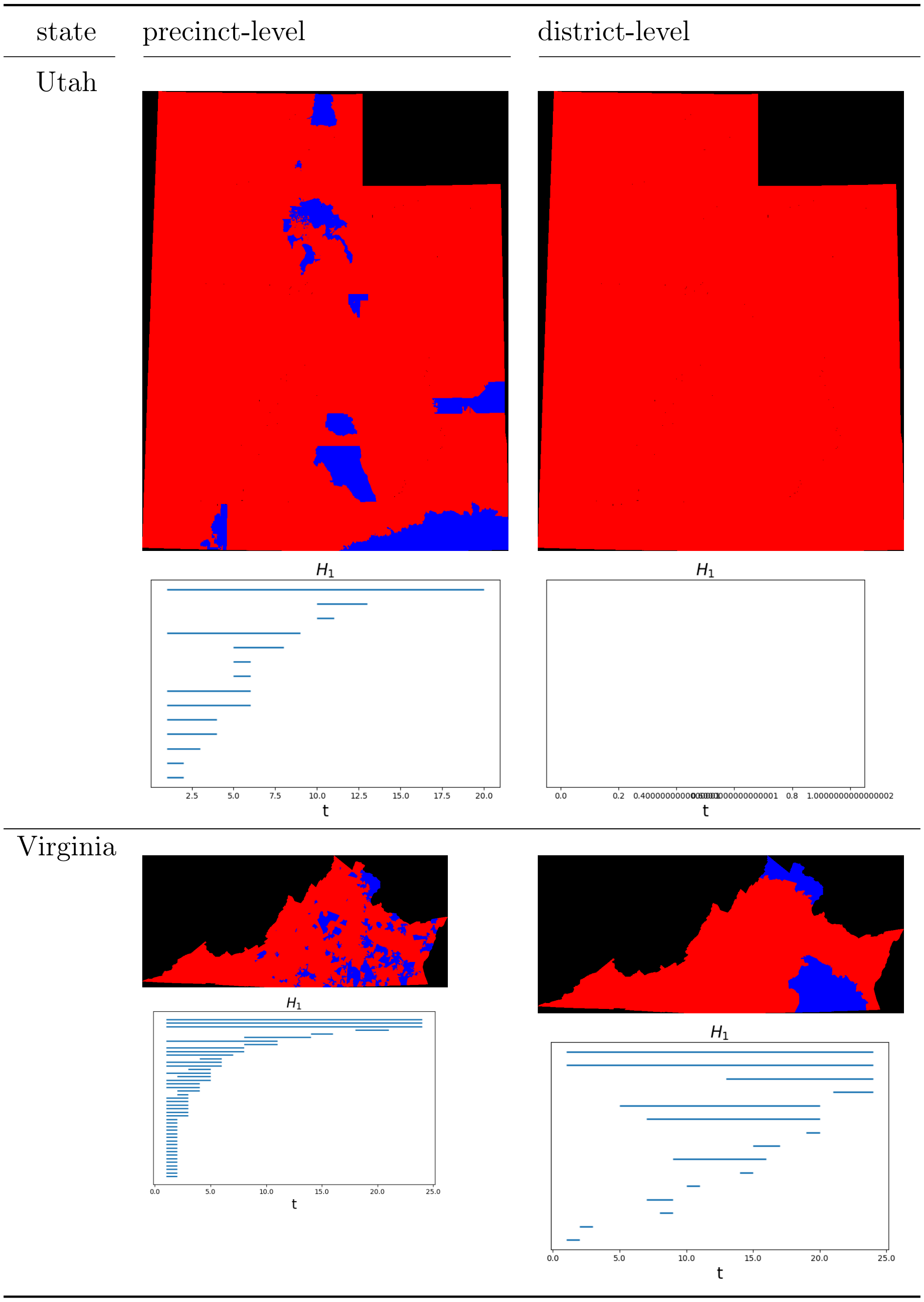


FIGURE 10. Barcodes and maps generated according to the procedures described in this paper.