THE HOMOLOGICAL PERSISTENCE OF POLICE VIOLENCE: ANALYSIS AND LIMITATIONS

DAWSON KINSMAN AND TIAN AN WONG

Abstract. Recently, collaborations between mathematicians and police departments have drawn criticism for their ethical implications and perpetuation of racial biases. More recently, in light of the Black Lives Matter movement, other researchers have used social networks and other methods to model police misconduct and violence.

This paper contributes to the quantitative study of police violence by applying for the first time persistent homology and topological data analysis (TDA) to police data. We utilize geospatial data from the Mapping Police Violence database in different approaches to modeling and analyzing fatal police violence in the US. While our analysis applies to national data, we consider several states as case studies and generate their Vietoris-Rips and Alpha filtrations in order to compute their persistent homologies. The results of our analysis do not reveal any novel trends or characterizations about our data, but we are able to determine clusters where police violence occurs and determine the stability of these clusters using persistence landscapes. In other words, our findings on the homological persistence of police violence confirms general expectations on where police violence is concentrated.

We also discuss the inherent methodological limitations of analyzing police data, also called Blue Data, and consequently the need from broader analyses involving other data sets in order to account from the full scale of the policing apparatus and the consideration of critical questions surrounding the function of policing and carcerality at large.

1. Introduction

On May 25, 2020, Derek Chauvin murdered George Floyd in the presence of three other officers, revitalizing criticism of the police and their excessive use of force. Police misconduct and violence has long been an issue in the United States, which has recently culminated in protests across the country and mathematicians questioning their work that support police departments. Predictive algorithms have been in use for decades now, but their application to criminal justice and mathematicians' collaborations with police departments remain highly controversial. These algorithms are meant to pick out patterns in the data and predict future incidents; however, the historical data fed to these algorithms are riddled with biases and racism that when paired with these algorithms, perpetuate the biases and racism that are already problems in communities. Furthermore, research has shown that the algorithms themselves contribute to the creation and worsening of negative feedback loops [8]. Both the data and choice of model are embedded in criminology and other disciplines, and to fully understand how to best model crime, collaboration and an understanding of historical biases are needed. At the same time,

Key words and phrases. Topological data analysis, police violence, persistent homology.

caution and care must be applied when interpreting the results of these models and placing them in context. While the adage "garbage in, garbage out" (GIGO) is well-known, many mathematicians argue that the use of these algorithms offer a "'scientific' veneer for racism" [4]. Moreover, since these predictive algorithms perpetuate pre-existing biases in data, there is the question of what these collaborations with the police and the use of police data, both collected by and about police departments, can tell us about crime that is not already intuitively known.

At the same time, the modeling of crime can be turned on its head, i.e., into modeling police behaviour. One recent topic of research in this area has been the modeling of police misconduct. Similar to predictive policing, police violence and excessive use of force are also crimes of a different kind, prompting the question of whether the methodologies police departments use to predict or analyze crimes can be used to analyze crimes perpetrated by members on the force. We do note however, that obtaining comprehensive data sets to model police violence and misconduct is a difficult ongoing effort. Previous research has utilized department and civilian complaints, incident reports, and 911 calls; however, because of the Blue wall of silence, in which police officers close rank to protect their own, it is frequently hard to establish a ground truth or extensive data set. Recent research by criminologists and applied mathematicians with social networks to model the spread of police misconduct have returned mixed results. A majority of the papers suggest that police deviancy spreads through social interaction [20, 23, 24, 21]; however, a recent paper utilizing survival analysis reports that police deviancy is dependent on an individual's traits.

In this paper, we analyse police misconduct in the particular form of fatal police violence, utilizing open-source police violence data from the Mapping Police Violence database, and taking hot spot analysis a step further, we conduct topological data analysis (TDA) to characterize patterns in police violence. TDA is a relatively recent computational tool that studies large scale features in potentially high-dimensional data. In the realm of social analysis, it has found applications such as in the study of gerrymandering [11, 13], tax property [10], and COVID-19 [16]. Notably, these studies analyze datasets that are include spatial characteristics, which we also do in our study. We conduct here an initial assessment of the potential uses of TDA in the study of police violence, and in doing so encounter a larger methodological question regarding the limits of working solely with police-related data, which we refer to as Blue Data.

In Section 2, we introduce works related to both collaborations between mathematicians and criminologists to model police deviance and geospatial applications of topological data analysis. Next, in Section 3, we discuss the main methods that we use, namely persistent homology, persistence diagrams, and persistence landscapes, as well as their results. We summarize the Vietoris-Rips and Alpha filtrations which are used to generate the filtered simplicial complexes whose persistent homologies are computed. In Section 4 we provide interpretations of the results and secondly, an extended discussion on the potential uses and methodological limitations in quantitative analyses involving police violence data. In particular, we argue that a broader, more comprehensive framework, both in terms of data and approach is needed to examine the carceral system at large.

2. Background

2.1. **Related Works.** Much of the recent research surrounding police misconduct have used civilian and departmental complaints against police officers to model misconduct on the force, such as Oullet et al. [20], Wood et al. [23], Zhao and Papachristos [24], and Quispe-Torreblanca and Stewart [21]. Most recently, Simpson and Kirk [22] utilize 911 calls in conjunction with complaints to model the spread, or lack thereof, of police deviancy. Furthermore, [20, 23, 24, 21] utilize social networks to model the police deviancy and spread, and conclude that police misconduct spreads through social interactions similar to a contagion; however, Simpson and Kirk [22] use survival analysis to suggest that it is the traits of individual officers that more clearly associate which officers are deviant, not social interactions.

Also, the application of mathematical and statistical models to predictive policing has been subject to much criticism from its inception from mathematicians and other professionals. Lum and Isaac [19] have shown that the PredPol algorithm picks up on biases existing in police crime data, creating feedback loops that can then lead to increased over-policing in areas that have historically experienced over-policing. While scholars are well familiar with the GIGO principle, Akpinar et al. [2] have shown that even with the use of victim crime reporting data, the misallocation of police resources can still occur as the displacement of predicted hotspots shift from areas with high crime and low reporting rates to areas with high to medium crime and high reporting rates. Ensign et al. offer a different model for predictive policing based on the partial monitoring framework from machine learning that allows them to obtain regret bounds. However, Chapman et al. [8] embed these algorithms in criminological theory and show that the choice of a model and the construction of these prediction algorithms can create biased feedback loops even with random data.

Lastly, Feng and Porter demonstrate an application of topological data analysis with voting data in [13]. They generate simplicial complexes from the geospatial data using traditional distance-based methods, such as the Vietoris-Rips and Alpha complexes, and also introduce two new methods for generating filtered simplicial complexes from geospatial data. Adjacency complexes are based on network adjacencies and have the advantage of retaining spatial information and a notion of contiguity that may otherwise be lost as a point cloud; however, the adjacency complex still requires each polygon to be contracted to a single point as opposed to considering the entire area the polygon covers [13]. On the other hand, level-set complexes more closely resemble underlying maps from which they were generated and persistence has intuitive interpretations, but still suffer from sensitivity to scale similar to the distance-based constructions [13]. Their results show that both adjacency and level-set complexes do a better job than traditional methods at identifying real features instead of noise when utilizing geospatial data or geographic maps.

2.2. **Data.** The data used comes from the open-source database Mapping Police Violence, one of the most comprehensive police killings databases.[1] Mapping Police Violence aggregates its data from Google News, the Washington Post, and Fatal

[1]Data was obtained on January 5, 2022 and consists of 10,927 records spanning from January 1, 2013 until October 18, 2022.

Encounters, another police violence database that also uses crowd-sourced data and public records requests. Police violence, defined by the Mapping Police Violence project is "Any incident where a law enforcement officer (off-duty or on-duty) applies, on a civilian, lethal force resulting in the civilian being killed whether it is considered 'justified' or 'unjustified' by the U.S. Criminal Legal System"[1]. The database also includes vehicular force and suicide by cop incidents. Suicide by cop occurs when a person having a mental health crisis or characterized as suicidal are killed by police [1]. Previous research comparing MPV and other police violence datasets concluded that the government database of police violence the National Vital Statistics System was missing or misclassified over half of the police related deaths estimated to have occurred from 1980-2018 [9]. Thus, we chose to use the MPV dataset since it is one of the most comprehensive databases with data from several years. In our research, we utilize the geographic coordinates, the state in which the incident occurred, and the date of each case. We assume each state to be a point cloud with each case being a point.

3. Methodology

3.1. **Persistent Homology.** To motivate our use of topological data analysis, we discuss persistent homology and the two different types of simplicial complexes used.[2] We first start by discussing simplicial complexes and simplicial homology upon which persistent homology is built. Given a set of affinely independent points X, we define the *simplex* σ spanned by X to be the convex hull of X, and the *dimension* of σ, denoted $dim(\sigma)$, is $|X| - 1$. For example, a single point is a 0-simplex, an edge is a 1-simplex, a 2-simplex is a triangle, etc. A *face* of σ is a simplex spanned by a nonempty subset of X. Combining these notions, define a *simplicial complex* as a space Σ that is a union of a list $\mathcal{L}$ of simplices such that if any simplex is included in $\mathcal{L}$, then so is any face of it, and for any two simplices in $\mathcal{L}$, the intersection is a face of both simplices [7]. Simplicial complexes can then be represented algebraically as chain complexes that are defined as follows. We denote the *chain complex* associated with Σ as

$$C(\Sigma) := (C_k(\Sigma), \delta_k)_{k\in\mathbb{Z}},$$

where δ_k are a sequence of linear transformations called *boundary maps* and $C_k(\Sigma)$ are a sequence of vector spaces called *chains* such that

$$C_k(\Sigma) \xrightarrow{\delta_k} C_{k+1}(\Sigma)$$

and $\delta_k \circ \delta_{k+1} = 0$. Finally, the k^{th} *homology group of* Σ ($H_k(\Sigma)$) is the quotient group of the k-boundaries, denoted $B_k(\Sigma) = \text{im}(\delta_{k+1})$, modulo the k-cycles or $Z_k = \ker(\delta_k)$; that is,

$$H_k(\Sigma) := \frac{Z_k(\Sigma)}{B_k(\Sigma)} = \frac{\ker(\delta_k)}{\text{im}(\delta_{k+1})}.$$

For example, $H_0(\Sigma)$ counts the number of connected components in Σ, $H_1(\Sigma)$ counts the number of loops in Σ, etc. That is, $H_k(\Sigma)$ counts the number of k-dimensional holes.

Given a simplicial complex Σ, a subcomplex of a simplicial complex is a collection of simplices belonging to Σ that also satisfy the aforementioned conditions. We can

[2]We turn readers to [15, 3] for an introduction to persistent homology and to [12, 7] for more details on persistent homology and topological data analysis.

then define a *filtered simplicial complex* $\{\Sigma_i\}$ as a finite sequence of subcomplexes such that $\emptyset = \Sigma_0 \subseteq \Sigma_1 \subseteq \ldots \subseteq \Sigma_k = \Sigma$. Now that we have a filtered simplicial complex, we can define the m^{th} persistent homology of X [13].

Definition 3.1 (m^{th} Persistent Homology, Persistent Homology)**.** *Let Σ be a filtered simplicial complex. The m^{th} persistent homology of Σ is*

$$(\{H_m(\Sigma_i)\}_{1\leq i\leq l}, \{f_{i,j}\}_{1\leq i\leq j\leq l})$$

where $0 \leq i \leq j$ and both i and j are less than the dimension of X. Additionally, let $f_{i,j}$ denote the map from $H_m(\Sigma_i)$ to $H_m(\Sigma^j)$ that is induced by the inclusion map $\Sigma_i \hookrightarrow \Sigma^j$. The persistent homology of Σ is the set of all m^{th} persistent homologies of Σ.

Using the persistent homology of Σ, it is possible to summarize homological features through barcodes, persistence diagrams, and other methods.

3.1.1. *Generation of Simplicial Complexes.* Given a point cloud $X \subset \mathbb{R}^2$, such that every point is an incident, it is possible to use the Vietoris-Rips (VR) complex and the Alpha complex to generate filtered simplicial complexes. Additionally, since both constructions are based on distance parameters or the pairwise distance between points and we are dealing with geospatial data, we only consider Euclidean distances. The VR complex is popular to use as it is straightforward to construct since we only need to compute pairwise distances, and has some nice theory from the Nerve Theorem relating it to the Čech complex. We present the definition of the VR complex from [13] as follows.

Definition 3.2 (Vietoris-Rips (VR) Complex)**.** *Given a real number $\epsilon > 0$ and a point cloud X, the Vietoris-Rips complex $VR_\epsilon(X)$ as*

$$VR_\epsilon(X) = \{\sigma \subset X : \forall x, y \in \sigma, d(x, y) \leq \epsilon\}.$$

Using a set $\{\epsilon_i : 0 < \epsilon_i < \epsilon_{i+1} \text{ for } i = 1, \ldots, k\}$, such that $X \subseteq VR_{\epsilon_1} \subseteq \ldots \subseteq VR_{\epsilon_k}$, we have generated a filtered simplicial complex whose persistent homology we can compute. However, given a point cloud X, in the worst case a VR complex can have up to $2^{|X|} - 1$ simplices and dimension $|X| - 1$. Consequently, similar to Feng and Porter [13], we decide to construct the Alpha complex for states with over 200 incidents instead of the VR complex, as it too relies on a distance parameter. For states with less than 200 cases, we generate both the Alpha and VR complexes to compare their persistence scales. The definition of the Alpha complex is as follows [13].

Definition 3.3 (Alpha Complex)**.** *Let $\epsilon > 0$ and $B(x, \epsilon) = \{y \in X : d(x, y) \leq \epsilon\}$ be the epsilon ball with center x and radius ϵ, where*

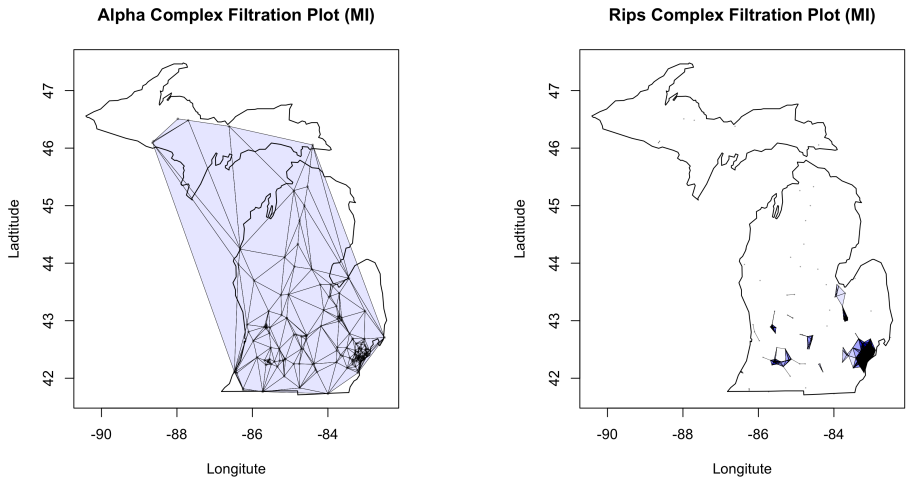
$$X_\epsilon := \bigcup_{x\in X} B(x, \epsilon).$$

Additionally, let $(V_x)_{x\in X}$ be the Voronoi diagram of X. Consider the intersection $V_x \cap B(x, \epsilon)$ for each $x \in X$, and note that the collection of sets covers X_ϵ. Thus,

$$A_\epsilon(X) = \{\sigma \subset X : \forall x_i \in \sigma, \bigcap_i (V_x \cap B(x, \epsilon) \neq \emptyset\}$$

.

As noted by Feng and Porter [13], since our data is in $\mathbb{R}^2$, the highest dimension simplex will be 2D simplices. An additional benefit to using these constructions is that there are many packages available to easily generate both types of complexes. The generation of simplicial complexes, persistence diagrams, and persistence landscapes were implemented using the `GUDHI` package for Python, and for visualizations, we use the `TDA` package for R which provides an R interface for `GUDHI` and other TDA packages. Consider the simplicial complexes of both the VR and Alpha complexes for Michigan, a state with 184 incidents, shown in Figures 1b and 1a, respectively. To limit the complexity of the VR complex, for this visualization,



(A) Alpha complex (B) VR complex

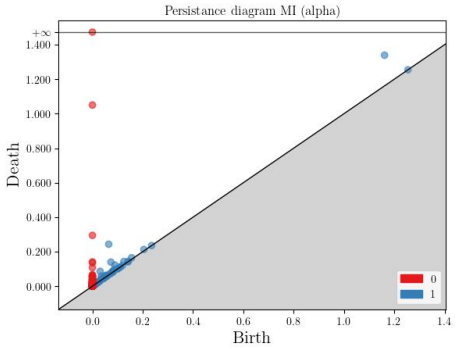
FIGURE 1. Visualizations of filtered simplicial complexes for Michigan

we restrict the maximum ϵ value to 0.25 and the maximum homology dimension to one.[3] Both plots of the complexes reveal where incidents are clustered, and return similar clusters around the cities of Detroit, Flint, Kalamazoo, and Grand Rapids. The location of the clusters are to be expected, since these are larger cities in Michigan known to have higher crime rates than other cities in Michigan.

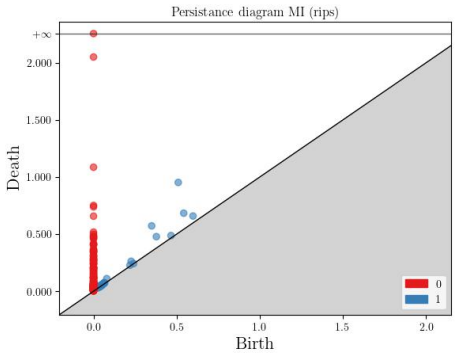
3.2. **Persistence Diagrams.** Persistence diagrams are one of the main tools in TDA based on the persistent homology of a filtered simplical complex. The diagrams lies in the first quadrant of the (x, y)-plane and consists of *persistence pairs* (b, d) where b and d are the steps from a filtered simplicial complex in which a homological feature is born and dies, respectively. The death time of the persistence pairs are determined when a younger feature is absorbed by an older homological feature and dies while the older persists. Persistence diagrams thus show the birth and death of homological features. We generate persistence diagrams for both the

[3]Code for the visualizations is adapted from the documentation for the TDA package for R [17], [18].

VR complex and Alpha complex since we expect that both the VR and Alpha complexes to produce a similar amount of holes and to have similar scales. However, we see that this is not the case for some states.



(A) Using the Alpha complex

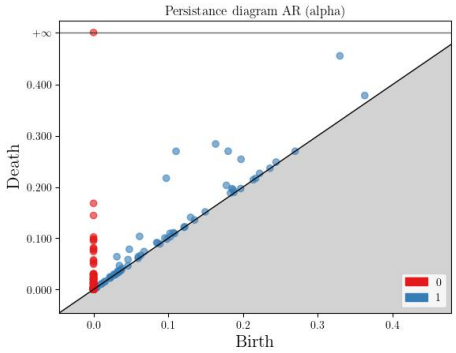


(B) Using the VR complex

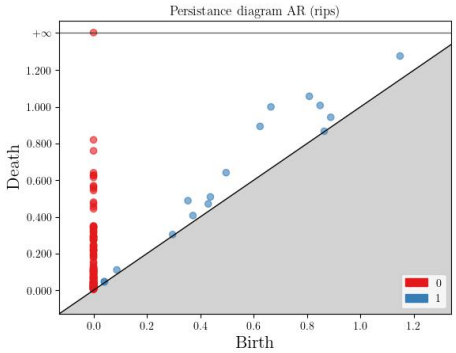
FIGURE 2. Persistence diagrams for Michigan

In Figure 2, we generate the persistence diagrams using both the Alpha and VR complexes for Michigan to compare their persistence scales. The persistence diagrams for Michigan have slightly different persistence scales, with homological features in the VR complex persisting longer than those in the Alpha complex. However, the Alpha filtration has a few holes that are born later in the filtration that do not persist long. Additionally, in Figure 2a the cycles lie close to the diagonal meaning that the cycles born die soon after. This suggests that the scale may be too large such that the holes never appear or that our data simply has no holes. Conversely, when looking at Figure 2b, there are more 1-dimensional homologies that persist for longer. That is, with the VR filtration, our holes do not close up as quickly as they do when considering the Alpha complex.

We also include the persistence diagrams for Arkansas, which has 142 cases, in Figure 3, as the homologies from the two filtrations display different behavior than those from Michigan. Notably, the persistence scales are much different between the Alpha and Rips filtrations. In Figure 3a the largest scale, besides infinity, is slightly above 0.4, but for the VR complex, the persistence scale goes up slightly



(A) Using the Alpha complex



(B) Using the VR complex

FIGURE 3. Persistence diagrams for Arkansas

past 1.2. Moreover, there are still many 1-dimensional homologies being born and dying almost immediately after. Again, this suggests that we might want to look at a smaller persistence scale which would cause the difference in scales to be exacerbated, and the cause of this difference in scales should be investigated in further research. Despite their difference in scales, we note that both the Alpha and VR filtrations for Arkansas have some 1-dimensional homological features that persist for some time after they are born and lie further from the diagonal.

Persistence diagrams are a traditional tool of TDA; however, persistence diagrams lack compatibility with many machine learning and statistical methods. Thus, persistence landscapes are additionally generated.

3.3. **Persistence Landscapes.** A persistence landscape, introduced by Bubenik in [5], is an alternative topological summary that is a piece-wise function which allows TDA to be combined with statistical analysis or machine learning. Similar to persistence diagrams, landscapes are also derived from the persistence module, and the construction of persistence landscapes from a persistence diagram can be intuitively understood as follows. Given a persistence diagram and disregarding any points that persist to infinity, at each point, we can draw vertical and horizontal lines from each point to the diagonal line $b = d$. These lines serve as our tent functions for each point. Next, we want to rotate the persistence diagram $-\frac{\pi}{4}$

such that the diagonal is now the x-axis. The k^{th} *persistence landscape* is the k^{th} pointwise maximum of these tent functions. The *persistence landscape* is more formally defined as follows [6].

Definition 3.4 (k^{th} Persistence Landscape)**.** *Given a real birth-death pair (b, d), the associated tent functions are the piece-wise linear functions $f_{(b,d)}(x)$ as follows.*

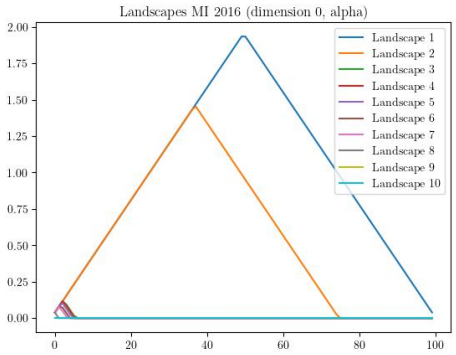
$$f_{(b,d)}(x) = \begin{cases} 0 & \text{if } x \notin (b, d) \\ x - b & \text{if } b < x \leq \frac{b+d}{2} \\ -x + d & \text{if } \frac{b+d}{2} < x < d \end{cases}$$

Given a set of birth-death pairs $\{(b_i, d_i)\}_{i=1}^n$, the tent functions form the set $\{f_{(b_i,d_i)}(x)\}_{i=1}^n$. The k^{th} persistence landscape *function, denoted $\lambda_k(x)$ is the k^{th} largest of $\{f_{(b_i,d_i)}(x)\}_{i=1}^n$, and for $k > n$ where no such maximum exists, $\lambda_k(x) = 0$ The sequence of functions λ_k form the* persistence landscape *of the given set of birth-death pairs.*

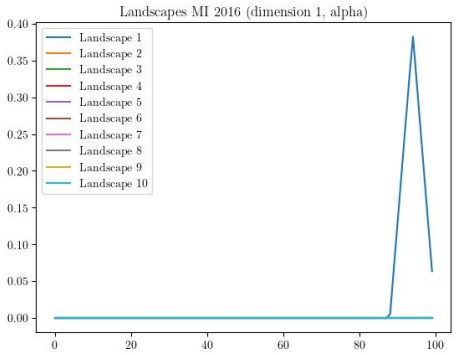
Persistence landscapes have some theoretical properties that give them an advantage over other topological summaries [5], and more specifically, we are interested in is their ability to be used with statistics. While persistence landscapes are functions, they can be "vectorized" for machine learning or statistical analyses by evaluating the landscape functions on a grid of points along the x-axis. In this section, we generate the top ten landscapes using the Alpha complex from 2016 and 2022. The Alpha complex was chosen since it can be generated for all states which gives us the ability to compare landscapes from all possible states in furture research. We then take the point-wise average of the top ten landscapes for Michigan in each year from 2013-2022 and compare all of the landscapes computed to comment on the stability of the homological features.

Figure 4 shows that only two 0-dimensional features persist for some time since all of the other landscapes return to the x-axis soon after they leave it. This means that connected components are quickly connecting with other components and dying. Similarly, for 1-dimensional homologies, there is only one loop that forms that persists for a short period of time. No other loops are created, and the one loop that persists, is born later in the filtration and its death occurs past the scale that all of the top 0-dimensional homologies persist in. Additionally, the scale of the two landscape plots for 2016 have different y-axis scales, and overall, the one cycle has a much shorter persistence that is more comparable with the components that are born and die very quickly than to the top two landscapes of the 0-dimensional features.

The persistence landscapes from 2022 in Figure 5 contain homologies with more variation in their birth and death times. While only one feature persists for a substantial amount of time, there is more variation among the other landscapes meaning that the homologies that these landscapes correspond to (and we know they correspond to certain homological features since the landscapes appear to be the tent functions) are born and die slower than the top 0-dimensional features in 2016, but they still die relatively quickly. Similarly, the persistence landscape plot for 1-dimensional features from 2022 is still dominated by one tall tent function corresponding to a feature that is born later in the filtration. However, we again note that the landscape plot for cycles has a scale much smaller than its 0-dimensional counterpart; moreover, comparing Figure 4b to Figure 5b, we note that the scale of the plot from 2016 to 2022 has also decreased in this comparison as well. Thus,



(A) H_0 features

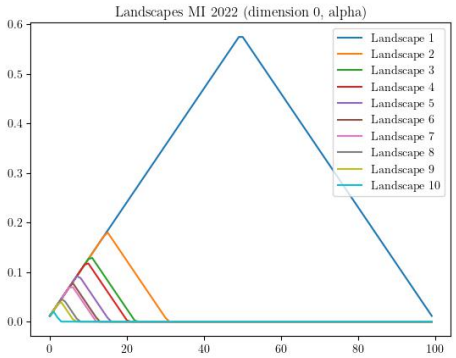


(B) H_1features

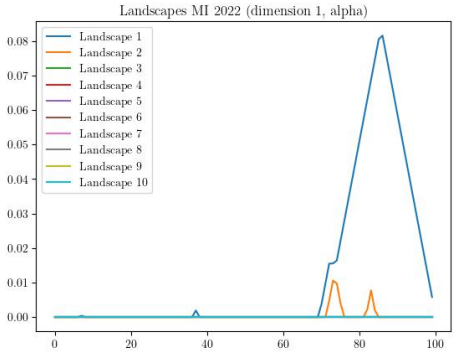
FIGURE 4. Top 10 persistence landscapes for Michigan in 2016

even the 1−dimensional features that persist the longest from 2022, die faster than the one feature of note from 2016.

Lastly, we compare the top ten landscapes from 2016 and 2022 to the point-wise average of the top ten landscapes from 2013-2022 (Figure 6). As with the two individual years that were examined, the average landscape plot for 0- dimensional features is dominated by a single component persisting much longer than any other. Besides the second average landscape, all other persistence landscapes quickly approach zero after becoming non-zero. This suggests that most of the components die soon after they are born. The second average landscape persists for some time but does not get very tall which suggests that in the average, there may be some years when the second persistence landscape substantially persists; however, based on the scale of Figure 6a, it is likely that more often the second landscape resembles the second landscape in Figure 5a or even persists for a shorter amount of time. Similar remarks can be made for the 1-dimensional landscape plot (Figure 6b). There is one cycle of note that appears to persist much longer than all other cycles; however, the scale of the plot is much smaller than that of the 0-dimensional plot; thus, when contextualizing the cycles among the total persistent homology of the filtered Alpha complex, they die rapidly after they are born.



(A) H_0 features



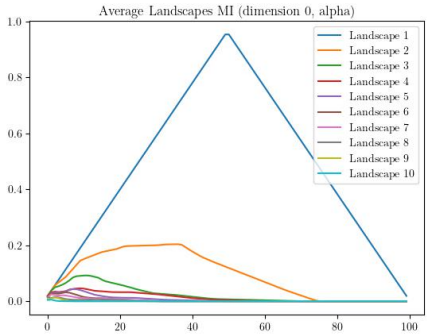
(B) H_1 features

FIGURE 5. Top 10 persistence landscapes for Michigan in 2022

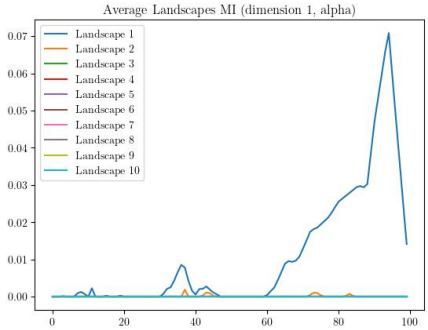
4. Discussion

4.1. **What data can tell us.** What do persistence scales and homological features dying soon after their born suggest? We start with discussing the latter. When homological features die soon after they are born, it means that soon after the feature is formed, it gets "absorbed" or connected to another feature, and since we are dealing with geospatial data, this suggests that the cases are occurring geographically close together. Thus, as soon as we start expanding the epsilon balls in our filtrations, the epsilon balls soon start intersecting and components start connecting. Any loops or cycles in our data are small enough that they are filling in soon almost immediately after formed, which once more implies the clustering of our data in certain areas. This gives further quantitative support to the general expectation that police violence tends to cluster and remain concentrated in specific areas.

Regarding the former, we expect the persistence scales to differ in two ways: (1) between states, and (2) between years for a state. More evidently, we expect different states to have different persistence scales as the geographic locations of police violence varies between states, and states with larger populations or big cities will likely experience more clustering of police violence than more rural states. Furthermore, for individual years, we expect more variation among persistence scales since



(A) H_0 features



(B) H_1 features

FIGURE 6. Average of Top 10 persistence landscapes for Michigan

we have a smaller data set. Our case study of Michigan had 184 total observations from $2013-2022$, and so, we suspect that analysis of states with more cases, such as California or Texas, may have more stable persistence scales amongst the individual years. And, while the average persistence landscape does offer a summary of the top persisting homological features across the years, we have shown that for states with a small number of cases each year, the persistence scale can widely vary between years, which while still displaying homological features that persist for a long period of time, can make some interpretations difficult to generalize. Finally, we also unexpectedly experienced a difference in persistence scales in our topological summaries when using the Alpha and Vietoris-Rips Complex. Experiments suggest that this difference can be mitigated by an adjustment of the respective persistent scales used, and hence this should not be attributed as an inherent feature.

4.2. **What the data can't tell us: the limits of Blue Data.** What can Blue Data tell us, and what are its limits? By Blue Data here we mean police data, broadly speaking, both the data obtained from and about police. As activists such as Data for Black Lives (D4BL) founder Yeshimabeit Milner has argued, the genealogy of data in modern policing can be traced back to slavery in the United States; in this sense, the use of data-centered approaches in policing should not be thought of as a new phenomenon but instead entirely resonant with older forms of policing.

At the same time, the advent of algorithmic approaches to surveillance and policing at large presents new challenges at an unprecedented scale. We focus here on three particular aspects of Blue Data: mystification, scale, and patterns.

4.2.1. *Mystification.* We begin first with a cautionary tale involving the PredPol itself, which may be considered to be the poster child of predictive policing in an algorithmic age. In a 2021 report 'Automating Banishment' produced by the Stop LAPD Spying coalition, it was revealed through a public records request from the LAPD that law professor Andrew Ferguson corresponded with then Deputy Chief Sean Malinowski in 2019 who was leaving to work in private and academic sectors. One of PredPol's founders Jeff Brantingham, according to Ferguson, faced heavy criticism because PredPol "misread the sentiment about how to think about predictive policing and didn't pivot to a more police accountability focus in time," and for the fact that Brantingham's UCLA affiliation was used by LAPD and PredPol in their marketing. Ferguson offered to help Malinowski in his transition and approach to predictive policing in the private sector.

Ferguson's principal argument was that the analysis used in predictive policing, viz., crime data, could also be applied to the effect of police accountability. The latter being more acceptable in the current political climate, it would be better to market the system under the veneer of police accountability, when in fact it also comes with predictice policing capabilities. Freguson made such a pitch in 2018 to HunchLab, now owned by ShotSpotter, asking "Have you guys thought about spinning out a new product (not predictive policing) but branded solely for police accountability?" Similarly, Ferguson suggested that Brantingham should brand his papers by "showing that you can balance race or other factors as a technical matter and it is all about how the police (not the companies) choose to calibrate the algorithm." Indeed, this was exactly what Ferguson suggested in his own paper, suggesting that "the same big data policing technologies built to track movements, actions, and patterns of criminal activity could be redesigned to foster data-driven police accountability" [14].

Several observations are in order: The first is that the promise of police accountability through the deployment of an internal system doubly mystifies the process of accountability: the algorithmic black box by which accountability is held, often through certain early-intervention systems, and the lack of external accountability and review of the accountability and intervention process. Thus Blue Data as such reinforces the 'Blue Wall' of silence. To be certain, this has more to do with the policy implications in response to the analysis of Blue Data rather than the analysis itself. Still, a related problem is the fact that an increase in crime is not easily distinguished between an increase in the occurrence or in the detection of certain incidents, such as by increased policing in a given neighborhoods.

Second, and perhaps more to the point, the supposed fungibility of police and crime data shows on the one hand the risks involved in narrow-minded approaches to accountability, namely that those who advocate for internal police accountability mechanisms, whether in the form of early intervention systems, body cameras, and the like, in fact expand police surveillance capabilities while at the same time disavowing it under the cover of accountability. (Body camera footage, for example, have been used more often as evidence in criminal courts than for addressing police misconduct.) Indeed, an easy review of citizen complaints of police misconduct

at any major police department readily reveals the small fraction at which such complaints translate into accountability.

On the other hand, a more subversive interpretation of the analysis of police data as crime data is perhaps more revealing: that is, that police misconduct, excessive use of force, and violence operates in the exact same manner as organized crime — only one is state sanctioned and one is not. We propose that such a hypothesis be tested far more extensively than has been done so far by researchers, in particular applying criminological analysis to police data such as modeling police violence as contagion, or as in the case of PredPol, as a self-exciting Hawkes process.

4.2.2. *Scale.* Our second argument regarding the limits of Blue Data is perhaps more important. First, it is readily acknowledged that just as crime data is often incomplete and unstructured, Blue Data — which includes not only crime data but police data at large — is even more so. Protected by the Blue Wall, Blue Data is often reported largely to the extent that it supports the function of continued policing, though again, as we have pointed out above, this is also dependent on interpretation and presentation. Importantly, we note that arguments for the expansion of policing need not follow an internally consistent logic — for example in a recent debate surrounding the contract renewal of ShotSpotter in Detroit in 2022, it was both argued that shootings in proximity to ShotSpotter detection locations will decrease because potential offenders will avoid such locations *and* that the locations where ShotSpotter is deployed are not revealed publicly for the security of police operations.

Second, much Blue Data that is available is made accessible only through expensive Freedom of Information Act (FOIA) requests. While in principle such information is free to be requested, in practice such information takes many months and thousands of dollars to procure, and unless the request is carefully specified to high precision, the resulting information that is provided is often lacking and insufficient for proper analysis.

Third, building on the previous points of the limited nature of Blue Data itself, is contrast with the actual scope of the present expansion of the policing apparatus. To take a simple example, while almost all analysis study only one or at most two types of datasets in conjunction with each other, for example the use of certain surveillance systems like ShotSpotter, predictive systems like PredPol, or police misconduct data, present analysis have fallen very short of confronting the fact that all these systems in communicate with each other, and the present push in policing technology is to integrate all these systems into large 'fusion systems' or 'correlation engines.' As the Stop LAPD Spying Coalition revealed about the LAPD, police departments maintain enormous information sharing networks, ranging from data obtained from the FBI and LA County departments such as the Department of Transportation, the Department of Health Services, the Department of Mental Health, and the Department of Child and Family Services. Hence the scale of the existing police network poses an immense methodological problem for the analysis of Blue Data: despite the limit amounts of information available, and not to mention the messiness of the available data, how can quantitative approaches analyze the interconnections of policing apparatus on the scale at which is exists in reality? How can such a vast network be effectively modelled and studied? What conclusions can be gained?

4.2.3. *Patterns.* The question of what new insight is there to be gained from the analysis of Blue Data is in fact an extremely difficult one. In the preceding analysis we have presented an application of topological data analysis to police violence data — though we have not included it here, we have also experimented with applying Lum and Isaac's reconstruction of the PredPol algorithm [19] to the same MPV data without any conceptual blocks — our basic analyses suggest that one should generally expect results to simply confirm general suspicions on how policing in the US operates. In other words, the analysis of Blue Data at large will not, we argue, yield many surprising results. (Of course, technical studies that confirm such expectations still play important roles at the level of policy and public health; a disproof of a null hypothesis has inherent value.) In some sense, this should not be surprising as many of the mathematical tools deployed can be understood as pattern recognition techniques, though it may be that certain outliers may prove interesting to study on their own.

Indeed, a possible interpretation of applications of topological data analysis to the very different problem of political gerrymandering is that the latter is too subtle a problem to be detected by such tools, whereas police violence lies on the other end of the spectrum: it is a largely known phenomenon that the patterns detected by persistent homology will generally be accurate but also expected.

On the question of scale, it is equally crucial to note that we have so far only discussed Blue Data in the strict context of policing, whereas critical theorists such as Ruth Wilson-Gilmore have argued, the policing apparatus sits in a much larger context of *abolition geographies*, which includes the prison industrial complex, capitalism, and the state. A simple example is the so called school-to-prison pipeline, wherein broken systems of education, policing, justice, and prison all cooperate to produce and reproduce criminality without reform.

The more pressing question remains: Can Blue Data be used not only as an effective critique of policing, but also in comparison with potential alternatives to policing? Can the sufficient study of Blue Data produce a tipping point with regards to public opinion and policy decisions concerning policing? Unfortunately, in order to propose alternative interventions to crises such as gun violence, domestic abuse, or acute mental health episodes, there needs to be sufficient empirical data for comparison which is not often available in Blue Data alone, or perhaps at all. In other words, Blue Data does not provide counterfactuals to policing per se.

5. Conclusion

Several algorithms and techniques have been implemented to model crime and police misconduct in recent years. In this paper, we explored police violence through the use of topological data analysis. Focusing on three key methodologies in TDA, we investigated police violence in Michigan as a case study, and found that police violence is concentrated in certain areas, or hotspots, similar to well-accepted characterizations of other types of crime. Moreover, while we acknowledge there exists some variability in the persistence scales for states with smaller data sets for each year, we offer the averaged persistence landscape as a way to determine which homological features persist the longest across the years, and the persistence of these features suggests the stability of these hotspots over time.

Finally, to reframe the preceding discussion in terms of the ethics, fairness, and accountability of policing systems, if we adopt the view, as many practitioners

do, of police violence as an ongoing and persistent public health concern, then there exists an ethical obligation to not only expand the availability and analysis of Blue Data, but moreover to formulate research questions aimed at the more difficult problem of counterfactual interventions as alternatives to policing. From a quantitative research standpoint, the analysis of Blue Data will be significantly improved by including data from other potential interventions, such as non-police crisis responses to emergency calls.

Acknowledgements. This research was partially supported by the UM-Dearborn Student-Faculty Mentored Research Grant and NSF Grant DMS 2212924. The authors thank Tom Fiore for guidance and Michelle Feng for helpful conversations.

References

[1] Mapping police violence: Data and methodology. https://mappingpoliceviolence.org/methodology. Accessed: 2023-01-27.

[2] Akpinar, N.-J., De-Arteaga, M., and Chouldechova, A. The effect of differential victim crime reporting on predictive policing systems. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021).

[3] Aktas, M., Akbas, E., and El Fatmaoui, A. Persistence homology of networks: methods and applications. *Applied Network Science 4* (08 2019).

[4] Aougab, Tarik, e. a. Boycott collaboration with police. Available at https://www.ams.org/journals/notices/202009/rnoti-p1293.pdf (2020/10).

[5] Bubenik, P. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res. 16*, 1 (jan 2015), 77–102.

[6] Bubenik, P., and Dłotko, P. A persistence landscapes toolbox for topological statistics. *Journal of Symbolic Computation 78* (2017), 91–114. Algorithms and Software for Computational Topology.

[7] Carlsson, G., and Vejdemo-Johansson, M. *Topological Data Analysis with Applications*. Cambridge University Press, 2021.

[8] Chapman, A., Grylls, P., Ugwudike, P., Gammack, D., and Ayling, J. A data-driven analysis of the interplay between criminological theory and predictive policing algorithms. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2022), FAccT '22, Association for Computing Machinery, p. 36–45.

[9] Collaborators, G. . P. V. U. S., et al. Fatal police violence by race and state in the usa, 1980–2019: a network meta-regression. *The Lancet 398*, 10307 (2021), 1239–1255.

[10] Damiano, D. B., and Wu, A. Topological data analysis in an economic context: Property tax maps. In *2021 IEEE International Conference on Big Data (Big Data)* (2021), pp. 3801–3811.

[11] Duchin, M., Needham, T., and Weighill, T. The (homological) persistence of gerrymandering. *Foundations of Data Science 4* (01 2021).

[12] Edelsbrunner, H., and Harer, J. *Computational Topology: An Introduction*. 01 2010.

[13] Feng, M., and Porter, M. A. Persistent homology of geospatial data: A case study with voting. *SIAM Review 63*, 1 (2021), 67–99.

[14] Ferguson, A. G. The exclusionary rule in the age of blue data. *Vanderbilt Law Review 72*, 2 (mar 2019), 561–646.

[15] Fugacci, U., Scaramuccia, S., Iuricich, F., and Floriani, L. D. Persistent homology: a step-by-step introduction for newcomers. In *Smart Tools and Applications in Graphics* (2016).

[16] Hickok, A., Needell, D., and Porter, M. A. Analysis of spatial and spatiotemporal anomalies using persistent homology: Case studies with covid-19 data, 2021.

[17] Kim, J. ripsfiltration: Rips filtration in tda: Statistical tools for topological data analysis, Oct 2022.

[18] Kim, J., and Rouvreau, V. alphacomplexfiltration: Alpha complex filtration in tda: Statistical tools for topological data analysis, Oct 2022.

[19] Lum, K., and Isaac, W. To predict and serve? *Significance 13*, 5 (2016), 14–19.

[20] Ouellet, M., Hashimi, S., Gravel, J., and Papachristos, A. V. Network exposure and excessive use of force. *Criminology & Public Policy 18*, 3 (2019), 675–704.
[21] Quispe-Torreblanca, E. G., and Stewart, N. Causal peer effects in police misconduct. *Nature Human Behaviour 3*, 8 (2019), 797–807.
[22] Simpson, C. R., and Kirk, D. S. Is police misconduct contagious? non-trivial null findings from dallas, texas. *Journal of Quantitative Criminology* (2022).
[23] Wood, G., Roithmayr, D., and Papachristos, A. V. The network structure of police misconduct. *Socius 5* (2019), 2378023119879798.
[24] Zhao, L., and Papachristos, A. V. Network position and police who shoot. *The ANNALS of the American Academy of Political and Social Science 687*, 1 (2020), 89–112.

Email address: dkinsman@umich.edu

Email address: tiananw@umich.edu

University of Michigan-Dearborn, 4901 Evergreen Road, Dearborn, Michigan, 48128