Lines: 327

Mapper graphs for voting analysis

Hazel N. Brenner
Cornell University, USA

Emarie De La Nuez
Tufts University, USA

Moon Duchin
Tufts University, USA

Jordan Phan
University of Virginia, USA

Abstract

Outputs from the Mapper algorithm are graph (or more generally cell-complex) representations of high-dimensional data, used to infer its "shape" and learn its structure. Here, we use the open-source package `KeplerMapper` to analyze voting patterns in Chicago mayoral elections. As part of this analysis, we create and refine techniques that enable dimension detection and feature selection with Mapper.

2012 ACM Subject Classification Applied computing → Law; Mathematics of computing → Discrete mathematics; Mathematics of computing → Topology

Keywords and phrases Mapper, topological data analysis, clustering, elections.

Funding *Moon Duchin*: partially supported by NSF DMS-2005512.

Acknowledgements The authors wish to thank the participants in the MGGG Redistricting Lab summer research cluster for their feedback and engagement throughout the development of this project. We thank the Democracy Fund for their generous support of the summer research group. Author order is alphabetical, following the convention in mathematical publication.

1 Introduction and background

Topological data analysis, or TDA, is a set of computational methods that provides a framework to simplify, visualize, and qualitatively describe high-dimensional datasets. Persistent homology is one dominant technique in the field, building simple structures called persistence diagrams to summarize features and their relationships. A second important method was initiated by Singh–Mémoli–Carlsson with their proposal of the Mapper algorithm [8]. Whereas linear regression implicitly assumes a linear "shape" is appropriate to describe your data, Mapper builds a cell complex—most often, a graph—where the nodes represent clusters in the data. In this project, we build a data analysis pipeline using the open-source Python package `KeplerMapper` to study Chicago's 2015 and 2019 mayoral elections.

1.1 Chicago mayoral elections

Chicago employs a two-round nonpartisan election system for its mayoral elections: a first round of voting is conducted in February, often with a dozen or more candidates competing, every four years. If no single candidate secures a majority of the vote, then the top two vote-getters compete in a runoff contest in April.

Chicago elections form an interesting dataset for analysis in a number of ways. First, Chicago mayoral contests historically provide a famous and extreme example of racialized voting behavior: in 1983, a majority of White Chicagoans, despite lifelong histories of supporting Democratic candidates, cast votes for a Republican in order to avoid supporting the first Black candidate to win a Democratic primary election in the city. That candidate, Harold Washington, narrowly won anyway, becoming the city's first Black mayor. Secondly, the stark racial segregation in residential housing allows us to visualize voting patterns detected below in the context of well-known neighborhoods. We will focus on the elections from 2015 and 2019; together with this year's (2023) election, these are the only three to have advanced to a runoff since that system was implemented in 1999. Two-round elections give us the added benefit of understanding shifts in patterns when fewer candidates are available, which gives a partial glimpse into voters' ranked preferences. Runoff elections are additionally clean because the candidates' support is complementary (summing to one). Finally, Chicago voting data is available in a clean and spatialized format.

1.2 Racially polarized voting

The Voting Rights Act of 1965 was intended by Congress to secure protection of minoritized groups from devices that "deny or abridge" the vote. A state or locality can be challenged when it enacts a system of election (including a districting plan) that can be shown to be responsible for reducing access to effective representation. In 1986, in a major case called *Thornburg v. Gingles*, the court adopted a three-pronged test that had been proposed in an academic paper to serve as a sort of checklist of "preconditions" before advancing to a lawsuit. These are now known as the three *Gingles factors* in voting rights law: plaintiffs must show that (1) it is possible to draw an additional majority-minority district while conforming to traditional principles; (2) the minority votes mostly as a bloc for the same candidates of choice; (3) the majority votes in a manner that prevents those chosen candidates from election. Together, Gingles 2-3 operationalize a notion of *racially polarized voting* (RPV). For instance, if a lawsuit is filed on behalf of Black voters, the RPV burden on the plaintiffs is first to show that Black voters vote cohesively for common candidates in recent elections, then to show that the (typically mostly White) majority is also cohesive, but supporting different candidates, with the effect that Black-preferred candidates are not prevailing.

Because American elections are conducted by secret ballot, the experts in voting rights cases will always lack direct evidence of voting by race. Instead, voting patterns are inferred from precinct-level results. In other words, voting records that are given to us with one aggregation—by precinct, which is typically a territorial area with a few thousand residents—are subjected to statistical inference to reaggregate them in a different way, namely by race/ethnicity. This kind of inferential regrouping is a long-studied problem sometimes called the *ecological inference problem*, which can lead to errors known as the *ecological paradox* or *Simpson's paradox*. There is no foolproof way to get around the missing data problem.

The standard method going back several decades is the first one we might expect: testing whether the racial balance of the precincts is correlated with the voting with a simple linear regression. Two *ecological regression* plots (as they are called) are shown in Figure 1.

Several facts about Chicago are visible from these plots directly. For instance, there are precincts with very high levels of Black population and others with very low levels, but relatively fewer that are between 15 and 85% BVAP. By contrast, the HVAP has much more even distribution across precincts over the levels from 15% to 100%. This is true despite the fact that the citywide share of BVAP in this dataset is not too far off from the HVAP share, at 33.5% and 27.3%, respectively. We also see that neither fit line is sharply sloped, but the slight upward trend in the BVAP plot indicates that Black voters were slightly more likely to support Preckwinkle than others, while Hispanic voters were estimated to prefer Lightfoot to Preckwinkle 75-25, just like non-Hispanic voters. A very clear preference is only evident in one of the four cases: Hispanic/Latino voters had a pronounced tendency to support Chuy

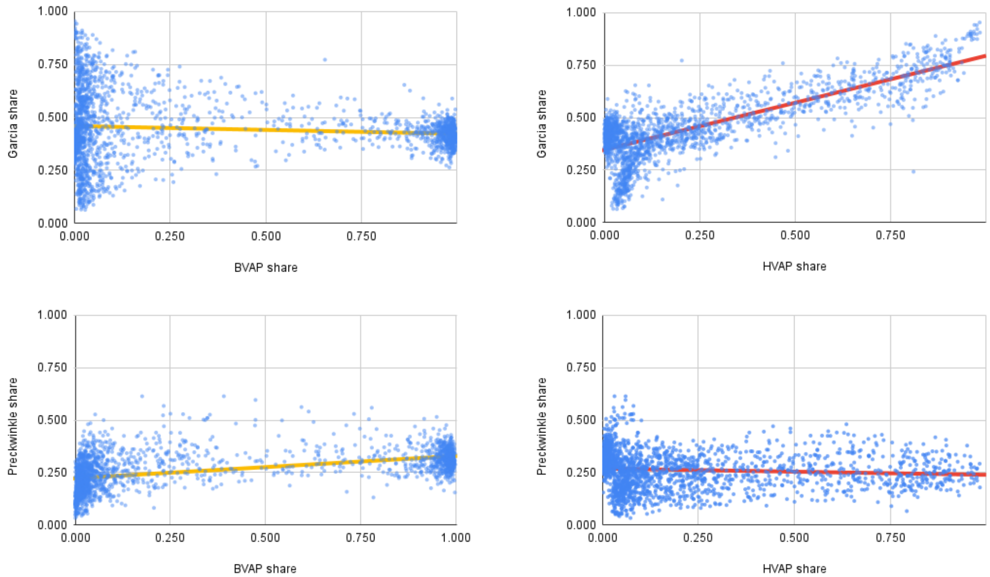


Figure 1 Four scatterplots whose points represent the 2069 precincts our Chicago dataset. The Black or Hispanic share of the voting age population (BVAP and HVAP, respectively) are plotted against the share of the vote that went to the runner-up in the 2015 and 2019 mayoral runoff elections. The slopes of the fit lines might indicate that Black voters were slightly more likely to support Preckwinkle than others

Garcia, who did not achieve a majority among other Chicago voters and indeed lost overall.[1]

What is not produced by this kind of analysis is any kind of assessment of how race/ethnicity variables interact with social, economic, geographic, and other factors to explain patterns of voting. And indeed ecological regression plots are most often conducted for one racial group at a time—while it is possible to conduct non-linear regressions on higher-dimensional data, it is not clear that regression analysis is the right tool for the job. In this paper we explore the use of Mapper as a tool for viewing racial polarization in context of a much richer picture of human geography.

1.3 TDA and Mapper

Mapper graphs are discrete objects that carry topological information thought to be helpfully descriptive of high-dimensional data. They are analogs of tools originally developed in Morse theory for the study of manifolds. Given a manifold and a function thought of as "height," a construction called the Reeb graph records the information of how the constant-height slices (i.e., level sets) change as the height varies over its range. The nodes of the graph correspond to birth, death, splitting and merging events for connected components of the level sets. The edges of the graph are drawn between nodes representing a given connected component and nodes representing events involving that component. The Mapper algorithm mimics the Reeb graph in the setting that the object of analysis is a point cloud. It proceeds

[1] Typically, it is only the slope and the intercepts with the $x = 0$ and $x = 1$ lines that are used in court. The intercepts are interpreted as predicted levels of support for the candidate by non-members and members of the indicated group, respectively. The closeness of fit is seldom mentioned.

with the following steps.

1. **Filter function.** Apply a continuous $f : \mathbb{R}^n \to \mathbb{R}$, called a *filter function*, to the dataset in $\mathbb{R}^n$.
2. **Interval cover.** Cover the image of the data in $\mathbb{R}$ by overlapping intervals $I_1, \ldots, I_n$. Commonly, this is done with a fixed number of intervals overlapping on a fixed fraction of their length
3. **Pullback cover.** The dataset is covered by the pre-images $f^{-1}(I_i)$ in $\mathbb{R}^n$. The collection of these is called the *pullback cover*.
4. **Clustering algorithm.** For each i, apply a chosen clustering algorithm to the pullback set $f^-1(I_i)$. This yields a family of subsets $\{C_{i,1}, \ldots, C_{i,k_i}\}$ corresponding to the k_i clusters in I_i.
5. **Mapper complex.** Construct a simplicial complex with a 0-simplex (nodes) for each cluster $C_{i,j}$, a 1-simplex (edge) between each pair of clusters that shares at least one common point, a 2-simplex (face) where a triple of clusters shares a member data point, etc. This is equivalent to the *nerve* of the cover $\{C_{i,j}\}$. The Mapper graph is the 1-skeleton of this complex.

Mapper is widely studied as a TDA tool because of its flexibility and interpretability. In his survey paper *Topology and Data* [1], Carlsson lists three main advantages of using Mapper: insensitivity to the choice of metric on the feature space, transparent reliance on parameters, and adaptability to multiple scales of resolution. In addition, graphs are a relatively simple and interpretable format for output. This is highly desirable for exploratory analysis of complex real-world datasets. The task of analysis becomes that of lining up several views of the dataset through various choices of Mapper parameters and then synthesizing a coherent narrative across them.

As the description of steps makes clear, there are numerous of choices that must be made for a given run. In our raw data, the data points correspond to precincts, embedded in a high-dimensional space of socio-demographic features. We have chosen the share of support for a particular candidate to serve as the filter function for most of the analysis presented here. This choice enables us to readily visualize patterns in voting behavior. There is no uniform choice for interval cover in Mapper applications, though there are selection regimes described by Carriere et al. [2]. In our application, we primarily used a combination of hand-tuning and the *adaptive cover* algorithm from Wang et al. [3]. Finally, the clustering algorithm is a crucial choice that often does not receive enough discussion in Mapper-related work. We used a combination of `DBSCAN`, `X-Means` and centroid-linkage agglomerative hierarchical clustering (AHC) in order to get strategically different views of the data, and will focus on `DBSCAN` and `X-Means` here.

The mapper pipeline here runs code built on open-source Python packages: `KeplerMapper`, the Optimal Transport (OT) Library, and Scikit-Learn [9, 5, 6]. We pulled social, economic, and demographic information from two data products of the U.S. Census Bureau: Decennial Census data via NHGIS and annual releases of the American Community Survey. Electoral data is sourced to the City of Chicago.

2 Clustering for shape detection

`KeplerMapper` is commonly used with Scikit-Learn clustering algorithms, defaulting to `DBSCAN`, which identifies core areas in the data by looking for well-separated regions of high point density. We make heavy use of a second alternative for clustering via `X-Means`, which runs k-means with a refinement step that can result in significant sub-cluster splitting.

Between DBSCAN and X-Means, each has advantages and disadvantages for gaining insight into data in our application, and we will use this section to begin to build up to interpreting summarized outputs such as we will create below in Figure 8. We include a test run showing how DBSCAN and X-Means Mapper graphs perform when trying to distinguishing data made by noising simple manifolds. We find that DBSCAN gives superior ability to distinguish the underlying manifold, but it does so without reliably learning the dimension of the manifold. In addition it discards data points in areas of data scarcity, which is helpful for a summary but undesirable for a finer-grained analysis of patterns. By contrast, X-Means gives a very satisfying picture of dimension but has less discernment of other topological features.

DBSCAN is run by specifying parameters ε and minPts. It identifies a *core point* as one whose ε neighborhood contains at least minPts other points; the algorithm works by processing the data into core points and their near-neighbors, and discarding the rest. A network of near-neighboring core points will be collapsed to a single cluster. This means that for well-chosen parameters, a DBSCAN mapper graph for uniform random noise in a unit d-dimensional cube will be essentially the nerve of the cubical cover—i.e., the Mapper graph is simply a path in the case of an $\mathbb{R}$-valued filter function, no matter the "true" dimension d of the point cloud.

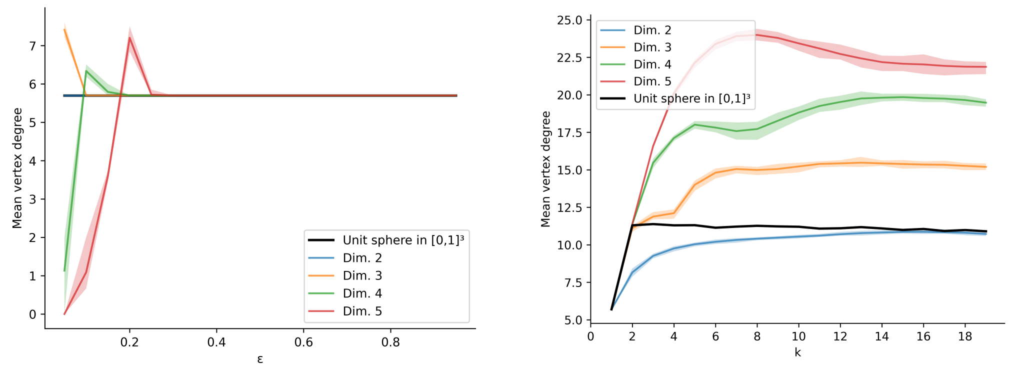
The X-Means algorithm is built as a a variant on classic k-means clustering, which proceeds by initializing some k sites in the feature space and assigning data points to the closest site to form k clusters; then updating sites as centroids for the newly formed clusters; then repeating the process until the sites have sufficiently converged. But the choice of how many clusters to use is made in advance. Qualitatively, choosing k too high tends to create a large number of extraneous nodes and edges, while too low a value can fails to capture the complexity of the data. X-Means avoids these artifacts by tuning k over each element of the pullback cover. This algorithm starts as in k-means but then considers possible refinements, such as by splitting a cluster and assessing whether the Akaike information criterion (AIC)/Bayesian information criterion (BIC) improves in comparison to the parent cluster. X-Means attempts to optimize the number of clusters in this way. In areas with dense points and detailed structure, X-Means will therefore do a better job of describing the detail, where DBSCAN would return a larger cluster.

2.1 Dimension

Figure 2 follows the idea proposed by Dłotko for use with the BallMapper algorithm [4], showing how mean vertex degree relates to the dimension of the manifold from which points were sampled. Both plots use uniform random noise in $[0,1]^d$, as well as a points sampled uniformly from a round sphere in $[0,1]^3$, as test data. The datasets are built from 10,000 points distributed uniformly at random from $[0,1]^d$ for $d = 2,3,4,5$. We run DBSCAN on these datasets 50 times at various levels of ε. We find that for each dimension, the vertex degrees collapse to a constant level when ε gets large enough; for lower values of ε, experimentation showed high volatility and no clear signal.

We then compare to runs with k-means for values of k between 1 and 20 and find, as we would expect, that higher-dimensional data produces higher-degree vertices—each cluster has more neighbors. In the plots, the baseline curves represent the mean of 50 trials and the shaded regions reflect the range for those trials.

We propose to use the k-means plot to infer that X-Means will perform well on dimension detection, because X-Means corresponds to k-means with "good" local choices for k—i.e., high enough k where needed.



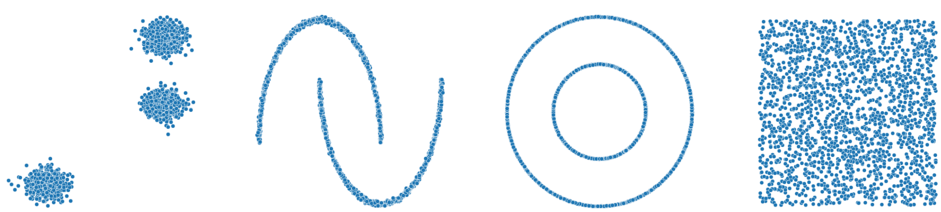
DBSCAN $\qquad$ k-Means

Figure 2 In the example plot generated with this methodology, we estimate that the experimental data has dimension strictly less than 3. This is consistent with our expectation for the intrinsic dimensionality of data on the surface of a 2-dimensional sphere.

2.2 Distinguishing shapes

In this section, we pull ideas from Mémoli [7] and Singh et al. [8] to deepen the comparison of `DBSCAN` and `X-Means`. We consider how each algorithm performs at distinguishing noised instances of toy datasets, using dissimilarity matrices as a visualization device.

Here, each individual Mapper graph is metrized via the path metric, where the length of an edge is taken to be the difference between the average value of the filter function at the endpoint nodes.



Spots $\qquad$ Arcs $\qquad$ Circles $\qquad$ Square

Figure 3 We will make three noised images from each of four manifolds: three spots (small disks); a pair of arcs; a pair of circles; and a square section of a plane.

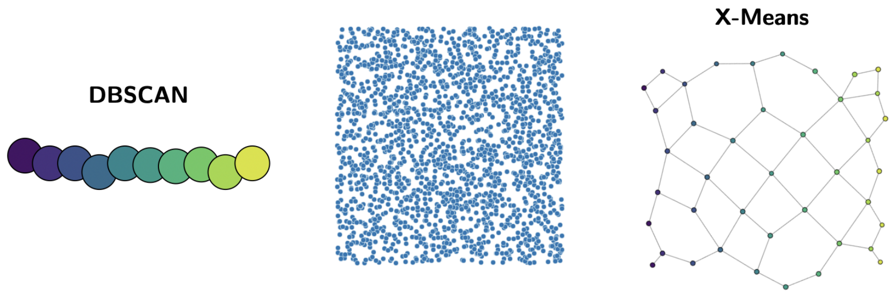


Figure 4 In particular, `DBSCAN` and `X-Means` handle the noisy planar square very differently, with `X-Means` succeeding beautifully at rendering a lattice-like Mapper graph, while `DBSCAN` simply returns the nerve of the interval cover.

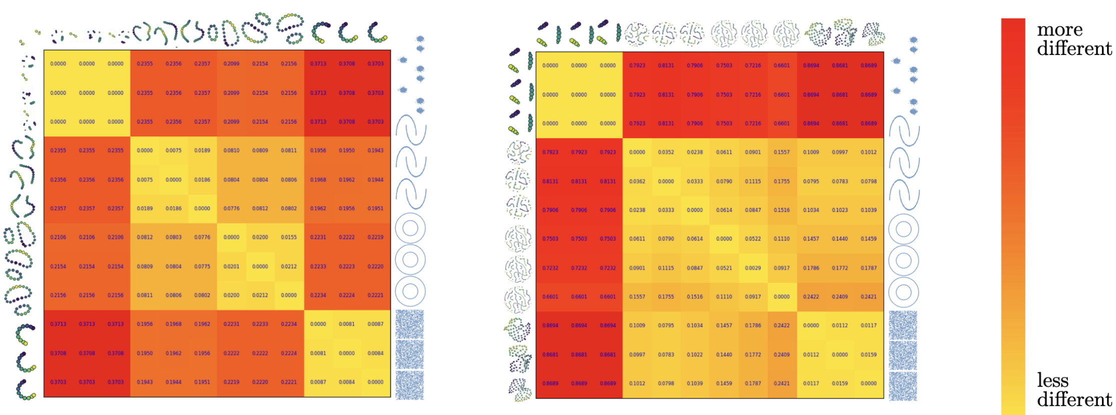


Figure 5 Dissimilarity matrices for Mapper graphs produced by `DBSCAN` (left) and `X-Means` (right) when four simple datasets are noised three times each. `DBSCAN` is much better at clearly distinguishing whether shapes come from the same source data. However, `X-Means` still passes the test that each noisy set is classified with its own cohort.

2.3 Feature importance

Our full dataset has 77 categorical variables (listed in Supplemental Table 1), which we classify into six buckets.

1. **Personal.** Gender, veteran status, marriage status, and insurance.
2. **Education.** Highest level of education achieved.
3. **Household.** Type of housing units, occupied housing units, and household characteristics, including language spoken at home.
4. **Income.** Brackets come in intervals of 10K from 0 to 200K+.
5. **Work.** Modes of commute, commuting time, occupation type (service, office, natural resources, transportation, and law enforcement), and rate of employment.
6. **Race/Ethnicity.** Voting age population (VAP) share that is White, Black, Hispanic, and Asian.

Figure 6 Pairwise Wasserstein distances between `DBSCAN` Mapper graphs based on the full dataset and the six alternatives made by holding back one bucket of variables at a time. This view shows that holding back race/ethnicity variables makes a bigger change to the Mapper graph in the 2015 election than in 2019.

To measure the explanatory power of categorical variables in voter behavior, we create an initial Mapper graph with the full set of variables, and create six ancillary graphs in which each one of the six buckets of variables, in turn, is held out.

The general picture is one that suggests major interrelations and correlations among the types of variables. This outcome is compatible, for instance, with the observation that the 2015 runoff was highly racially polarized.

3 Case study: 2019 runoff election

3.1 Overview

Chicago's 2019 mayoral election culminated in a runoff between the top two candidates: relative political outsider Lori Lightfoot against city council stalwart Toni Preckwinkle, both Black women. The first-round vote had been divided many ways, with 560,701 votes cast of 1,581,755 registered voters. Lightfoot won the first-round plurality with only 17.54% of the vote, followed by Preckwinkle's 16.06%, and the remaining two-thirds of votes divided many ways among the other candidates. The runoff had only slightly lower turnout of 526,886, and the outcome was not close: Lightfoot won in a landslide, ending with 73.7% of the final vote and a majority in every one of Chicago's 50 wards. Preckwinkle's 23.6% runoff performance put her well behind the runner-up in the 2015 race, Jesùs "Chuy" Garcia, who clocked about 33.6% against Rahm Emanuel. In Figure 7 we can see that Preckwinkle's support was also remarkably constant over the precincts of Chicago, in sharp contrast to the racial variation observable at the precinct level.

We set up two kinds of mapper runs to study the 2019 runoff: one mapper graph using `DBSCAN` clustering and one with `X-Means`. Both versions use an adaptive cover after the nodes are filtered by the share of support for Preckwinkle. In all colorations, yellow represents high levels of the variable, while dark purple represents low levels. The data in this run did not include race, ethnicity, or geographic variables because we are interested in seeing whether clustering on the other variables will recover those as distinctive structures. That is, it would not be informative to find that high-BVAP precincts are clustered together in a Mapper graph when BVAP is one of the variables: in this case, the metric on feature space will necessarily consider racially similar precincts to be closer than racially dissimilar ones, so will be more likely to keep them together in a cluster.

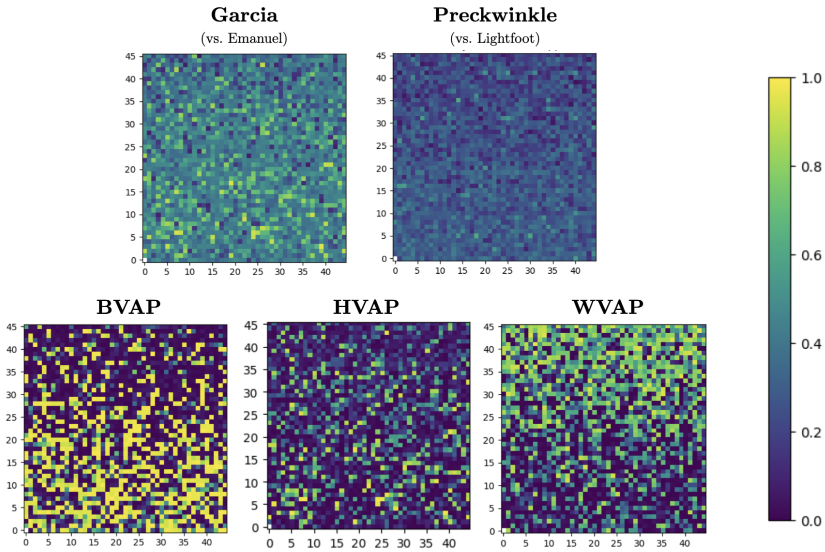


Figure 7 Nodes in the mapper graphs below fuse collections of these 2069 precincts, arrayed here in an arbitrary lexicographic order to give a sense of the level of variation.

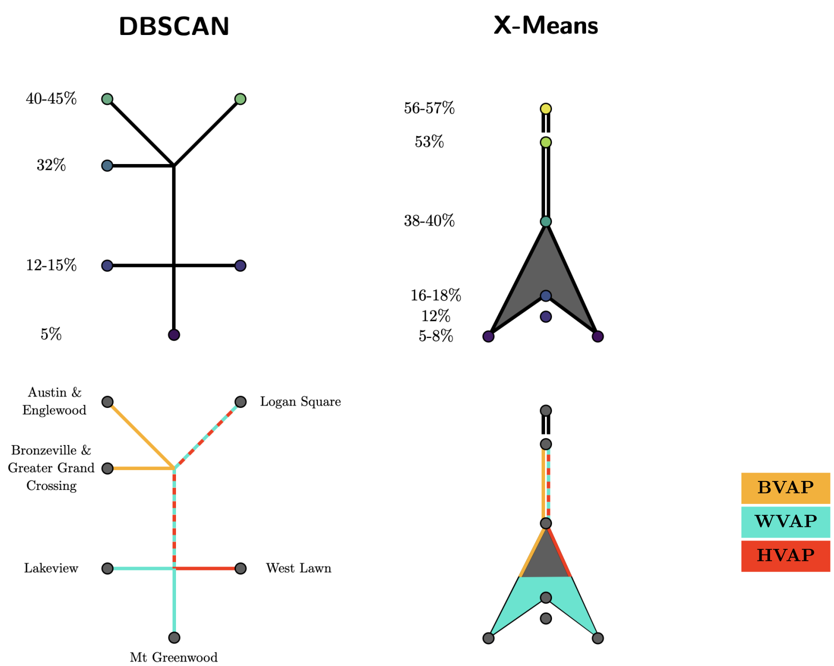


Figure 8 High-level summary of the information contained in the topology of the Mapper graphs on the 2019 runoff election, filtered by Preckwinkle support (Figs 9-11). Race trends appear in the outputs even though the election was not highly racially polarized and race variables were not included in the point cloud embedding. The endpoints of the `DBSCAN` arms pick out geographically recognizable neighborhoods, even though geography variables were also excluded. The `X-Means` plot tells us that there is one main progression of clusters tending to higher Preckwinkle support, while Lightfoot support is fundamentally two-dimensional.

3.2 Enumerating themes with dbscan

Figure 9 shows a typical `DBSCAN` mapper graph when the runoff election is used with a large and varied set of variables. The same graph is colored in turn by Preckwinkle support, WVAP share, HVAP share, and BVAP share. From this graph, we can identify six key directions or themes among Chicago precincts: two trending to (relatively) high Preckwinkle support, one trending to Lightfoot, and three spurs at constant support levels.

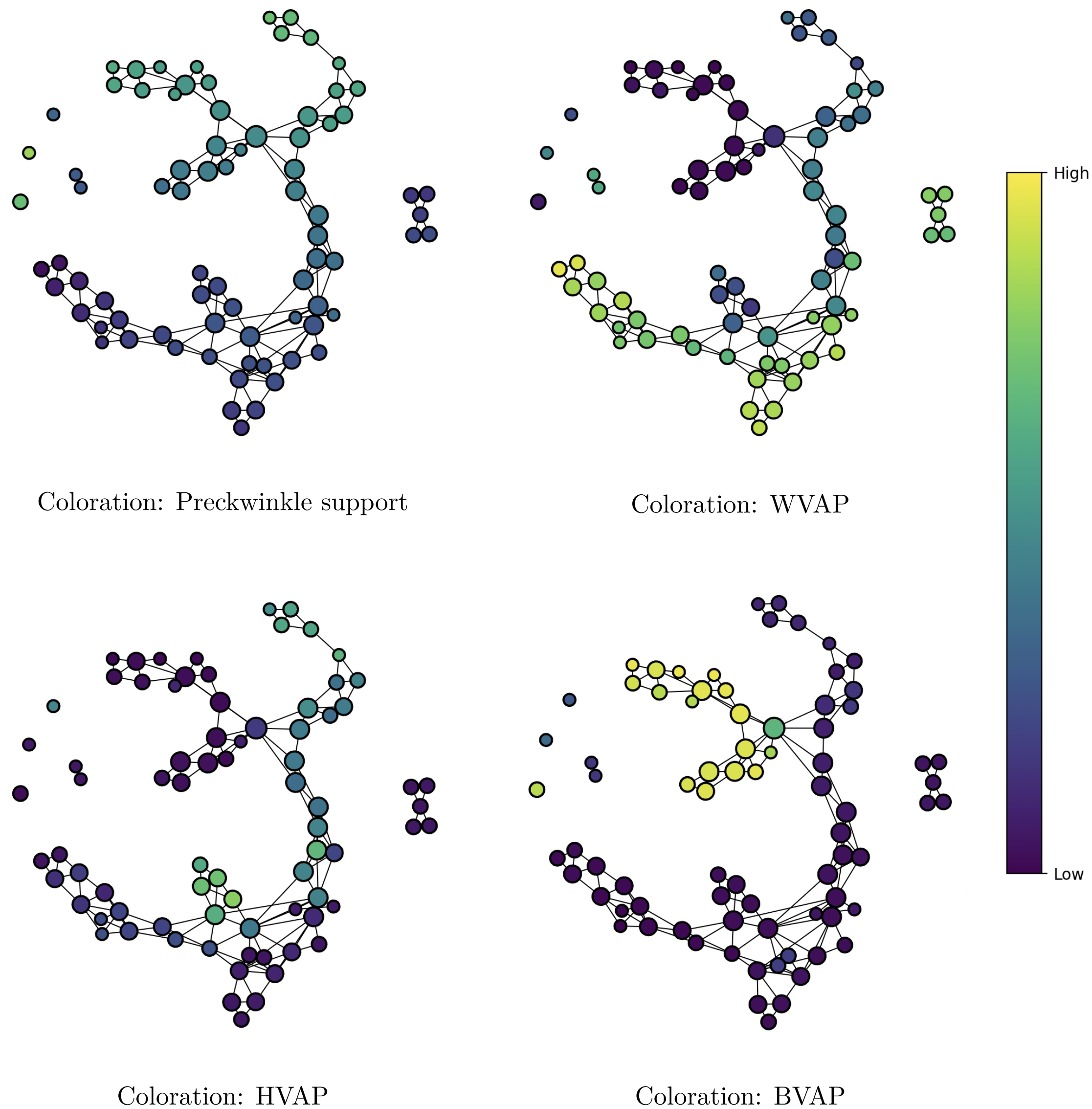


Figure 9 `DBSCAN` Mapper graph on 2019 runoff data. In all cases, the filter function is the share of Preckwinkle support in the precinct; the coloring function varies as indicated.

The high-BVAP path in Figure 9 encompasses much of Chicago's South Side, as well as neighborhoods on the West Side around the Garfield Park region. Both these areas are historically majority-Black and the extreme concentration of Black Chicagoans in these regions has gone far to earn Chicago's reputation for high segregation.

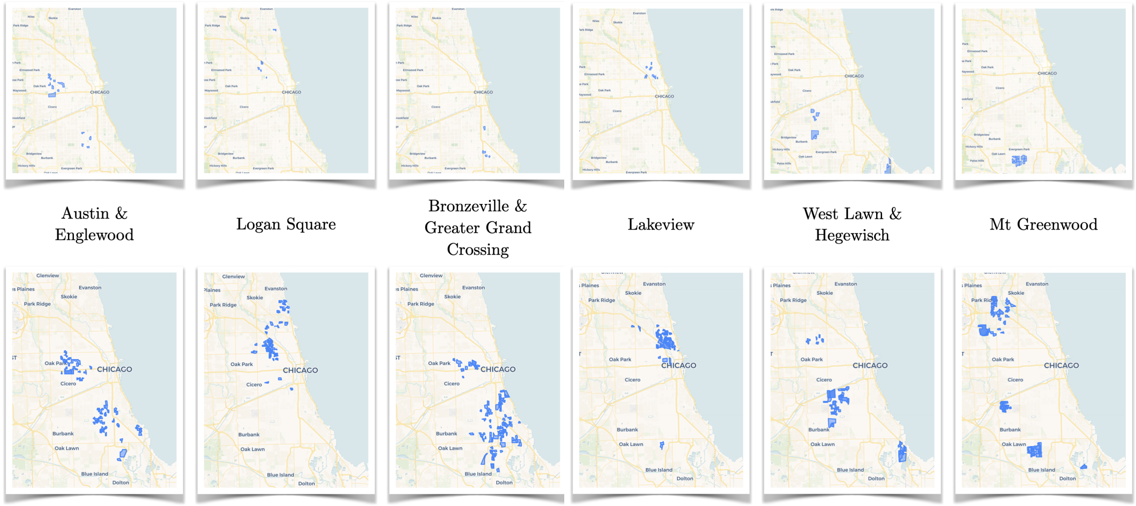


Figure 10 Geolocating the precincts that constitute the extreme nodes in the six major directions picked out by the `DBSCAN` Mapper graph. Top row: the single most extreme nodes, representing 5-16 precincts each. Bottom row: several nodes collected from along each branch, until roughly 100 precincts are included.

Besides the six dominant directions, it can also be interesting to look at nodes that are high-degree and centrally located. Coloring by BVAP in Figure 9 highlights only a single node as having an intermediate level of BVAP—it is a large node constituting 186 precincts. Its degree is nine, and it serves as a hub connecting the high-BVAP and high-HVAP spines. This node represents areas throughout the South and East Sides of Chicago, not concentrated in any particular neighborhood but scattered throughout. While the high degree of this node indicates that many of these precincts also belong to surrounding nodes, these 186 precincts have enough common webbing in their social, economic, and demographic features to be clustered together by `DBSCAN`. This node does not have a one-to-one counterpart in the `X-Means` version of this graph, but splits into several nodes along the area next to both the HVAP and BVAP regions. For community organizers, this might merit study as an emergent area with common issues and needs.

3.3 Learning dimensionality with xmeans

Above, we argued that `X-Means` clustering is especially well suited to detecting dimensionality of data. Accordingly, most notable feature of the `X-Means` graphs is the difference in form between the tapered end at the high-Preckwinkle side to the lattice-like triangle of high Lightfoot support. As we see from the WVAP coloration panel of Figure 11, the whole flaring end is characterized by being composed of mainly high-WVAP precincts.

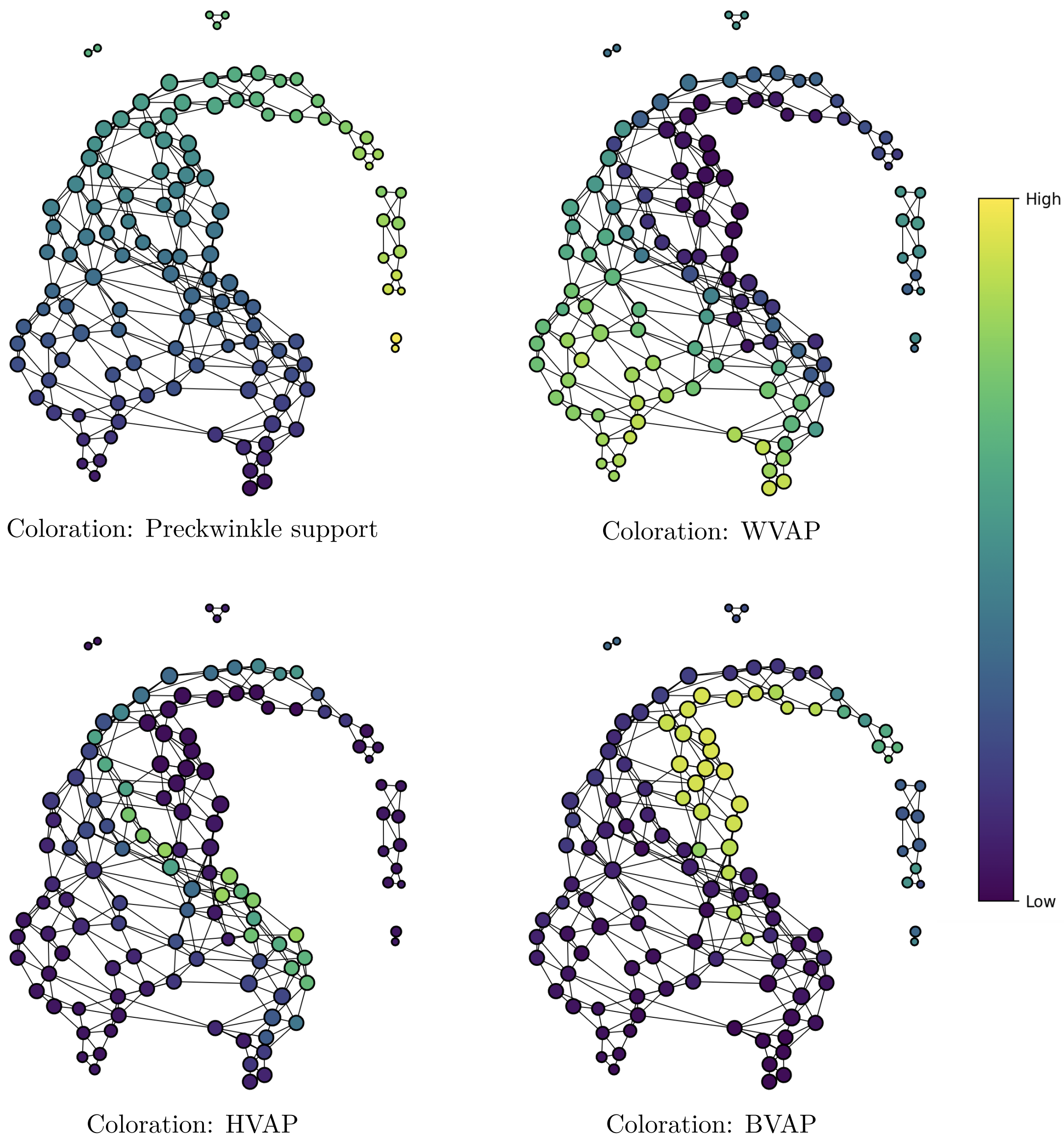


Figure 11 `X-Means` Mapper graph on 2019 runoff data. In all cases, the filter function is the share of Preckwinkle support in the precinct; the coloring function varies as indicated.

The two extreme tips of the high-Lightfoot triangle represent distinct subsets of White-majority neighborhoods in the Chicago landscape. One side corresponds to Lakeview, Lincoln Park, the Near North Side, and the Loop, which are generally affluent. The other side picks up several lower-income neighborhoods that turn out to be characterized by sharply higher-than-average police residency, including the three main police neighborhoods of Norwood Park, Garfield Ridge, and Mount Greenwood.

These two directions—one more wealthy and one less—are pulled together by one "hub" node that is well connected to much of the high-Lightfoot edge of the graph. This hub maps onto part of the area around Norridge, and together with its neighbors covers parts of the Near North Side, Garfield Ridge, and Southeast Chicago near the Southeast police neighborhoods. From manual inspection of the data, these precincts in the hub node seem to have been clustered together because of the income variable and its lower police population compared to the rest of the high-police clump. Generally, the popularity of Lightfoot in the police neighborhoods emerges strongly from the analysis here.

4 Conclusion

In geography, a *choropleth* is a geographical map where the units have been divided up and shaded according to the levels of some variable. Mapper graphs function as a kind of reverse-choropleth; rather than using a geographic map to reference voter preferences, our Mapper graphs are a representation of voter preferences that can be used to identify trends in geography, but also in race and socio-economic variables. A finding of geographically coherent areas in the Mapper graph conveys information about significant communities, bonded by some set of shared features, without having to guess in advance whether geography, economics, or race/ethnicity variables will be the most important and explanatory.

This paper offers a solid, if preliminary, theoretical and example-driven pitch for TDA-enabled analysis of how voting behavior draws out patterns in the human geography of one major American city. We hope this mainly serves as grounding and provocation, and that other authors will take up this research direction to develop this promising tool.

References

1 Gunnar Carlsson. Topology and data. *Bull. Amer. Math. Soc. (N.S.)*, 46(2):255–308, 2009. doi:10.1090/S0273-0979-09-01249-X.

2 Mathieu Carrière, Bertrand Michel, and Steve Oudot. Statistical analysis and parameter selection for mapper. 19(12):1–39. URL: http://jmlr.org/papers/v19/17-291.html.

3 Nithin Chalapathi, Youjia Zhou, and Bei Wang. Adaptive covers for mapper graphs using information criteria. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3789–3800. doi:10.1109/BigData52589.2021.9671324.

4 Paweł Dłotko. Ball mapper: a shape summary for topological data analysis. URL: http://arxiv.org/abs/1901.07410, arXiv:1901.07410[math], doi:10.48550/arXiv.1901.07410.

5 Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. 22(1):3571–3578. URL: http://jmlr.org/papers/v22/20-451.html.

6 Oliver Kramer. Scikit-learn. In Oliver Kramer, editor, *Machine Learning for Evolution Strategies*, Studies in Big Data, pages 45–53. Springer International Publishing. doi:10.1007/978-3-319-33383-0_5.

7 Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. 11(4):417–487. URL: http://link.springer.com/10.1007/s10208-011-9093-5, doi:10.1007/s10208-011-9093-5.

8 Gurjeet Singh, Facundo Memoli, and Gunnar Carlsson. *Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition*. The Eurographics Association.

Accepted: 2014-01-29T16:52:11Z ISSN: 1811-7813. URL: https://diglib.eg.org:443/xmlui/handle/10.2312/SPBG.SPBG07.091-100, doi:10.2312/SPBG/SPBG07/091-100.

9 Hendrik Jacob van Veen, Nathaniel Saul, David Eargle, and Sam W. Mangham. Kepler mapper: A flexible python implementation of the mapper algorithm. 4(42):1315. URL: https://joss.theoj.org/papers/10.21105/joss.01315, doi:10.21105/joss.01315.

A Historical information on Chicago elections

In the 1980s, Chicago earned an ugly reputation for racially rigid voting when White voters with lifetime Democratic voting history crossed over in large numbers to avoid voting for Harold Washington, the first Black man to earn the Democratic nomination for mayor. That reputation has continued into the present day, and ties in to the stark racial segregation prevalent in the city.

Today, Chicago has become a deeply multiracial. As of the 2020 census, Chicago had a total population of 2,746,388, and population trends had brought White, Black, and Latino residents into near parity. Nearly a third of the population, 33.1%, identified as non-Hispanic White alone on the Census; 29.2% as non-Hispanic Black alone; and 28.7% as Hispanic or Latino. Additionally, 6.8% identified as Asian, and 7.4% as belonging to two or more racial groups.

In the 2015 mayoral election, the first round was held between eight most significant candidates: incumbent Rahm Emanuel, Jesús G. "Chuy" Garcia, Robert Fioretti, William "Dock" Walls, Willie Wilson, William H. Calloway, Christopher Ware, and Mary Vann. In the first round, no candidate received a 50% majority: Emanuel received 46% of the vote and Garcia received 34%, with the others trailing Garcia by 20 or more percentage points. Therefore, a runoff was held between Emanuel and Garcia. Despite both candidates' membership in the Democratic party, Emanuel and Garcia's politics displayed more differences than similarities. A staunch centrist, Emanuel began his career representing Illinois in the House of Representatives from 2003 to 2009. He then served as White House Chief of Staff in the Obama Administration from 2009 to 2010. Similarly, Garcia began his career in local politics, first as a member of Chicago's city council in 1986, later becoming the first Mexican-American member of the Illinois State Senate in 1992. Garcia continued pursuing a career in local Chicago politics in 2010 as county commissioner on the Cook County Board.

The Garcia campaign launched initiatives to increase turnout and support among Hispanic voters in the runoff, as well as endorsements by Black community leaders. Garcia also earned endorsements from Congressman Danny Davis, Jesse Jackson's Rainbow/PUSH Coalition, Willie Wilson, and labor unions. Ultimately, Emanuel won the election with 56% of the runoff vote.

The 2019 mayoral election saw 14 candidates running for the seat, with front-runners Lori Lightfoot and Toni Preckwinkle. Preckwinkle was a firmly established Chicago politician, having served on the Chicago City Council from 1991 to 2010, where she was seen as an ally to the club of politicians that were holdovers from the Machine Era of Chicago politics. Preckwinkle then transitioned to the Cook County Board of Commissioners, where she has served as President ever since. In contrast, Lightfoot was considered a political outsider, as her candidacy for mayor was the first time she had made a bid for public office. She was the first openly lesbian candidate for the mayor seat, and had previously served as appointee of the Emanuel administration to the Chicago Police Accountability Task force and the Chicago Police Board. Her years of work with members of the police force would later lead progressive groups in Chicago to complain vocally that she was too close to the police, and ill-suited to hold their feet to the fire.

B 2015 Mayoral contest

2015's mayoral contest was a contentious race that was publicly viewed through a racialized lens. Figure 12 shows a run of Mapper graphs on the first round using the `X-Means` clusterer with adaptive cover, filtered by Garcia support.

B.1 First-round

A common perception of the 2015 mayoral election is that Hispanic voters voted in a solid bloc for Chuy Garcia, but one interesting observation in Figure 12 is that the three major racial groups are represented with pathlike structures that reach from high to low levels of Garcia support.

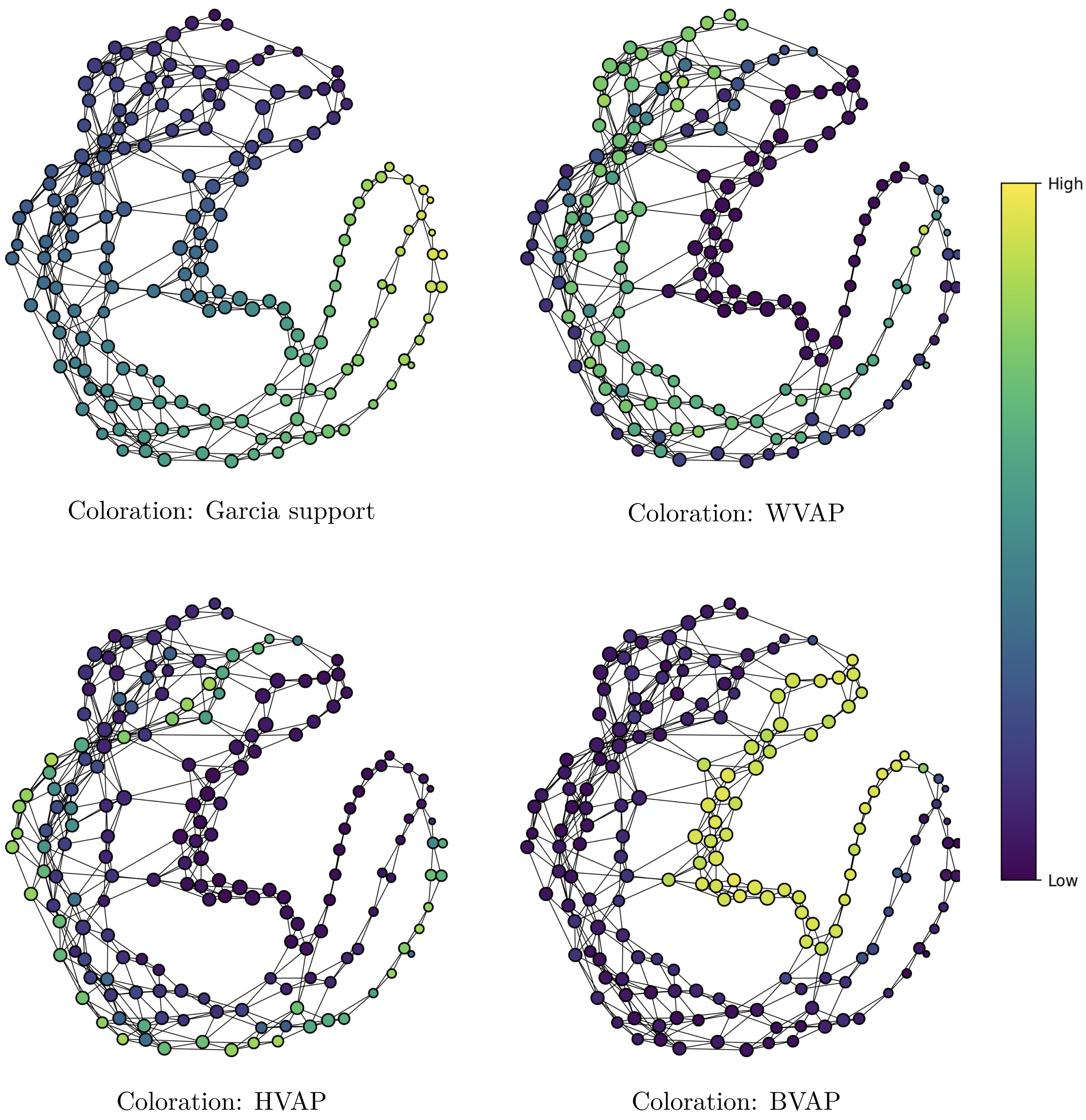


Figure 12 `DBSCAN` Mapper graph on 2015 first-round data. In all cases, the filter function is the share of Garcia support in the precinct; the coloring function varies as indicated.

These structures indicate that there were Hispanic-majority precincts falling at all levels of support for Garcia, and highlights more variety in Latino voting behavior than is suggested by the regression plots.

B.2 Runoff

The election wen to a runoff. In the resulting graphs, support for all candidates sums to one. This graph uses the same variables and procedures as the first-round graph, and was made with `X-Means`.

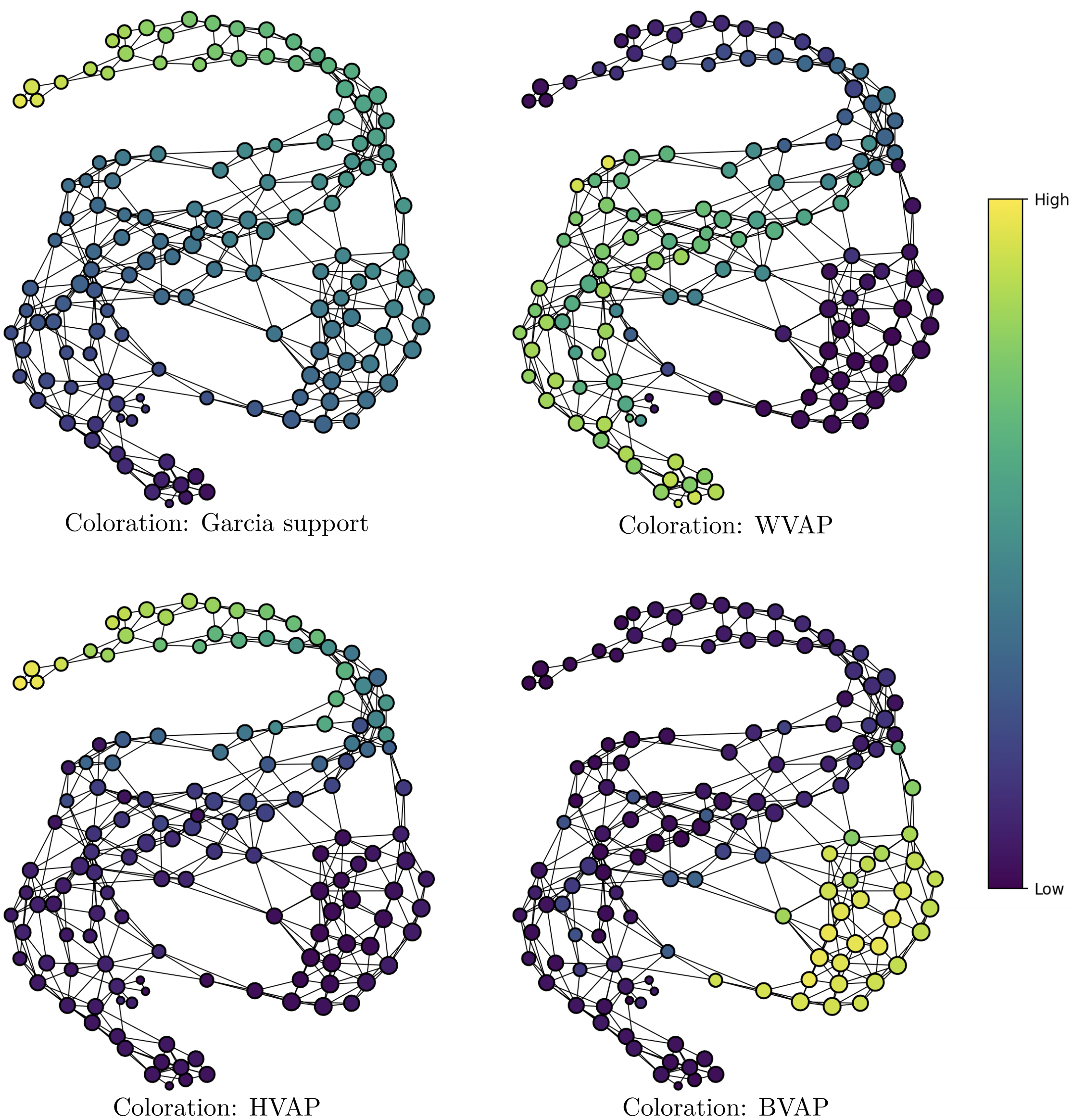


Figure 13 `DBSCAN` Mapper graph on 2015 runoff data. In all cases, the filter function is the share of Garcia support in the precinct; the coloring function varies as indicated.

The three race and ethnicity colorations of Figure 13 reveal three separate segments of the graph with varying levels of support for Garcia. This time, as opposed to our primary study of the 2019 runoff, the signs of racial polarization are unmistakable. One candidate's pole of support is heavily White while the other's is heavily Hispanic/Latino. And heavily Black precincts are very clearly off to the side, ranging over middling levels of support.

C Data dictionary

Variable	Definition
full_text	Ward and precinct IDs
precinct	Precinct ID
ward	Ward ID
ward_prec	Alternate code for Ward and precinct IDs
shape_area	Precinct area in square feet
shape_len	Precinct perimeter in feet
TOTPOP	Total population from 2010 census
NH_WHITE	Non-hispanic White population from 2010 census
NH_BLACK	Non-hispanic Black population from 2010 census
NH_AMIN	Non-hispanic American Indian and Alaska Native population from 2010 census
NH_ASIAN	Non-hispanic Asian population from 2010 census
NH_NHPI	Non-hispanic Native Hawaiian and Pacific Islander population from 2010 census
NH_OTHER	Non-hispanic population of other race from 2010 census
NH_2MORE	Non-hispanic population of two or more races from 2010 census
HISP	Hispanic population from 2010 census
H_WHITE	Hispanic White population from 2010 census
H_BLACK	Hispanic Black population from 2010 census
H_AMIN	Hispanic American Indian and Alaska Native population from 2010 census
H_ASIAN	Hispanic Asian population from 2010 census
H_NHPI	Hispanic Native Hawaiian and Pacific Islander population from 2010 census
H_OTHER	Hispanic population of other race from 2010 census
H_2MORE	Hispanic population of two or more races from 2010 census
VAP	Voting age population from 2010 census
HVAP	Hispanic voting age population from 2010 census
WVAP	White voting age population from 2010 census
BVAP	Black voting age population from 2010 census
AMINVAP	American Indian and Alaska Native voting age population from 2010 census
ASIANVAP	Asian voting age population from 2010 census
NHPIVAP	Native Hawaiian and Pacific Islander voting age population from 2010 census
OTHERVAP	Voting age population of other race from 2010 census
2MOREVAP	Voting age population of two or more races from 2010 census
TOTHH	Total number of households from 2013-2017 ACS
LESS_10K	Number of households w/income under $10,000 from 2013-2017 ACS
10K_15K	Num households w/income between $10,000 and $14,999 from 2013-2017 ACS
15K_20K	Num households w/income between $15,000 and $19,999 from 2013-2017 ACS
20K_25K	Num households w/income between $20,000 and $24,999 from 2013-2017 ACS
25K_30K	Num households w/income between $25,000 and $29,999 from 2013-2017 ACS
30K_35K	Num households w/income between $30,000 and $34,999 from 2013-2017 ACS
35K_40K	Num households w/income between $35,000 and $39,999 from 2013-2017 ACS
40K_45K	Num households w/income between $40,000 and $44,999 from 2013-2017 ACS
45K_50K	Num households w/income between $45,000 and $49,999 from 2013-2017 ACS
50K_60K	Num households w/income between $50,000 and $59,999 from 2013-2017 ACS
60K_75K	Num households w/income between $60,000 and $74,999 from 2013-2017 ACS
75K_100K	Num households w/income between $75,000 and $99,999 from 2013-2017 ACS
100K_125K	Num households w/income between $100,000 and $124,999 from 2013-2017 ACS
125K_150K	Num households w/income between $125,000 and $149,999 from 2013-2017 ACS
150K_200K	Num households w/income between $150,000 and $199,999 from 2013-2017 ACS
200K_MORE	Number of households w/income over $200,000 from 2013-2017 ACS
JOINID	Unique ID
TOTV_19	Number of votes cast in the 2019 mayoral general election
JOYCE_19	Number of votes for Jerry Joyce in 2019 mayoral general election
VALLAS_19	Number of votes for Paul Vallas in 2019 mayoral general election
WILSON_19	Number of votes for Willie Wilson in 2019 mayoral general election
PRECK_19	Number of votes for Toni Preckwinkle in 2019 mayoral general election
DALEY_19	Number of votes for Bill Daley in 2019 mayoral general election
MCCART_19	Number of votes for Gary McCarthy in 2019 mayoral general election
CHICO_19	Number of votes for Gery Chico in 2019 mayoral general election
MEND_19	Number of votes for Susana Mendoza in 2019 mayoral general election
ENYIA_19	Number of votes for Amara Enyia in 2019 mayoral general election
FORD_19	Number of votes for La Shawn Ford in 2019 mayoral general election
SALGRIF_19	Number of votes for Neal Sales-Griffin in 2019 mayoral general election
LHGTFT_19	Number of votes for Lori Lightfoot in 2019 mayoral general election
FIORETTI_1	Number of votes for Bob Fioretti in 2019 mayoral general election
KOZLAR_19	Number of votes for John Kozlar in 2019 mayoral general election
TOTV_RO15	Number of votes cast in the 2015 mayoral runoff election
RO_E15	Number of votes for Rahm Emanuel in 2015 mayoral runoff election

RO_G15	Number of votes for Jesus “Chuy” García in 2015 mayoral runoff election
TOTV_G15	Number of votes cast in the 2015 mayoral general election
EMAN_G15	Number of votes for Rahm Emanuel in 2015 mayoral general election
WILS_G15	Number of votes for Willie Wilson in 2015 mayoral general election
FIORET_G15	Number of votes for Robert Fioretti in 2015 mayoral general election
GARCIA_G15	Number of votes for Jesus “Chuy” García in 2015 mayoral general election
WALLS_G15	Number of votes for William Walls in 2015 mayoral general election
TOTPOP19	Total population from 2015-2019 ACS
NH_WHITE19	Non-hispanic White population from 2015-2019 ACS
NH_BLACK19	Non-hispanic Black population from 2015-2019 ACS
NH_AMIN19	Non-hispanic American Indian and Alaska Native pop from 2015-2019 ACS
NH_ASIAN19	Non-hispanic Asian population from 2015-2019 ACS
NH_NHPI19	Non-hispanic Native Hawaiian and Pacific Islander pop from 2015-2019 ACS
NH_OTHER19	Non-hispanic population of other race from 2015-2019 ACS
NH_2MORE19	Non-hispanic population of two or more races from 2015-2019 ACS
HISP19	Hispanic population from 2015-2019 ACS
H_WHITE19	Hispanic White population from 2015-2019 ACS
H_BLACK19	Hispanic Black population from 2015-2019 ACS
H_AMIN19	Hispanic American Indian and Alaska Native population from 2015-2019 ACS
H_ASIAN19	Hispanic Asian population from 2015-2019 ACS
H_NHPI19	Hispanic Native Hawaiian and Pacific Islander population from 2015-2019 ACS
H_OTHER19	Hispanic population of other race from 2015-2019 ACS
H_2MORE19	Hispanic population of two or more races from 2015-2019 ACS
MAY19LL	Number of votes for Lori Lightfoot in 2019 mayoral runoff election
MAY19TP	Number of votes for Toni Preckwinkle in 2019 mayoral runoff election
GARCIA_G15_pct	GARCIA_G15 as a percent of TOTV_G15
EMAN_G15_pct	EMAN_G15 as a percent of TOTV_G15
RO_GARCIA_G15_pct	RO_GARCIA_G15 as a percent of TOTV_RO15
RO_EMAN_G15_pct	RO_EMAN_G15 as a percent of TOTV_RO15
LL_19_pct	LHGTFT_19 as a percent of TOTV_19
TP_19_pct	PRECK_19 as a percent of TOTV_19
RO_LL_19_pct	MAY19LL as a percent of MAY19LL + MAY19TP
RO_TP_19_pct	MAY19TP as a percent of MAY19LL + MAY19TP
normalized_first_round_garcia	GARCIA_G15 as a percent of GARCIA_G15 + EMAN_G15
normalized_first_round_eman	EMAN_G15 as a percent of GARCIA_G15 + EMAN_G15
normalized_area	Precinct area normalized by the area of the largest precinct
normalized_log_area	Precinct log-area normalized by the log-area of the largest precinct
centroid_x	The longitude of precinct centroid
centroid_y	The latitude of precinct centroid
normalized_centroid_x	Min-max normalized longitude of precinct centroid
normalized_centroid_y	Min-max normalized longitude of precinct centroid
tot_pop_acs	Total population
tot_vap_acs	Total voting age population
civ_vap_acs	Total civilian voting age population
cvap_acs	Total citizen voting age population
gt_19_uninst_civs	Total population of uninstitutionalized civilians older than 19
gt_25_pop	Total population older than 25
gt_16_working_pop	Total working population older than 16
poverty_ratio_ref_pop	Total working population older than 16 for whom poverty status is determined
gt_15_pop	Total population older than 15
tot_h_units_acs	Total housing units
tot_hh_acs	Total households
tot_occ_h_units_acs	Total occupied housing units
uninsured_pct	Pct of gt_19_uninst_civs with no form of health insurance
medicare_medicaid_pct	Pct of gt_19_uninst_civs enrolled in Medicare and/or Medicaid
tricare_va_pct	Pct of gt_19_uninst_civs enrolled in TRICARE or VA health insurance
female_pct	Pct of tot_vap_acs who are female
veteran_pct	Pct of civ_vap_acs who are veterans
married_pct	Pct of gt_15_pop who are married
divorced_pct	Pct of gt_15_pop who are divorced
lt_highschool_pct	Pct of gt_25_pop who have less than a high school education
highschool_pct	Pct of gt_25_pop who have a high school education
some_college_pct	Pct of gt_25_pop who have some college education
associates_pct	Pct of gt_25_pop who have an associates degree
bachelors_pct	Pct of gt_25_pop who have a bachelors degree
grad_and_professional_pct	Pct of gt_25_pop who have a grad or professional degree
drives_alone_work_pct	Pct of gt_16_working_pop who drive alone to work
public_transit_work_pct	Pct of gt_16_working_pop who take public transit to work
walk_to_work_pct	Pct of gt_16_working_pop who walk to work
bike_to_work_pct	Pct of gt_16_working_pop who bike to work
lt_10_min_pct	Pct of gt_16_working_pop whose commute is less than 10 minutes
10_to_30_min_pct	Pct of gt_16_working_pop whose commute is 10 to 30 minutes
30_to_60_min_pct	Pct of gt_16_working_pop whose commute is 30 to 60 minutes

gt_60_min_pct	Pct of gt_16_working_pop whose commute is greater than 60 minutes
receiving_public_assistance_pct	Pct of tot_hh_acs received public asst or SNAP in past 12 months
eng_only_pct	Pct of tot_hh_acs where English is the only language spoken at home
esp_lim_pct	Pct of tot_hh_acs with primary Spanish, limited English
esp_not_lim_pct	Pct of tot_hh_acs with primary Spanish, not limited English
other_lang_lim_pct	Pct of tot_hh_acs w/another lang (not Spanish) primary, limited English
other_lang_not_lim_pct	Pct of tot_hh_acs w/another lang (not Spanish) primary, not limited English
non_computer_pct	Pct of tot_hh_acs without a computer
internet_pct	Pct of tot_hh_acs with a computer and internet
family_pct	Pct of tot_hh_acs consisting of two or more individuals who are related
living_alone_pct	Pct of tot_hh_acs consisting of one person living alone
non_family_multi_member_pct	Pct of tot_hh_acs consisting of multiple unrelated people
mbsa_occupation_pct	Pct of gt_16_working_pop in management, business, science or arts
service_occupation_pct	Pct of gt_16_working_pop in the service category
sales_and_office_occupation_pct	Pct of gt_16_working_pop in sales or office
nrcm_occupation_pct	Pct of gt_16_working_pop in natl resources, construction, maintenance
ptmm_occupation_pct	Pct of gt_16_working_pop in production, transportation, moving
cop_pct	Pct of gt_16_working_pop who are law enforcement workers
poverty_ratio_lt_p50_pct	Pct of poverty_ratio_ref_pop under 0.50 of poverty level
poverty_ratio_p50_p99_pct	Pct of poverty_ratio_ref_pop between 0.50 and 0.99 of poverty level
poverty_ratio_1p00_1p24_pct	Pct of poverty_ratio_ref_pop between 1.00 and 1.24 of poverty level
poverty_ratio_1p25_1p49_pct	Pct of poverty_ratio_ref_pop between 1.25 and 1.49 of poverty level
poverty_ratio_1p50_1p84_pct	Pct of poverty_ratio_ref_pop between 1.50 and 1.84 of poverty level
poverty_ratio_1p85_1p99_pct	Pct of poverty_ratio_ref_pop between 1.85 and 1.99 of poverty level
poverty_ratio_gt_2p00_pct	Pct of poverty_ratio_ref_pop greater than 2.00 of poverty level
occ_per_room_lt_p50_pct	Pct of tot_occ_h_units_acs w/occupancy per room less than 0.50
occ_per_room_p51_1p00_pct	Pct of tot_occ_h_units_acs w/occupancy per room btw 0.51 and 1.00
occ_per_room_1p01_1p50_pct	Pct of tot_occ_h_units_acs w/occupancy per room btw 1.01 and 1.50
occ_per_room_1p51_2p00_pct	Pct of tot_occ_h_units_acs w/occupancy per room btw 1.51 and 2.00
occ_per_room_gt_2p00_pct	Pct of tot_occ_h_units_acs w/occupancy per room greater than 2.00
built_after_2014_pct	Pct of tot_h_units_acs built after 2014
built_2010_2013_pct	Pct of tot_h_units_acs built between 2010 and 2013
built_00s_pct	Pct of tot_h_units_acs built between 2000 and 2009
built_90s_pct	Pct of tot_h_units_acs built between 1990 and 1999
built_80s_pct	Pct of tot_h_units_acs built between 1980 and 1989
built_70s_pct	Pct of tot_h_units_acs built between 1970 and 1979
built_60s_pct	Pct of tot_h_units_acs built between 1960 and 1969
built_50s_pct	Pct of tot_h_units_acs built between 1950 and 1959
built_40s_pct	Pct of tot_h_units_acs built between 1940 and 1949
built_pre_40s_pct	Pct of tot_h_units_acs built before 1940
tot_h_units10	Total housing units from 2010 census
occ_h_units10	Total occupied housing units from 2010 census

Table 1 The full list of variables joined to our dataset. Subsets of these were used in the paper, as described above.