Topological Analysis of Congress

Zach Willert
Math 471: Topology
Macalester College
St. Paul, MN 55105

May 10, 2017

Abstract

In this paper, we examine the topological structure of the legislative branch of the United States Federal government in search of evidence for or against a partisan political divide in Congress. We use roll call data for every member of a given Congress from. To analyze this data, we consider each member of congress to be a data point sampled from a larger space, and we employ the topological method persistent homology. Persistent homology is topological tool used to uncover the shape of data. Applying this technique to our dataset, we find that there is a relationship between political party and the shape of the data. We then propose several novel techniques for comparing the results of the persistence analysis across time. Examining Senate roll call data since WWII, we find that the the structure of our government is quite partisan. Furthermorewe find that the relationship between the homological structure of the Senate and political parties has changed with time, as divisions increasingly exist along partisan lines.

1 Introduction

Recent years have shown an increase in political polarization in America (Duca, Saving, 2014). This polarization has permeated every level of society: the places we live (Massey, Domina, 2009), the TV we watch, the people we choose to date are all increasingly along partisan lines. Additionally, this political polarization has affected every level of government from the local to the federal level. Most notably and perhaps most problematic has been the increasingly partisan behavior in the United States Congress (Mann, Ornstein, 2012). It is in this context that it seems proper to apply the tools of topological data analysis to the legislative branch of the federal Government.

First and foremost, we seek topological evidence relating to theories about polarization. Specifically, we would like a quantifiable and general method to study legislative data. While many claims about polarization analyze rhetoric, bitterness, or the extent to which we disagree about policy, we seek a method which does not depend on such subjective or contextual measures. Using roll call data for all members of any given congress, We employ the topological method persistent homology for our study. Persistent homology is a topological tool used to uncover the shape of data which has proved valuable in a range of applications. While persistent homology can be used to describe manifolds of structures in any finite dimensional space, we will only be using it to find connected components (in a 1 dimensional space).

In addition to a basic application of persistence, we seek methods of comparing the results of this analysis between similar but distinct datasets. Persistent homology is relatively dependent on the accessible data, and there is so far no standard way to compare the results of persistent

	Vote.1	Vote.2	Vote.3	Vote.4	Vote.5	Vote.6	Vote.7	Vote.8	Vote.9	Vote.10	Vote.11	Vot
SESSIONS (R AL)	1	1	2	1	2	1	1	1	1	1	1	
SHELBY (R AL)	1	1	1	2	2	1	1	1	1	1	2	
MURKOWSKI (R AK)	2	2	1	2	2	1	2	1	2	1	2	
BEGICH (D AK)	2	2	1	2	2	2	2	2	2	2	2	
FLAKE (R AZ)	1	2	2	1	2	1	1	1	2	1	1	
MCCAIN (R AZ)	2	2	1	1	2	1	1	1	2	1	2	
PRYOR (D AR)	2	2	1	2	2	2	2	2	2	2	2	
BOOZMAN (R AR)	2	2	2	1	2	1	1	1	1	1	1	
BOXER (D CA)	2	2	1	2	2	2	2	2	2	2	2	
FEINSTEIN (D CA)	2	2	1	2	2	2	2	2	2	2	2	
UDALL (D CO)	2	2	1	2	2	2	2	2	2	2	2	
BENNET (D CO)	2	2	1	2	2	2	2	2	2	2	2	
BLUMENTHAL (D CT)	2	2	1	2	2	2	2	2	2	2	2	
MURPHY (D CT)	2	2	1	2	2	2	2	2	2	2	2	
COONS (D DE)	2	2	1	2	2	2	2	2	2	2	2	
CARPER (D DE)	2	2	1	2	2	2	2	2	2	2	2	

Figure 1: Example Senate Data Taken From The 113th Senate

homology between data sets of different composition or size. There exists no discrete topological invariant similar to persistence in multiple dimensions (Carlsson, Zomorodian, 2009), but we find ways in which persistence can be compared over time as well as many varied opportunities for future work.

2 Data Description

The data we use comes from the website voteview.com, developed by Keith Poole (Poole, 2015). Specifically, we look at roll call data from the United States Senate and House of Representatives. This data describes the vote taken by every member of congress on every bill introduced during a given session of congress. While we do not discuss all of the data included, this website has data on every Senate and House from 1789-2014. All of the computational work described below takes place in the R programming language. We use the readKH function provided by the pscl package to download the data from the source, and then a series of functions we developed to process the data into the preferred format.

Our data for Congress c (either the House or the Senate in a given year) consists of every roll call vote of every congress person within the session. Every data entry describes a "yea", "nea", or "absent" status for every member on every vote. Thus for a Senate with 100 members which voted on 600 bills, the data for that Senate would be a 100 by 600 matrix with a "yea", "nea", or "absent" in every entry. In order to value "yea", "nea", and "absent" on independent scales, we have expanded the data such that every column describes "yea", "nea", or "absent" for a specific vote (before we had v1, v2 ..., now it is v1yea, v1nay, v1absent, v2yea, ...), and every entry is a binary 0 or 1 describing if Congress person i had the specified vote on the specified bill. Thus for a Senate with 100 members which voted on 600 bills, the data for that Senate would be a 100 by 1800 matrix with a 0 or 1 in every entry. A sample is shown in Fig. 1.

We can get a sense for how the whole data set looks by taking a heat map of values for the matrix. Figure 2 shows the heat map matrix for the 113th Senate. Every square represents how

the legislator of the given row voted on the bill of the given column.

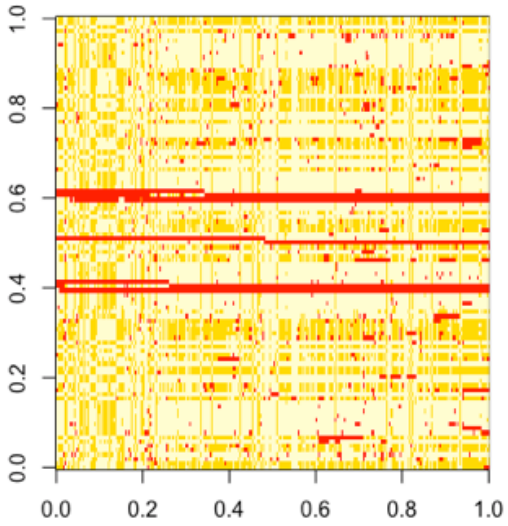

Figure 2: Every vote of the 113th Senate where yellow indicates "Yea" vote, white means "Nay", and red signifies abstention

To better understand the data expansion, Figure 3 is the heat map for the expanded, binary votes. The first third represents "Yea" on all votes, the second vertical third is "Nay" for all votes, and the last third represents abstentions for all votes.

3 Persistent Homology Review

Persistent homology is a topological tool used to estimate the shape of some data. Persistent homology assumes that the observed data is some noisy sample of a higher-dimensional manifold. Under this assumption, we can build a simplicial complex from the data if we give some breathing room (open neighborhood) epsilon around every point to account for the noise. To accomplish this, we must define some distance metric between the data points, and a method for interpreting the metric information as a simplicial complex. The persistence part comes from a continuous deformation of the open neighborhood epsilon, and an encoding of the changing homology in what is called a persistence diagram. The details of this math, though beautiful, are not relevant to this paper, as we stay to the first dimension.

We will be using homology to find the connected components of our data at various epsilon values. This requires data (which we have described above), a metric, and a filtration process. We would like to choose the most simple and intuitive metric to measure distance between these points, so we use the Euclidian metric. This is defined as square root of the inner product of the difference between any two vectors. Every congress person can be thought of as a data point in a high dimensional space. We showed above how the hypothetical Senate with 100 members which voted on 600 bills can be considered a set of 100 points in an 1800 dimensional space. To find the distance between each of these legislators, we simply find the length (2-norm) of the difference

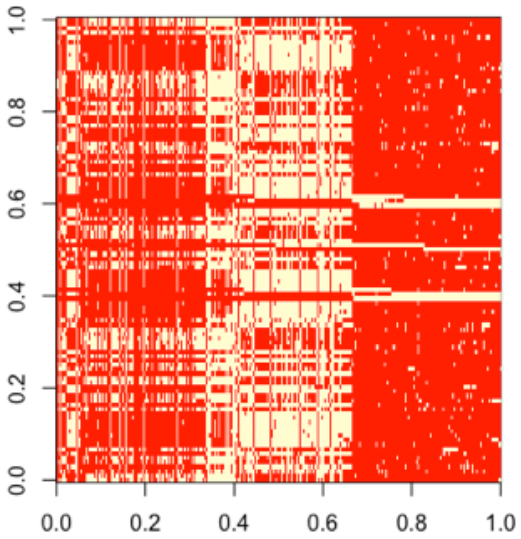

Figure 3: Every vote of the 113th Senate where white indicates a 1 (logical positive), and red signifies 0 (logical negative)

between each of the 1800 dimensional vectors.

In a similar methodology, we would like to use the most simple and unassuming filtration method. We are using the common Rips filtration (as described by Grist, 2007). This process takes a set of points with a defined metric and a given metric distance ϵ, and defines a k-simplex wherever $k+1$ points are pairwise within ϵ. This process forms a simplicial complex which can then be analyzed homologically. Persistent homology is computed when the ϵ is continuously changed over an interval. As ϵ is increased, the simplicial complex changes, indicating the homology of the implied manifold over a range of ϵ values. With our data, metric, and filtration defined, we can move on to the computing.

4 Computing Homology

To start, we will use the TDA package in R. To compute persistence under the rips filtration, we call the ripsDiag function, which requires a data set, a metric (euclidian is default), a maximum epsilon value, and a maximum dimension (to search for manifolds, boundaries). Fig. 4 shows a sample call and output of this function.

We observe one particularly long lasting component which dies at epsilon just over 25. Unfortunately, the TDA package does not share information on the composition of the connected components, but an implementation of the union find algorithm enables us to calculate this information ourselves. Doing so shows us that this final component is comprised of the Senators with the highest abstention rates for this session of Congress. Disregarding these outliers (which dont inform us about partisanship), we can extract the names and parties of the Senators from each of the final two connected components (as shown in Figure 5). This serves as a preliminary test of the value of using persistence to study partisanship.

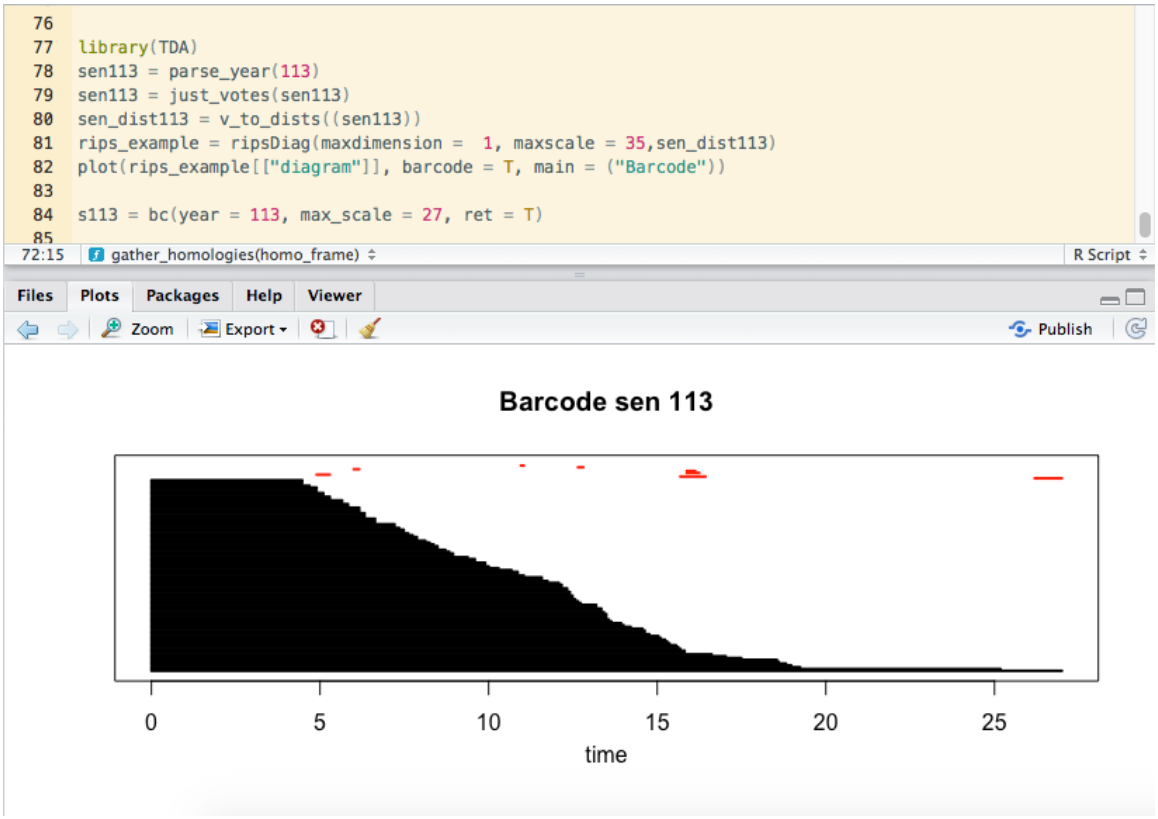

Figure 4: Example function call and barcode plot: 113th Senate

```
> give_groups(v3, as.character(h_test[147,3]))[1:2]
[[1]]
 [1] "SESSIONS (R AL)"  "SHELBY (R AL)"    "FLAKE (R AZ)"     "MCCAIN (R AZ)"    "BOOZMAN (R AR)"
 [6] "RUBIO (R FL)"     "ISAKSON (R GA)"   "RISCH (R ID)"     "CRAPO (R ID)"     "COATS (R IN)"
[11] "GRASSLEY (R IA)"  "MORAN (R KS)"     "ROBERTS (R KS)"   "PAUL (R KY)"      "MCCONNELL (R KY)"
[16] "VITTER (R LA)"    "WICKER (R MS)"    "BLUNT (R MO)"     "JOHANNS (R NE)"   "FISCHER (R NE)"
[21] "HELLER (R NV)"    "AYOTTE (R NH)"    "BURR (R NC)"      "HOEVEN (R ND)"    "PORTMAN (R OH)"
[26] "INHOFE (R OK)"    "TOOMEY (R PA)"    "SCOTT (R SC)"     "GRAHAM (R SC)"    "THUNE (R SD)"
[31] "CORKER (R TN)"    "ALEXANDER (R TN)" "CORNYN (R TX)"    "CRUZ (R TX)"      "LEE (R UT)"
[36] "HATCH (R UT)"     "JOHNSON (R WI)"   "ENZI (R WY)"      "BARASSO (R WY)"   "KIRK (R IL)"
[41] "CHAMBLISS (R GA)" "COCHRAN (R MS)"

[[2]]
 [1] "MURKOWSKI (R AK)"   "COLLINS (R ME)"     "MANCHIN (D WV)"     "BEGICH (D AK)"
 [5] "PRYOR (D AR)"       "BOXER (D CA)"       "FEINSTEIN (D CA)"   "UDALL (D CO)"
 [9] "BENNET (D CO)"      "BLUMENTHAL (D CT)"  "MURPHY (D CT)"      "COONS (D DE)"
[13] "CARPER (D DE)"      "NELSON (D FL)"      "HIRONO (D HI)"      "SCHATZ (D HI)"
[17] "DURBIN (D IL)"      "DONNELLY (D IN)"    "HARKIN (D IA)"      "LANDRIEU (D LA)"
[21] "KING (Indep ME)"    "MIKULSKI (D MD)"    "CARDIN (D MD)"      "WARREN (D MA)"
[25] "STABENOW (D MI)"    "LEVIN (D MI)"       "KLOBUCHAR (D MN)"   "FRANKEN (D MN)"
[29] "MCCASKILL (D MO)"   "TESTER (D MT)"      "REID (D NV)"        "SHAHEEN (D NH)"
[33] "MENENDEZ (D NJ)"    "HEINRICH (D NM)"    "UDALL (D NM)"       "GILLIBRAND (D NY)"
[37] "SCHUMER (D NY)"     "HAGAN (D NC)"       "HEITKAMP (D ND)"    "BROWN (D OH)"
[41] "MERKLEY (D OR)"     "WYDEN (D OR)"       "CASEY (D PA)"       "WHITEHOUSE (D RI)"
[45] "REED (D RI)"        "JOHNSON (D SD)"     "SANDERS (Indep VT)" "LEAHY (D VT)"
[49] "KAINE (D VA)"       "WARNER (D VA)"      "CANTWELL (D WA)"    "MURRAY (D WA)"
[53] "ROCKEFELLER (D WV)" "BALDWIN (D WI)"
```

Figure 5: Final Connected Components in the Persistence Diagram of the 113th Senate: Homology Indicates Partisanship

Figures 6 and 7 show the heat maps of the data after removing the abstainers:

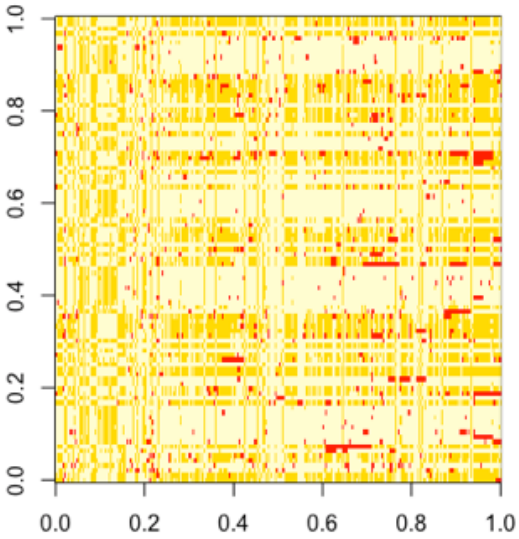

Figure 6: Heat map of 113th Senate, Outliers Removed

5 Comparing Persistence

These initial results look promising, but they are hardly exciting. That voting patterns in the Senate reflect partisanship is common knowledge. So we move now to time series comparison. In the absence of hard mathematical methods for these questions, we will rely somewhat on graphical interpretation. All of the below plots required that I custom-program all of the code– the TDA package was not used in the creation of any of the below plots. This programming project grew quite massive, so Ill not be sharing any particular code in this document. However, the author can be reached via email for questions regarding the code.

First, we will color in the barplots to reflect any fully partisan connected components. This means that for a given epsilon, if any connected components are comprised entirely of one party, the color of the barcode at that point will reflect the party (red for Republicans, blue for Democrats). The barcode will be black wherever a connected component is bipartisan. To clarify, I have outlined the death boundary of the persistence diagram with a purple line. Lastly, I have plotted a the number of senators who are in partisan connected components with respect to epsilon in green. This line gives a rough idea of how partisan the government is for how much of the population is fully partisan for a given epsilon. Additionally, this green line gives us an idea of how responsive partisan separation is relative to total separation: if there were no relationship between connectivity under our metric, then we would expect the number of politically divided Senators to move roughly in step with the number of divided Senators (disconnected components). Here in Figure 8 are barplots from every fourth year between 1949-2015.

At first look, this plot brings mixed results: Partisanship appears to accelerate in the 90's (when most connected components remain partisan until the bitter end), then reach a height in 2001 (when there is no bipartisan connectivity until all Senators join one connected component),

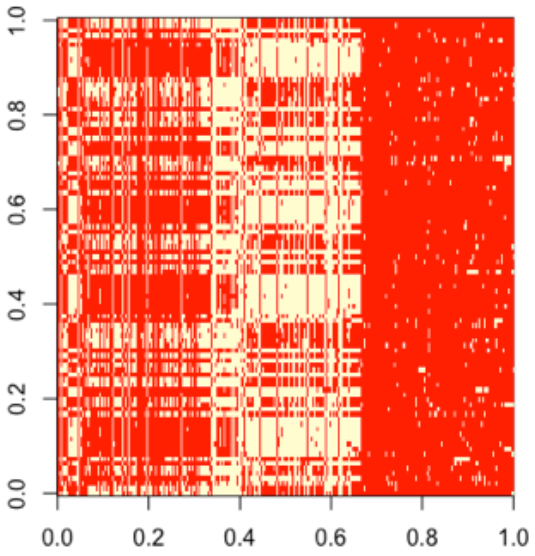

Figure 7: Expanded, Binary Heat map of 113th Senate, Outliers Removed

and finally improve slightly in the Obama years. While it is generally understood that partisanship accelerated through the 90's (Hacker, Pierson, 2010; Mann, Ornstein, 2012), the general consensus is that partisanship stayed the same or got worse in the Obama years. The problem is the inclusion of Independents: Independent Senators (neither Democrat nor Republican) are able to live somewhere in the partisan divide. Because the plot is showing the number of Senators grouped exclusively with their own party, the distinction of Independents (in this case, Bernie Sanders) distorts the divide between Democrats and Republicans. We can assign Independent Senator to the Democratic party (Bernie Sanders historically has been more aligned with Democrats than Republicans) and redraw the plot as Figure 9.

This plot indicates, as expected, that the partisan divide in the Senate didn't improve in the Obama years.

6 Conclusion

As we have seen, persistent homology is a possibly helpful tool in uncovering the structure of legislative bodies. Unfortunately, furthermore, there is substantial topological evidence that our Senate is highly partisan, and there is no evidence that this problem is getting any better.

In this investigation, we have seen some ways in which persistence can be compared across datasets. Unfortunately there isn't time or space to fit in all of my uses of persistence and various findings into this paper. The good news is that this is merely the prequel for an upcoming paper which will address some of the work that remains.

This includes: Adding more data on the government for topological analysis, as well as fully investigating the House of Representatives. Incorporating Bayesian methods to find the most likely manifold shape (requires an understanding of the expected distribution of sampled points given an underlying manifold shape, then generating a hierarchical model which describes the overarching distribution of individual simplex parameters– more theoretical understanding necessary), exploring

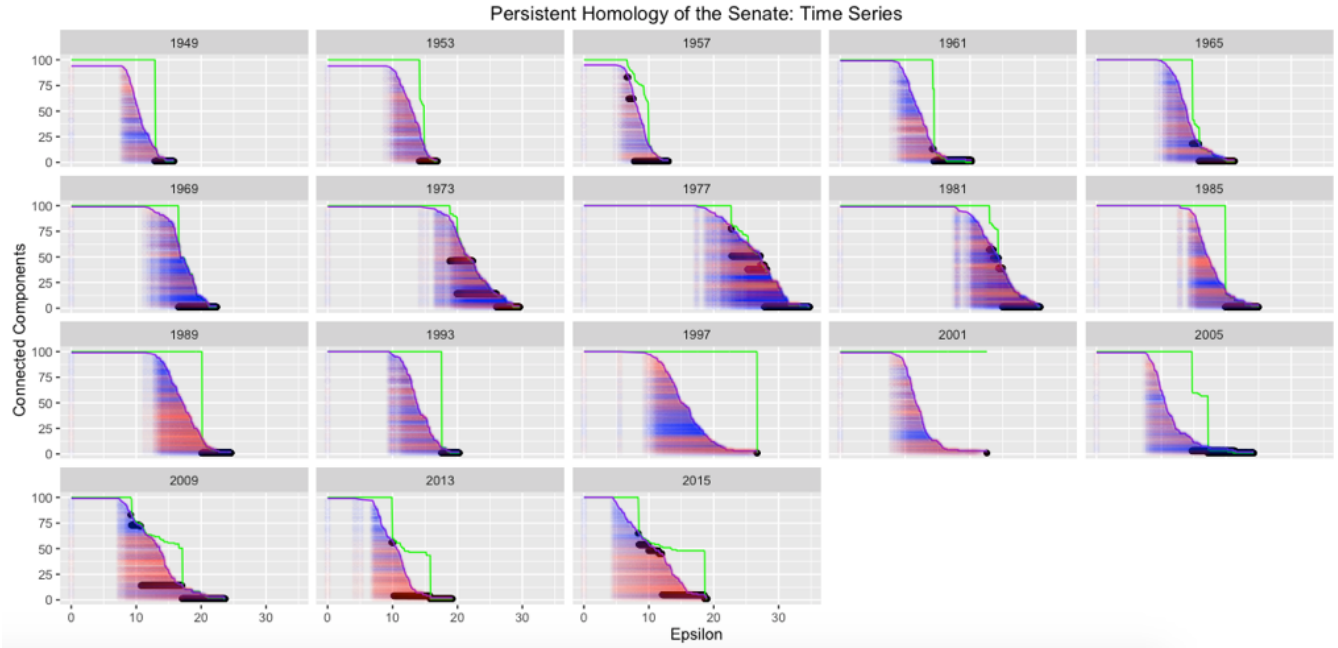

Figure 8: Colored Barplot of Historical Partisan Shape of Senate

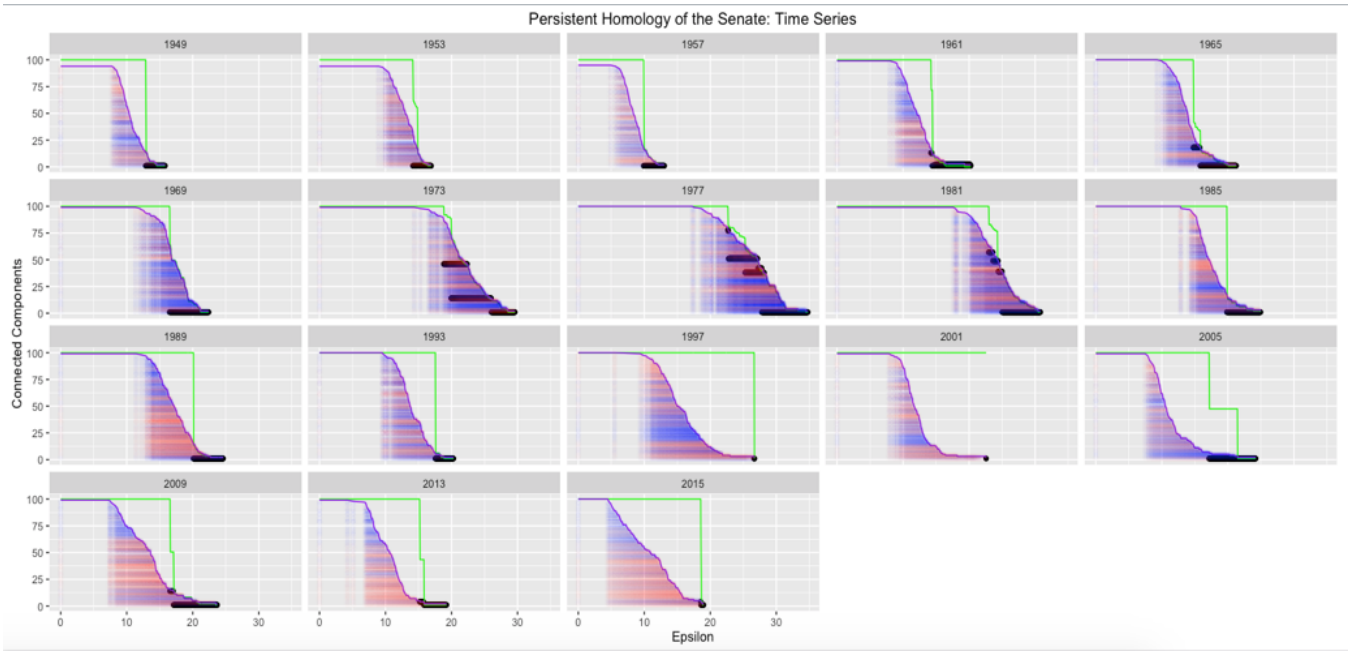

Figure 9: Time-Series Barplots Without Independents

the sigfnificance of the area between the green and purple curves from Figs. 8, 9, developing an understanding of how persistence diagrams would behave under randomly distributed points in various dimensional spaces & with various noise parameters as a potential basis for comparison (of how a given space deviates from a norm), exploration of how we can combine the above ideas into a better understanding of data.

References

[1] Carlsson, Gunnar, and Afra Zomorodian.(2009). *The Theory of Multidimensional Persistence.* Discrete & Computational Geometry 42.1 : 71-93.

[2] Duca, J., & Saving, J. (2014). *Income Inequality and Political Polarization: Time Series Evidence Over Nine Decades.*

[3] Ghrist, Robert.(2007). *Barcodes: The Persistent Topology of Data.* Bulletin of the American Mathematical Society 45.01 61-76. Web.

[4] Mann, T., Ornstein, N. (2012). *It's even worse than it looks: How the American constitutional system collided with the new politics of extremism.* New York: Basic Books.

[5] Massey, D. S., Rothwell, J., & Domina, T. (2009). *The Changing Bases of Segregation in the United States.* The Annals of the American Academy of Political and Social Science.

[6] Poole. (2015). *voteview.com*

7 Bonus plots

I regretted running out of space and not getting to share all these awesome graphs. So I threw two extras in here which I think are rather interesting.

Bonus plots:

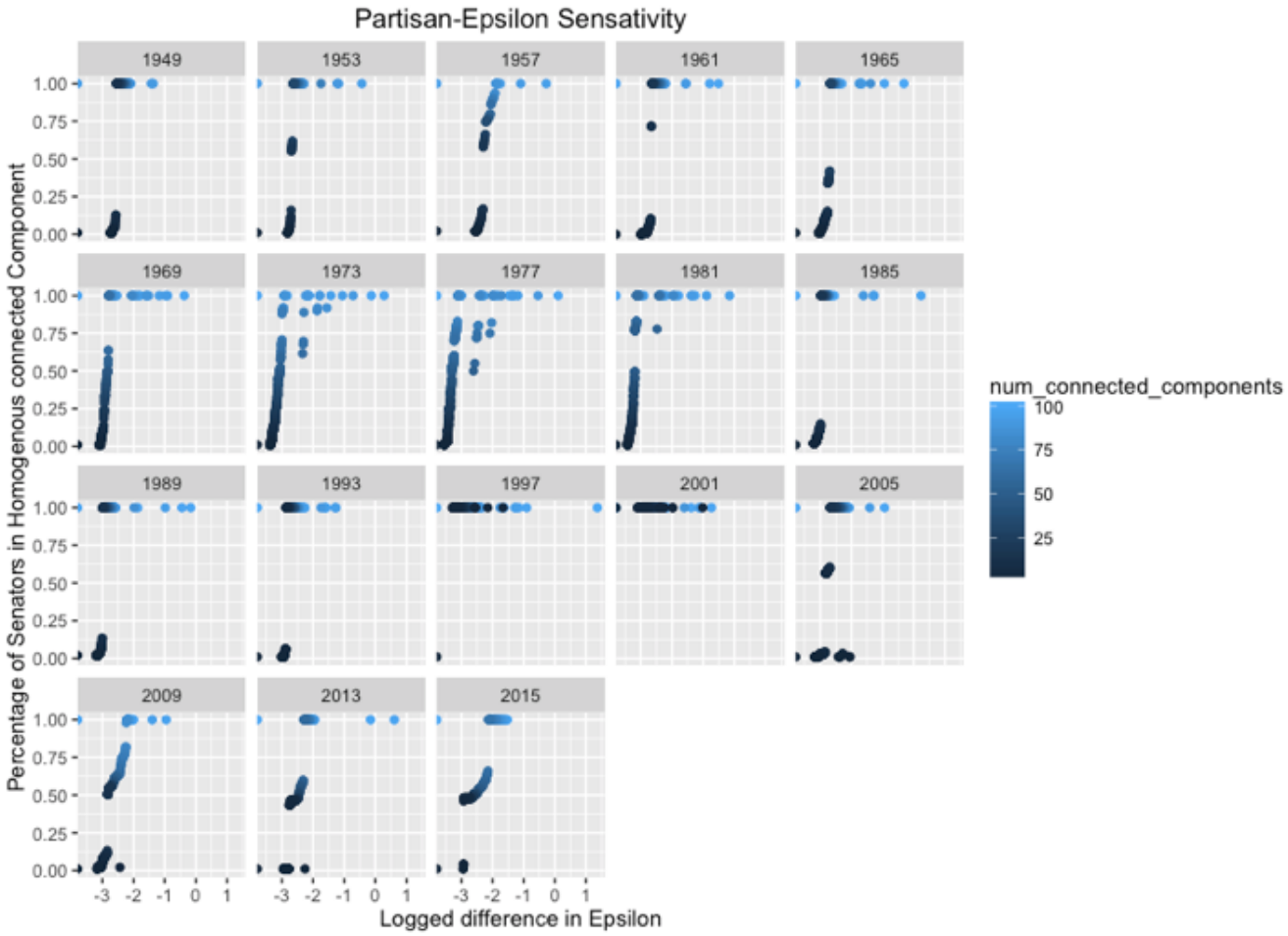

Figure 10: Normalized Partisan-Epsilon Sensitivity by Year

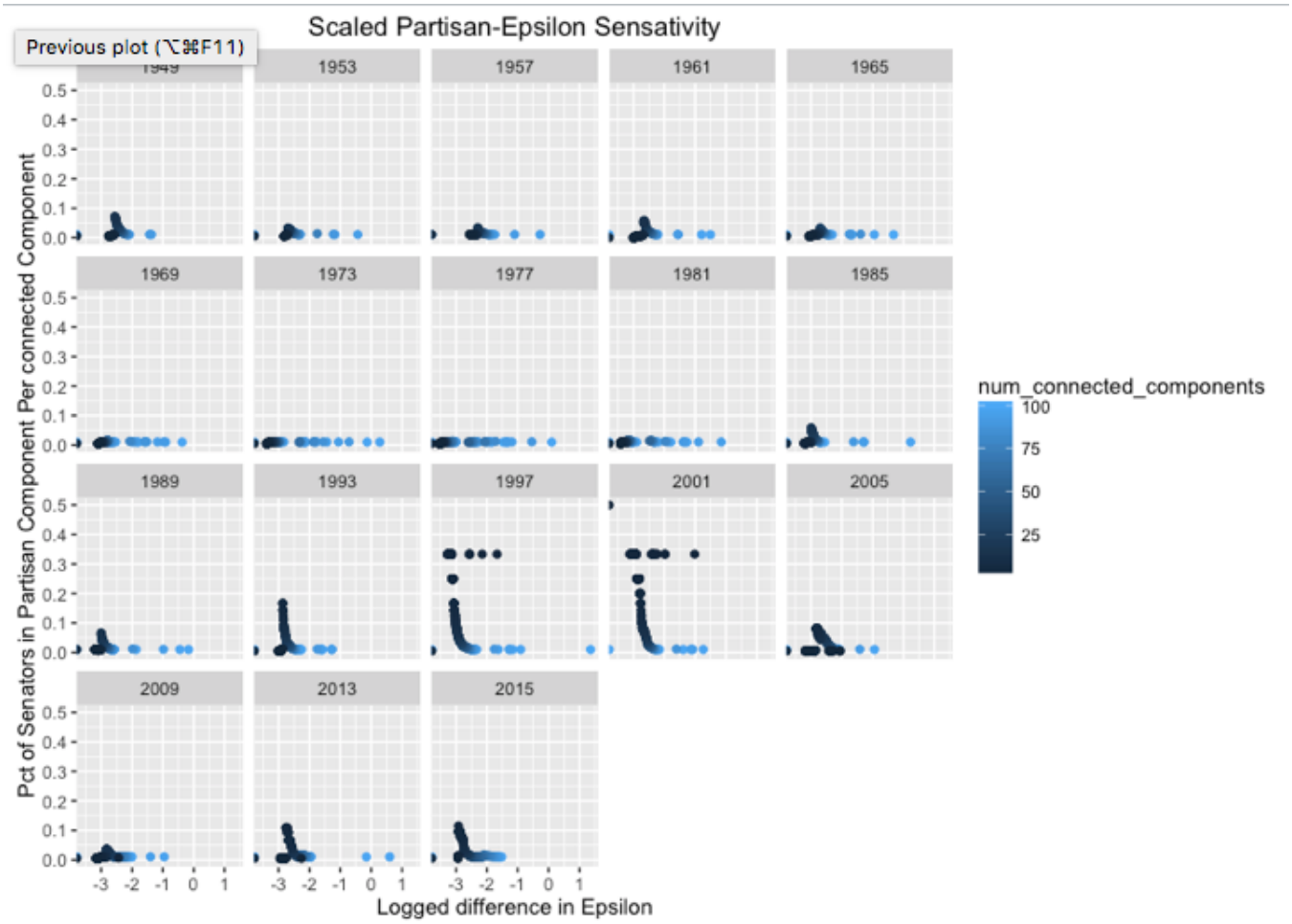

Figure 11: Normalized Partisan-Epsilon Sensitivity by Year