Mapping the Shape of the U.S. Economy: A Topological Data Analysis Approach with BallMapper

*Ryan Johnson**

Graduate, Department of Mathematics, University of Alaska Anchorage, Anchorage, AK

Student: *johnson.ryan-0@pm.me**
Mentor: *scook25@alaska.edu*

KEYWORDS

Topological Data Analysis; Topology; TDA; Ball Mapper; Mapper; Data Science; Economics; Macroeconomics;

ABSTRACT

Topological Data Analysis is (TDA) is a developing data analysis method which gained popularity starting in the early 21st century. Currently, a large body of TDA research utilizes the traditional Mapper algorithm. Some goals of this paper is to show and entry level path into TDA as well as expand the body of literature using Ball Mapper–a traditional Mapper adjacent algorithm. Using U.S. macroeconomic data from the Bureau of Economic Analysis (BEA), the Federal Reserve Bank (FRB), and the Bureau of Labor Statistics (BLS), we achieve our goal by first showcasing Ball Mapper's use in Exploratory Data Analysis. We then show how we can effectively use Ball Mapper in combination with other derived economic measures: the Fisher Equation, PPI-CPI Gap, and Wage Growth. Our findings suggest that Ball Mapper is a useful data analysis tool to look at large multidimensional datasets and extract meaningful relationships in insights about our data quickly [...]

INTRODUCTION

Topological Data Analysis (TDA) is an emerging field of data analysis that is increasing in popularity. Broadly, traditional TDA applications use two tools: an algorithm called Mapper whose output is represented by a mathematical graph and an analysis technique called Persistence Homology.[1][A] These two tools are not necessarily dependent on one another and do provide information about data in their own right. Moreover, the combination of Mapper and Persistence Homology is what forms the central argument for TDA: data contains an underlying shape, and this shape can provide us with qualitative, and sometimes quantitative, insights about large multidimensional datasets.[2]

As a helpful general framework, Mapper can be thought of the visual side of TDA, and Persistent Homology as a the underlying engine. For this analysis we will be focusing on the "visual side" of TDA, the graph creation. Specifically, we are examining various macroeconomic indicators of the United States of America (U.S.) using a Mapper-adjacent algorithm called Ball Mapper(BM). Ball Mapper is of particular interest to us because it reduces the parameters compared to traditional Mapper, allowing for easier implementation. Further, this reduction in parameters removes some of the barriers to start learning TDA; i.e. needed background and coding knowledge.

In traditional Mapper, one must pass the data through three stages, all with their own parameters, in order to generate a graph. Ball Mapper however only needs one parameter before producing a graph.[3] An important trade off between Ball Mapper and traditional Mapper is the following: BM reduces the number of steps needed to produce similar outcomes of traditional Mapper, but one loses control of fine tuning these output. Additionally, traditional Mapper presents itself as needing a minimum mathematical background, introductory course in Topology, or conceptual knowledge of data science methods and algorithms to understand it. Although it is not impossible to learn TDA without the aforementioned backgrounds, without them, TDA might seem unnecessarily complicated for data analysis, or impossible to decipher.

One major motivation for this paper came from a paper by Dłotko, et al.[4] This paper was a two pronged macroeconomic analysis using Ball Mapper. The first was a dataset of five macroeconomic variables from 16 countries from the 1870s-2017. The authors were

---

[A]A mathematical graph, from Graph Theory, is a diagram that consits of nodes and edges.

interested in comparing how countries have evolved over time, their transformation from the Great Depression Era, and various views on wealth and inequality. The second prong was to look at relationship between private credit growth and GDP of various countries. This paper was the only on in the literature that we could find which combines both macroeconomics and Ball Mapper.

If successful this analysis will expand the small body of literature whose main focus is applications with Ball Mapper.[1] We also hope that it will serve as an on-ramp for anyone interested in TDA but feels inundated with jargon upon early stage researching.

Using U.S. macroeconomic data from the Bureau of Economic Analysis (BEA), the Federal Reserve Bank (FRB), and the Bureau of Labor Statistics (BLS), we plan to achieve these objectives in two stages. The first is to show Ball Mapper's use in exploratory data analysis (EDA) by looking at structural observations, coloration, as well as size variation of our graph. Then using what we showed in our first stage, we extend that framework to other derived macroeconomic measures: The Fisher Equation,[5,6] PPI-CPI Gap,[B7] and Real-Wage Growth.[8C]

METHODS AND PROCEDURES

Data Selection & Preparation

We pulled data from three publicly available data from U.S. government sources: The Bureau of Economic Analysis (BEA), The Federal Reserve Board (FRB), and The Bureau of Labor Statistics (BLS).[9,10] The data were gathered using R using two application programming interfaces (APIs): one for data from the BEA, and the other from the Federal Reserve Economic Data (FRED) API.[11,12D] We additionally used recession dates based on business cycle contractions and expansions provided by the National Bureau of Economic Research (NBER). These agencies were selected because of they are authoritative sources for U.S. economic data. Their widespread use in both the private and public sectors gives us high confidence in the accuracy and integrity of the data.[13]

Data Series	Series Abbreviation	Year Start	Year End	Frequency	Series	API
Gross Domestic Product	GDP	1930	2024	Annual	Table 1.1.1	BEA
Personal Income & Its Disposition	PID	1948	2024	Annual	Table 2.1	BEA
Foreign Direct Investment	FDI	2014	2024	Annual	Direct Investment and MNE	BEA
Federal Funds Rate	FFR	1955	2025	Annual	RIFSPFFNA	FRED
Employment Cost Index	ECI	2001	2025	Quarterly	ECIALLCIV	FRED
Consumer Price Index	CPI	1947	2025	Monthly	CPIAUCSL	FRED
New Privately-Owned Housing Units Started	Housing Starts	1959	2025	Monthly	HOUST	FRED
Producer Price Index - Finished Goods	PPI	1947	2025	Monthly	WPSFD49207	FRED
Unemployment Rate	Unrate	1948	2025	Monthly	UNRATE	FRED

Table 1: *Housing Starts is considered a flow because it is provided as a Seasonally Adjusted Annual Rate (SAAR).

Table 1 shows all the data series we considered for this analysis. We excluded Employment Cost Index (ECI) and Foreign Direct Investment (FDI) due to their limited availability of years. Of the remaining data series, New Privately-Owned Housing Units Started (Housing Starts) had the smallest range and so set our lower bound for the range years used (1960-2024). Our upper bound was the latest full calendar year that was available. We wanted to represent the economy from various vantage points. i.e. the broad economy, policy makers, consumers, business owners, etc. Table 2 shows the data series that we felt were representative of the economy as a whole and had data for our time period of interest. Within each of these series we used a subset of columns to construct our final dataset for analysis. To see what parts of the economy we determined each series plays, the column Functional Description, can be found Table 2.

For this analysis we are focusing on an annual time frame and some data transformation was needed. Looking at Table 1, you can see that not all our data was provided on an annual basis. Table 2 shows a summary of the transformation taken and the details for each follow.

[B]Producer Price Index: Finished Goods less Consumer Price Index. Details on data transformation in Methods and Procedures section.
[C]Compesation less Inflation (CPI). Details on data transformation in Methods and Procedures section.
[D]FRED aggregates data from national and international sources, as well as public and private sources.

The first transformation we did was on Personal Income and Its Disposition (Personal Income). Personal Income, consisting of a few different lines of income according to the BEA, is reported in nominal dollars. Nominal dollars do not account for inflation, so using Consumer Price Index (CPI) we transformed our values to Real Dollars (Equation 1).[14] We then took the the log difference (Equation 2) from the output from Equation 1 to get a percentage change from the preceding year.

$$\text{Real Dollars} = \frac{\text{Nominal Dollars}}{\text{CPI}} * 100 \tag{1}$$

This log difference helps to make our data more linear and although not the exact change, is a good approximation.[15] An interesting point that has not been studied to our knowledge is the relationship between linearized data and underlying distance metric that Ball Mapper uses, the Euclidean distance. That is to say, if we did not linearize our data, we do not know how different our results would be, but this is later touched on in the discussion.

$$\Delta \ln (\text{Level}) = [\ln (\text{Level}_t) - \ln (\text{Level}_{t-1})] * 100 \tag{2}$$

Our next transformation was on Housing Starts, Producer Price Index - Finished Goods (PPI) and CPI. We should note that PPI tracks only physical goods that businesses purchase either for their operations or to sell to consumers. Some examples are: fleet vehicles, appliances, groceries, gas, and clothing.[16] All three of these sources were only provided on a monthly time frame, so to get an annual value we used a simple arithmetic mean. Once these data are in annual form, we then found the percent change from the previous year using Equation 2 for their final data in our analysis.

Series Abbreviation	Economic Role	Functional Descriptioin	Transformation Applied
GDP	Growth Composition	Consumption, Investment, Gov., and Trade	None (Source in % Change)
PID	Income Structure	Wages, Entrepreneurship, Gov. Transfers	Real Adjustment & Log Difference
UnRate	Labor Market Dynamics	Labor Distress Level and Trend	None (Annual Rate) & Simple Difference
FFR	Monetary Policy	Cost of Capital and Borrowing Conditions	None (Annual Rate) & Simple Difference
PPI	Supply-Side Signal	Producer Costs & Measures Cost-Push Inflation	NA
CPI	Demand-Side Signal	Cost of Living & Measures Demand-Pull Inflation	12-month Avg. & Log Difference
Housing Starts	Leading Indicator	Physical Residential Production	12-month Avg. & Log Difference

Table 2: All years for analysis are 1960-2024.

BallMapper

In depth descriptions of Ball Mapper's theory have been covered in various papers.[3,17,18] As the aim of our paper is to provided a introductory foray into BallMapper and TDA in general, we will omit some of the more technical details below but provide reference to more in-depth explanations.

Mathematically, the concept of Ball Mapper supposes we are given a dataset X in K dimensions with N observations. Then for some point $x \in X$ and given $\epsilon > 0$, we create a ball, $b(x, \epsilon)$, centered around x with radius ϵ. From there our aim to create set of balls that cover the point cloud.

[REFINE THIS, such that $B = \bigcup_{i=0}^{n} b(x, \epsilon)$ for all $x \in B$.

We call our dataset X a Point Cloud where each of the N rows represents a point in our cloud. Our dimension, K, are each column of our data and commonly our point cloud has dimensions $K > 2$. When we are selecting a point, $x_i \in X$, to draw a ball, we are randomly selecting a row of our point cloud and then taking the Euclidean Distance in K dimensions such that $i \neq j$ and $x_i, x_j \in X$ to every other x_j in our point cloud. Then with our given epsilon, $\epsilon > 0$, we assign any points where the distance is less to epsilon to a ball with center x_i (Equation 3).

$$d(x_i, x_j) = \sqrt{(x_{i_1} - x_{j_1})^2 + (x_{i_2} - x_{j_2})^2 + ... + (x_{i_k} - x_{j_k})^2} < \epsilon \tag{3}$$

As an algorithm, Ball Mapper is as follows:

1. Select a random point x_i from point cloud X.
2. Given $\epsilon > 0$, construct a ball $b(x_i, \epsilon)$ with a center x_i by associating all other points where $d(x_i, x_j) < \epsilon$ and $i \neq j$. (Equation 3)
3. Place (b_i, ϵ) in a set of balls B.
4. Repeat steps 1-3 until all $x \in X$ belong to some $b(x, \epsilon)$.
5. Draw an edge between $b_i, b_j \in B$ if they contain the same x, weighting the edge based on number distinct x values b_i and b_j both contain.

Implementation

As we saw above we need three parts for BallMapper: a Point Cloud, a variable to color by, and an epsilon. For this analysis, our point cloud was made up of 16 variables (dimensions), two of which were created using raw data: *Unemployment Change* and *Fed Rate Change*. We elected to create these first because Unemployment and the Federal Funds Rate were the only two dimensions which did not have a transformation and thus did not have a year-over-year change. Having the rate of change for both Unemployment and the Fed Rate is also important because it gives us additional information beyond the intrinsic value of each of the measures. Both measures on their own give us a current state of the economy at large, the Unemployment Rate telling us how many who are not working but want to,[19] and the Fed Rate signals the structural cost of capital.[20]

To our knowledge there is no generally accepted way of choosing epsilon in the literature. Since Ball Mapper constructs maps by selecting a point and creating a ball with a radius of epsilon, this is should in theory, given an epsilon small enough, produce a graph such that every point is a ball (node) of only itself. Conversely, this should lead to and epsilon big enough such that all of our points are included in one ball (node). For our analysis here, this is indeed the case. So to find an appropriate epsilon we first start by finding these lower and upper bounds.

What we look for in our lower bound is a graph with an epsilon small enough make every point in our point cloud a singleton node, not connected to any other. For our upper bound we do the converse, we look for an epsilon so large where it creates a singular node that houses all of our points. Additionally, to reduce some of the computational time, we look for a lower bound where a connected component started to form, and an upper bound where we had only two connected nodes.

For our data we first found a lower and upper bound which created an interval of $[0.38 - 0.90]$. We then applied a function to generated around 100 graphs and reviewed them, narrowing our interval to $[0.40 - 0.70]$. During this narrowing process we look for interesting features such as connected components forming or dissolving, flares coming off of any components, or notable sizing or coloration patterns. We finally decided on the value 0.474 because it presented a parts of all the aforementioned features we look for: coloration and size patterning, and connected components.

Before describing some visual changes we made to the Ball Mapper output graph, it is important to note that Ball Mapper will output the same graph unless at least one of two things changes: the size of the point cloud changes or the order of the point cloud. The former intuitively makes sense, the algorithm is creating epsilon radius balls so adding or subtracting any data will yield a different amount of points in at least one ball. Depending on the data being changed or number of points, the structure of the output might not change much. In our case, when we were choosing one variable, PPI, we had a multitude of choices. Wanting to look at the the larger picture of the US we first tested PPI - All Commodities. We found a good epsilon at 0.474, however for part of our analysis we wanted to look at the gap between PPI less CPI. As we saw above, PPI tracks finished goods including capital expenditures. In plain language this translates to, physical equipment or machines that help a business produce their product; think, company cars, bottling machines, computers, etc. On the other hand CPI tracks consumer goods which will leave out some of these items tracked in PPI.[21] However, there is some overlap between the two indexes and this gap can serve as a proxy for a leading inflation indicator. Thus we decided to use PPI - Finished Goods because using all commodities will lead to double counting. [MAKING CLEAR ABOUT THE DATA SWITCH]

(a) Ball Mapper output using standard function (ColorIgraphPlot)   (b) Ball Maper output using new graph output (ggraph)
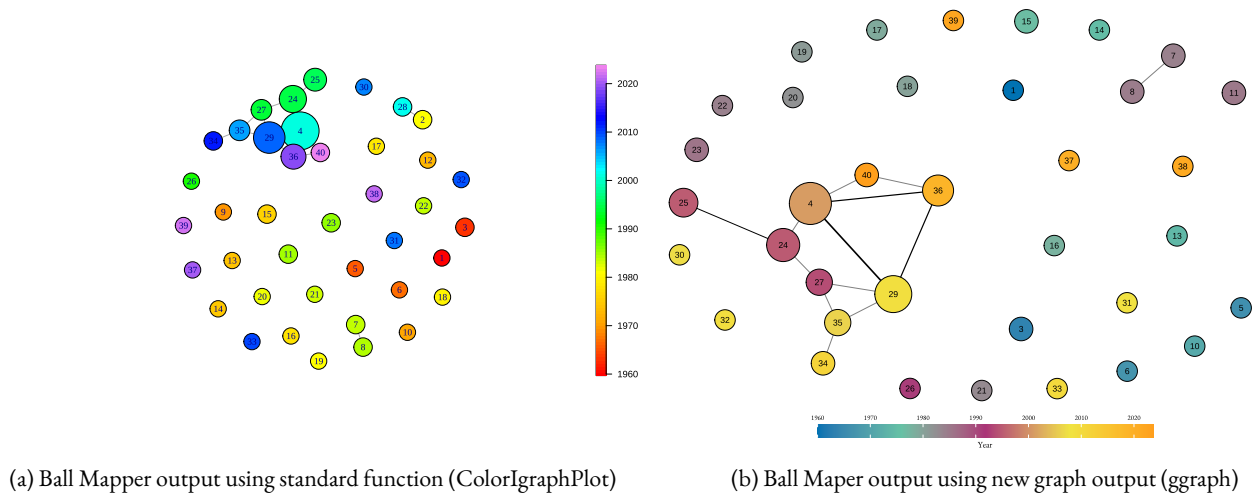
Figure 1: Ball Mapper Graph colored by Year with Epsilon set to 0.511.

The standard Ball Mapper package in R produces a sufficient graph for analysis. It's graphing function uses R base plot function and default coloring scheme. We however, felt that this graph output could be improved upon and recognized that we could use alternative, and more comprehensive packages to create our graph. That is, the output of the BallMapper algorithm could use an alternative package specifically for a network-like graph such as ours. Thus, translating the direct output from the BallMapper algorithm, we created the same graph but changed the visual representation and added another helpful piece of information that we were given but not used previously. Looking at Figure 1, we have the standard Ball Mapper graph output on the left ,Figure 1a, and our new graph output on the right, Figure 1b. The most prominent changes is the spread of the network and coloration.

Having the nodes spread out helps us see the structure of our graph more clearly as well as the new piece of information that was given to us from the Ball Mapper algorithm but not seen in Figure 1a, the edge strength. We see that in Figure 1b, highlighting the edge strength presents an insight about the relation between certain nodes. Although, our data consists of 65 rows, one could imagine the advantage this graph might have with a data set much larger that this one (e.g. looking at this data on a quarterly or monthly basis). Additionally, we changed the color palette for accessibility and readability. Looking at Figure 1a, any nodes colored in the royal blue (years approximately 2010-2015) mask the number identifier of the node.

For Figure 1b and Figure 2, we use Year as our initial coloring variable. We started with this variable because Dłotko, et al.[4] provided a framework for this type of analysis. The aforementioned paper started by coloring their graphs by Year to initially understand the shape of their data. Later, in the second half of their analysis, they colored their graphs by each of the dimensions of a different point cloud than the first. Because we have 16 dimensions to our point cloud, a more in depth analysis is needed and is the the subject for another paper. Additionally, when Dłotko, et al. colored by Year, it was not in their point cloud, and seeing usefulness of this technique, we extended this idea to our point cloud. Specifically, our calculated variables that represented different economic lenses: the Fisher Equation (Figure 3), PPI-CPI Gap (Figure 4), and Real Wage Growth (Figure 5).
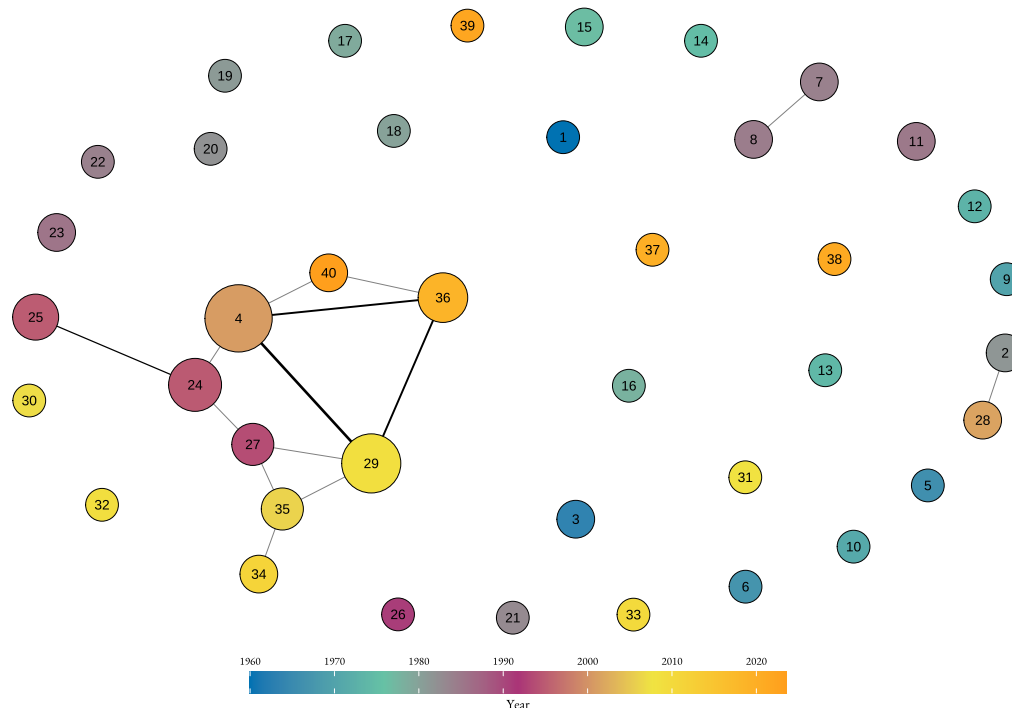
RESULTS

Exploratory Data Analysis

Figure 2: Ball Mapper Graph colored by Year with Epsilon set to 0.474.

There are two general categories of nodes when interpreting Ball Mapper outputs: connected or not connected. Figure 2 shows three different connected nodes (connected components), and many nodes not connected to any others (singletons). These singletons suggests that the data is unlike all the other data points and indicates possible outliers in the dataset.[4] The three connected components, there is one large and two small components. Moving clockwise, starting with the large component and will label them C1,[E] C2,[F] and C3[G] for conversational ease.

Looking at C1, we note that one large difference between it and C2 and C3 is that it has more than two nodes. Since it does, any nodes which are not directly connected by an edge (e.g. Nodes 24 and 4), can be interpreted as alike in some fashion, but also signal a significant difference within the component.[4] What is unclear, though is if path length between two nodes is commensurate with the similarity in data and something that could be explored for future research. [THINK ABOUT MOVING THIS PARAGRPAH TO DISCUSSION, AT LEAST THIS SENTENC(BEFORE)] This also applies the nodes of large size that are connected. We have not found any literature that supports this idea that proximity is commensurate with relatedness. What we can conclude about size is that the size of the node indicates the number of points that nodes has. Similarly, our edge strength is shown this ways with a thicker line representing more shared points between two nodes.

Most often coloration of Ball Mapper graph is used to highlight a variable of interest, or certain know structures in the data. However, we also will look for any patterning (e.g. a smooth gradient from one end of the graph to the other), or nodes which do not conform to patterning. For our map some coloration of interest is the three sub-groups of colors in C1 as well as the mismatching colors for C3. It is important to remember when looking at coloration that the value of the color is an average value and is susceptible to large variances.

Since out graph consists of 40 nodes and three components we will provide all the singletons in a table (Table 3), and talk more in depth about our components below.

[E]Nodes: 4, 24, 25, 27, 29, 34, 35, 40
[F]Nodes: 7 and 8
[G]Nodes: 2 and 28

Node	Year(s)	Node	Year(s)
1	1960	19	1981
3	1962, 1965	20	1982
5	1966	21	1983
6	1967	22	1984
9	1970	23	1985, 1986
10	1971	26	1991
11	1972, 1998	30	2007
12	1973	31	2008
13	1974	32	2009
14	1975	33	2010
15	1976, 1977	37	2020
16	1978	38	2021
17	1979	39	2022
18	1980		

Table 3: Singlton nodes and the years they contain.

Still using Figure 2, we first will examine C2 closer. Adding to some of our initial observations we see that C2's nodes are colored very similarly, and are of the same size. As a relative measure, using Table 3, we look at our singletons and notice that many only hold one point. Thus comparing C2 nodes we should expect around one to two points in each of the nodes, as well as C3. When looking at what points C2 contains we find the years 1968 and 1969 in nodes 7 and 8, respectively, and 2000 being in both nodes, creating the edge. C2 is a good example where coloring by average value can be misleading. Conversely, looking at C3, we first notice that both nodes are colored differently. This inconsistent coloration could be meaningful if we had a strong global coloration pattern, however since C3 is such a small component, much like a small sample size, does not provide us with a lot of further detail than observation. Looking at the nodes we find that node 2 contains 1961 and node 28 contains 2001, with 2002 being our bridge year between the two.

As mentioned above, C1 contains three subgroups of different colors. For the left flare in C1 (nodes 25, 24, 27), all the years are in the range of 1987-2006. These three nodes have 13 points total and a interquartile range (IQR) being between 1990 and 1996, and is consistent with this flare's coloration. Moving towards the bottom flare and connection with the main body of C1 (nodes 34, 35, 29), visual inspection tells us we should expect to find years from the mid 2000's to the 2010's. Upon inspection we also find that these three nodes contain 13 points, however the IQR and range are larger with the IQR being 2004-2014 and range being 1993-2018. Next looking at nodes 36 and 40 we find that they consist of 7 points with a range 2016-2024, a much tighter range than the two previous sub groups. Our last node, node 4, has a bit of all of the sub groups. It points range from 1963 to 2024, with and IQR of 1996-1026, much larger. Indeed, we find the coloration consistent with its wide range, but we also see how wide ranges can be misleading.

Focusing on the structure of C1 there are two distinct features that we notice: the three largest nodes are all connected to each other (nodes 4, 29, and 36), and the sweeping arm on the left side of the component that contains two flares at each end (nodes 25, 24, 27, 35, and 34). The three large nodes that form a triangle, they also have the strongest strength between one another relative any other two nodes. The next strongest connection is between nodes 25 and 24. Between the three nodes we find that 2017 and 2018 are the only years which reside in all three nodes. The middle portion of our sweeping arm (nodes 24, 27, 35) contain most points during the late 80s and the early-mid 90s. The flares (25 and 24) are rather different than the middle nodes. Node 25 has a wide range of years, skipping from 1990 to 1995 and then to 2006. Node 34 on the other hand has years 2011 and 2012, a very tight grouping of years.

Fisher Effect

Using the Fisher Equation we colored our graph with the approximated real interested rate (Equation 4).[5] Albeit, the Fisher Equation is an approximation and does not hold well when nominal interest rates (Fed Rate) are relatively high and with time spans are short, Figure 3 shows us a stark picture of our economic history and important time periods.[6]

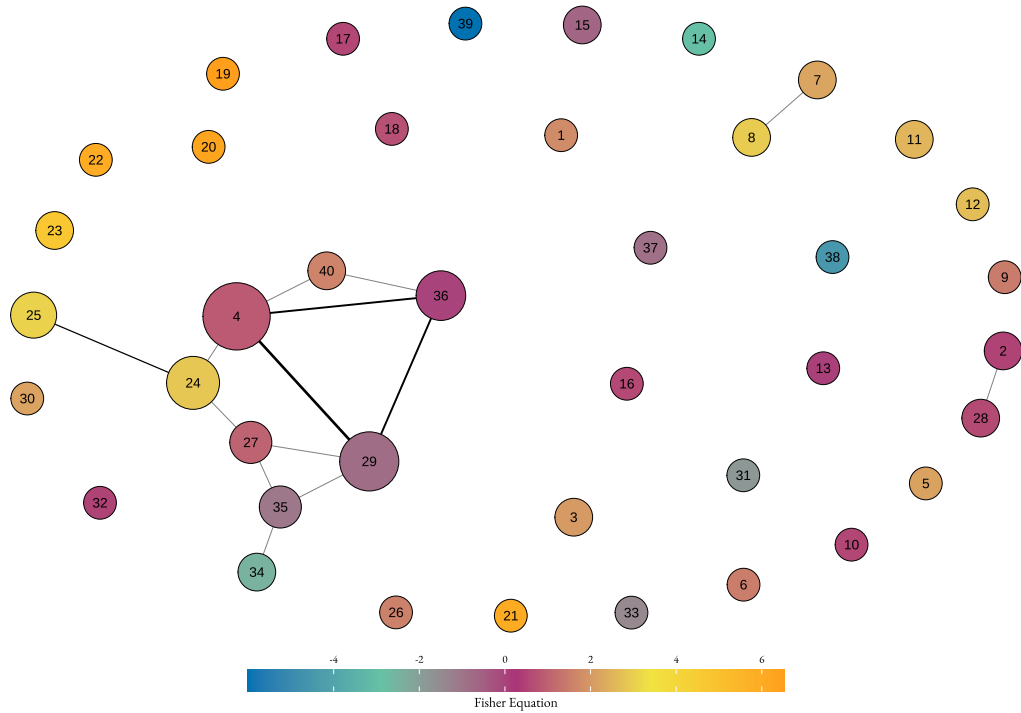$$\text{Real Interest Rate} \approx \text{Fed Rate} - \text{Inflation} \tag{4}$$

Figure 3: Ball Mapper Graph colored by Fisher Equation with Epsilon set to 0.474.

The most prominent nodes that stand out will be those on the far ends of our scale on the bottom of Figure 3. Looking at Equation 4, we see that when the Fed Rate is lower than Inflation, we get negative values, and vice versa when the Fed Rate is later than Inflation.

Singleton nodes 19-22 have relatively high values (average of 6.178%) consist of years 1981-1984. This directly lines up with the experience "Volcker Shock".[22] On the opposite end, moving from the twelve o'clock position clockwise, we can visually see nodes 39, 14, 38, 31, 33, 34 (2022, 1975, 2021, 2008, 2010, 2011-2012 respectively) show a negative value (Equation 4).

We also observe that in C1, most nodes along with C3 seem to have Real Interest Rates showing between -1% and 2%. To contrast this, we see that nodes 25, 24, and C2 look to be above 2% but not above 4-5%.

PPI-CPI Gap

Figure 4 shows us the difference between PPI - Finished Goods and CPI (consumer inflation). More specifically, this graph shows us percent change from the previous year of at-cost prices (before adding margin) that producers are paying. Consumer Price Index (CPI) is the "inflation" measure. It is important to note that although CPI is used for general inflation, the Federal Reserve prefers Personal Consumption Expenditures because it tracks more normal consumer behavior such s substitution of products.

Similar to Figure 3, and subsequently Figure 5, we see that singleton nodes tend to be colored at the extreme ends and our components having mostly a consistent coloring. Node 39, 39, and 13 (2022, 2021, 1974) show high positive values indicating higher PPI values relative to Inflation. On the opposite end, the nodes which represent highest inflation relative to PPI values are nodes 23, 22, 20, 37, 9, 2, 28, 26, and 32 (1985-86, 1984, 1982, 2020, 1970, 1961 and 2002, 2001-02, 1991, and 2009 respectively).
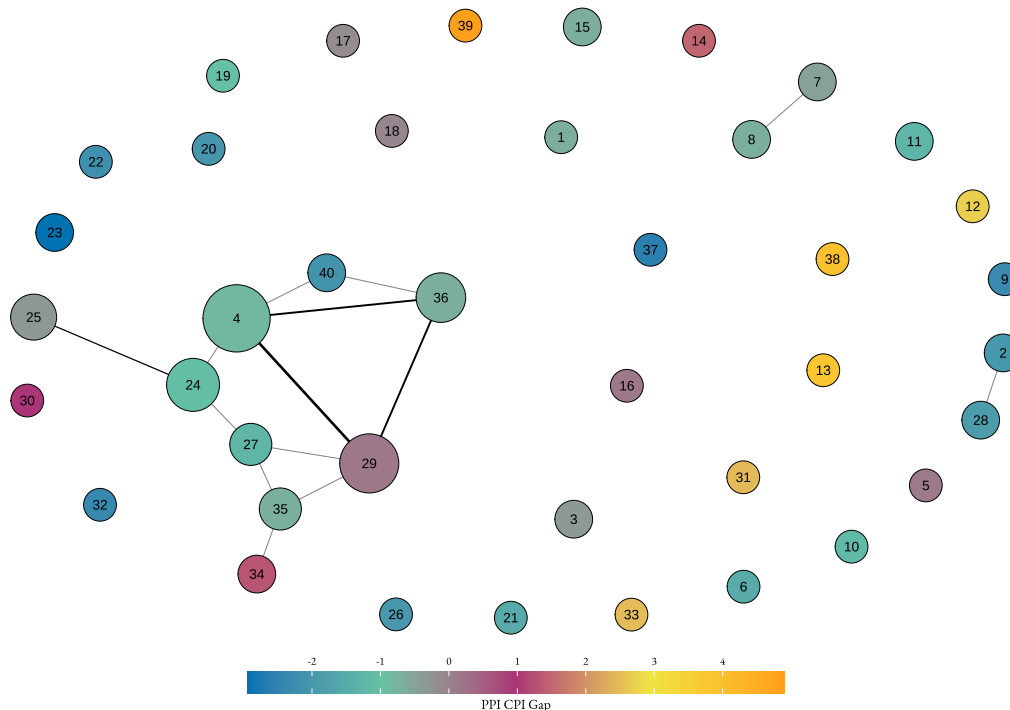
Figure 4: Ball Mapper Graph colored by PPI less CPI with Epsilon set to 0.474.

Wage Growth

For Figure 5, we are looking at the difference between Compensation and Inflation. With this measure we were looking at if people annually were getting paid faster than inflation - i.e. wages were keeping up with inflation. Similar to The Fisher Equation and the PPI-CPI spread, positive numbers indicate that wages are growing faster than inflation, and if inflation is outpacing wages, then people are effectively taking a pay cut.

Looking at nodes 19, 17, 18, 39, 14, and 13,[H] we see that these are the worst in terms of inflation outpacing people's compensation. Indeed, when we look at the years contained in the nodes, they are years which had record breaking inflation - The 1970s Oil Embargo, The Volcker Shock, and 2022 once all the stimulus cash made its way through the economy. On our positive end we have nodes 11, 5 and 3 which consist of years 1972 and 1998, 1966, as well as 1962 and 1965. What is interesting that we don't see in Figure 3 or Figure 4 is that the legend shows a long tail with our zero point being on the far right end. Notice, most of our points are shades of yellow (around zero or little below) but we have very few values that represent negative double digits.

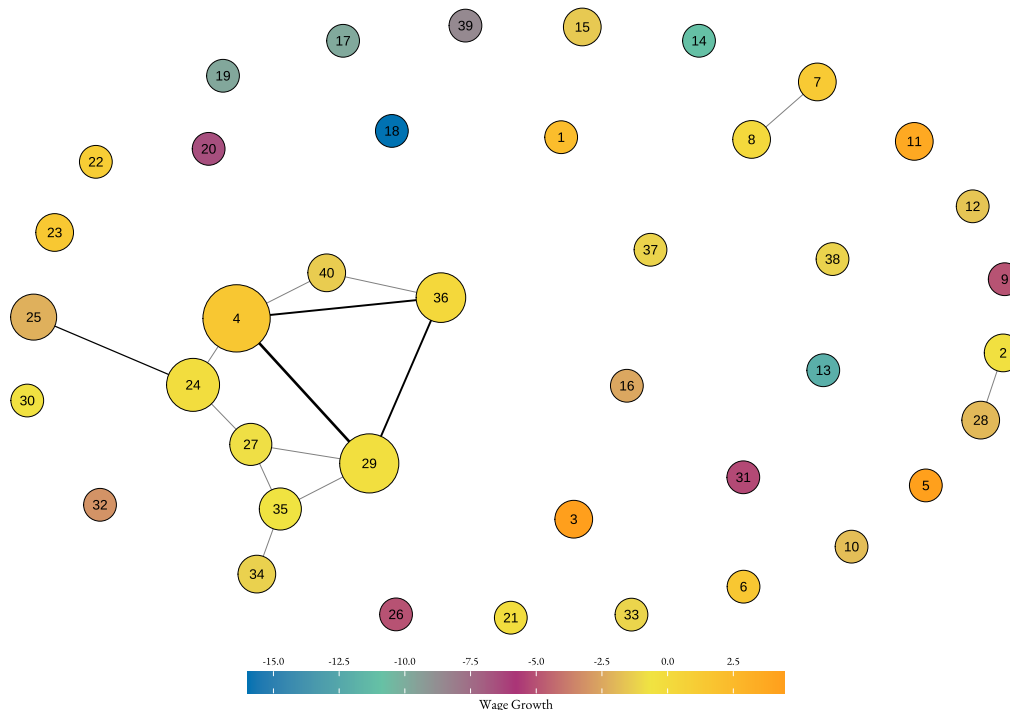[H]1981, 1979, 1980, 2022, 1975, 1974 respectively

Figure 5: Ball Mapper Graph colored by Wage Growth with Epsilon set to 0.474.

DISSCUSSION

When first looking our graph, it is hard not to notice C1 and then see that there are two other components. What was so striking about C1 was that there seem to be two shapes, a triangle and a arc that seems to be attached to the triangle's left side. Although we have not found any literature that explores the relationship between nodes and edge strength (how many points they share), it was something immediately of interest. Since our point cloud was made up of economic data, we considered that these nodes (4, 29, and 36) could represent a certain extended time period or "phase" of the economy. [EXTEND PARAGRAPH TO EXTEND PHASE OF ECONOMY IDEA] Further, we extended this idea to the arc on the left side of C1 with the two flares, as well as C2 and C3.

Starting with C1 we found that nodes 4, 29, 36 largely represented different expansionary states of the economy. Looking at our complete data set [UPDATE: need to put in blurb about "full" data set in the data section.] we find that for all points in our triangle, all years

To hark back to our discussion of C1, when we investigate nodes 4, 24, and 26 we indeed notice a pattern with the three generally: for the whole year the US was in an expansionary state. Node 4 shows us an "ideal" version of the economy. We have personal income and compensation slightly higher than inflation, sitting around 2%. Unemployment is a little elevated but historically not unusually high. For Nodes 24 and 26 we see two different sides of large economic shocks. Node 24 shows us years where the economy is starting to get "over it's skis". That is, it is starting growing faster than it can keep up with but it's not necessarily too late. Node 26 on the other hand consists of years following right after recessions or points where the economy was over heating. Most of the years in Node 26, follow what economist call a "Jobless Recovery".[23] Usually marked by growth returning to the economy but unemployment being "sticky" and not falling as business starts to take off. For the flares and smaller points that are connect each of our large nodes in C1, we find that some years are in both the smaller node and larger one, which creates our edge. However they also contain additional years which are related nodes 4, 24, or 26 but instead describe transition years to these large nodes.

C2 represents a textbook definition of demand-pull inflation. The formation of this component seems to be driven by the distinct and parallel macroeconomic structure of each point. 1968, 1969, and 2000 were a culmination of factors which led the economy into a state of "too much money chasing too few goods."[24] In the late 1960s, the demand was primarily driven by an injection of cash from Johnson's "Great Society" policies and spending on the Vietnam War during the mid-1960s.[25] Conversely, in 2000 demand accelerated due to the private sector increasing their capital expenditures in anticipation of the digital age.[26]

Additionally, demographic and psychological factors played a role during these years' demand shocks. The 1960s saw the Baby Boomer generation hitting their prime working-age years. This rise in the workforce and subsequent creation of households led to an increase in consumption and lower unemployment - signs of a healthy economy. 2000 presented a similar outcome of high consumption via a different mechanism. As mentioned previously, corporate spending was accelerating. This led to high stock valuations and, in turn, high stock market returns – and a general rise in asset values. Known today as the "Wealth Effect", this led consumers to increase their spending, placing more upward pressure on demand.[27]

The small component (C2), nodes 2 and 27, are of interest because it tells us us that these related nodes are somehow distinctly different from C1 and all other nodes. Seeing a small components like this sometimes can indicate outliers in data. In our case, this could be years where large economic events happened such as a recessions or extraordinary growth. However, we also can see this behavior in singleton nodes much like we have Figure 2. When we look at the coloration of the singleton nodes, we see indication of outlier events such as The Great Recession and the COVID-19 Pandemic.

Further this short time period not only is represented by interesting about why these nodes are colored is that these nodes see to show the impact of Volker's decision to raise interest and the lag that took place when raising rates.[28]

In these year rates were already extremely low, or the economy just came off a huge economic shock. In 1975 the fed aggressively lowered rates due to climbing inflation and high unemployment rates from the 1973-'74 oil embargo.[29] In 2021-'22 the COVID-19 Pandemic stimulus checks were going to cause inflation but the fed did not want to raise rates too aggressively.

Positive values indicated that businesses are having their revenue rise higher than consumer inflation. This could mean various things: corporate margins are expanding, they are passing through cost to customers, or.... # CONCLUSION
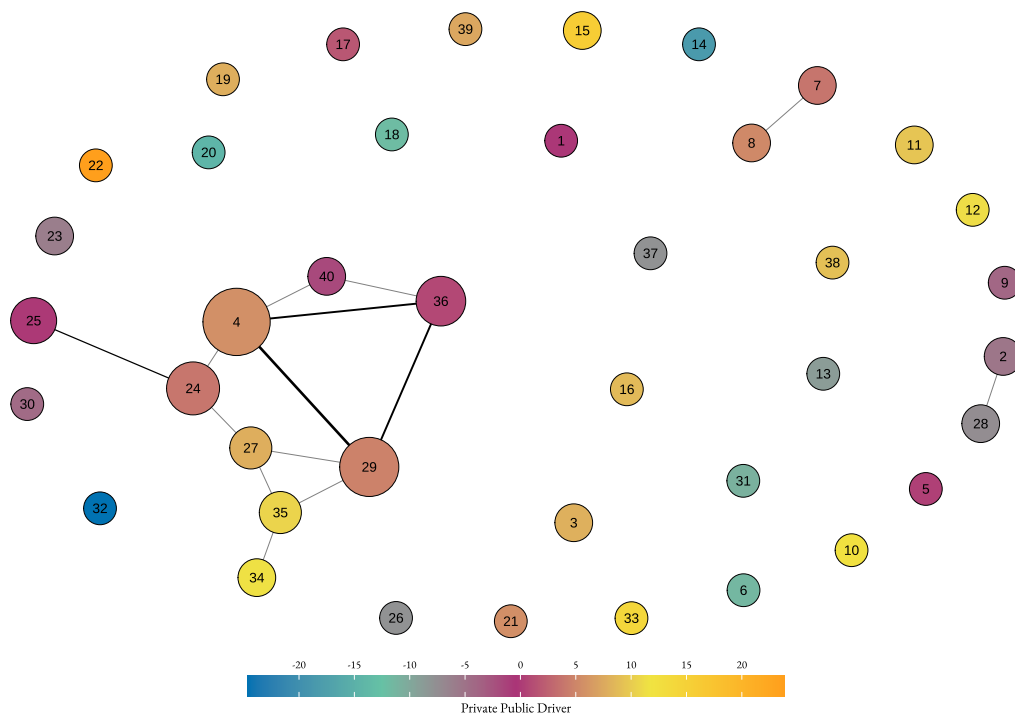

Figure 6: Ball Mapper Graph colored by Private Public Engine with Epsilon set to 0.474.

ACKNOWLEDGEMENTS

REFERENCES

[1] V.N. Madukpe, B.C. Ugoala, and N.F.S. Zulkepli, "A comprehensive review of the Mapper algorithm, a topological data analysis technique, and its applications across various fields (2007–2025)," arXiv Preprint arXiv:2504.09042, (2025).

[2] F. Chazal, and B. Michel, "An introduction to topological data analysis: Fundamental and practical aspects for data scientists," Front. Artif. Intell. 4, 667963 (2021).

[3] P. Dłotko, "Ball mapper: A shape summary for topological data analysis," arXiv Preprint arXiv:1901.07410, (2019).

[4] P. Dłotko, S. Rudkin, and W. Qiu, "Topologically mapping the macroeconomy," arXiv Preprint arXiv:1911.10476, (2019).

[5] S. McClung, "Fisher effect," (2024).

[6] P.G. Michaelides, "The fisher equation," in *21 Equations That Shaped the World Economy: Understanding the Theory Behind the Equations*, (Springer, 2024), pp. 113–120.

[7] Bureau of Labor Statistics, "How does the producer price index differ from the consumer price index? Comparing the personal consumption PPI with the CPI," (2023).

[8] D.G. Sullivan, "Trends in real wage growth," Chicago Fed Letter (115), (1997).

[9] Bureau of Economic Analysis, "Interactive data application," (n.d.).

[10] Federal Reserve Bank of St. Louis, "Federal reserve economic data (FRED)," (n.d.).

[11] A. Batch, J. Chen, and W. Kampas, *Bea.r: Bureau of Economic Analysis API* (2018).

[12] S. Boysel, and D. Vaughan, *Fredr: An r Client for the 'FRED' API* (2021).

[13] E. Hughes-Cromwick, and J. Coronado, "The value of US government data to US business decisions," J. Econ. Perspect. **33**(1), 131–146 (2019).

[14] Federal Reserve Bank of Dallas, "Deflating nominal values to real values," (n.d.).

[15] J.D. Hamilton, and M. Chinn, "Use of logarithms in economics," (2014).

[16] Bureau of Labor Statistics, "Producer price indexes: concepts," (2025).

[17] S. Rudkin, "An introduction to topological data analysis ball mapper in r," arXiv Preprint arXiv:2504.14081, (2025).

[18] W. Qiu, S. Rudkin, and P. Dłotko, "Refining understanding of corporate failure through a topological data analysis mapping of Altman's Z-score model," Expert Syst. Appl. **156**, 113475 (2020).

[19] Bureau of Labor Statistics, "Current population survey: concepts," (2018).

[20] Board of Governors of the Federal Reserve System, "Monetary policy: What are its goals? How does it work?" (2021).

[21] Bureau of Labor Statistics, "Consumer price index: concepts," (2025).

[22] M. Bryan, "The great inflation," (2013).

[23] N.A. Kolesnikova, and Y. Liu, "Jobless recoveries: Causes and consequences," The Regional Economist, 18–19 (2011).

[24] Investopedia, "Demand-pull inflation: Definition, how it works, causes, vs. Cost-push inflation," (n.d.).

[25] D. Marsh, "Look at 1960s, not 1970s, to learn how US inflation took hold," OMFIF, (2021).

[26] Evercore Wealth and Trust, "Booms and busts: A brief history of capital spending cycles," (2024).

[27] A.L. Sussman, "New estimates of the stock market wealth effect," NBER Digest, (2019).

[28] Federal Reserve Bank of St. Louis, "What are "long and variable lags" in monetary policy?" (2023).

[29] M. Corbett, "Oil shock of 1973–74," (2013).

Old Paper

Discussion

In our section we briefly went over two ways to interpret TDABM graphs. As seen in many other, longer papers written by Dlotko and colleagues, there is a lot of room to expound our TDABM graph here,though that is beyond the scope of this paper. In general, though, we can see that TDBM can give insight into our data that we might not see otherwise. Seeing that there is data that is not like the others gives an indication that there is something to investigate. Consider our comparison of the 08' - 09' financial crisis and the 2020 COVID Pandemic: We see learned that the years following the financial crisis were similar economically. Another way to view this is, clustering revels that the financial crisis was a singular problem that spurred the collapse of the financial system. Conversely, the COVID pandemic was a singular issue but exposed many different weaknesses in our current world.

For the future development of TDABM, and TDA in general, there are many possibilities of future research and one glaring downside. The downside to TDA is that it is a high barrier to entry to understand the methodologies inner workings. Those with a technical background will have an easier time, but nonetheless a higher barrier than methods such as linear regressions. Subject wise, TDA has an advantage in some of the social sciences because it can be somewhat of a bridge between qualitative and quantitative analyses. TDABM has been used to analyze the Brexit Vote data, seeking to understand quantitative and qualitative motivations behind its outcome. TDA also has the advantages of dimension reduction which could be useful to the life-sciences. Fields such as genetics, chemistry, and biology have many fields of data, and being able to consolidate it into an understandable form could benefit the body of knowledge greatly.

Dimension/Variable	Data Series
Year	All
PPI Finished Change	Producer Price Index - Finished Goods
Personal Income	Personal Income & Its Disposition
Compensation	Personal Income & Its Disposition
Entrepreneurship	Personal Income & Its Disposition
Transfers	Personal Income & Its Disposition
Consumption	Gross Domestic Product
Domestic Investment	Gross Domestic Product
Government Spending	Gross Domestic Product
Exports	Gross Domestic Product
Imports	Gross Domestic Product
Inflation	Consumer Price Index
Unemployment	Unemployment Rate
Unemployment Change	Unemployment Rate
Fed Rate	Federal Funds Rate
Fed Rate Change	Federal Funds Rate
Housing Change	New Privately-Owned Housing Units Started

Table 4: Variables inculde in BallMapper Point Cloud

Draft Review

INTRODUCTION - frame tda goals regarding statistics rather than provide the analogy, change how am presenting TDA, instead of maiking claim betwen analysis and stats - lens (end of paragraph), change how I am portraying

METHODS AN DPROCEDUES - clairify functiontaion decription