Topological Data Analysis

Utkarsh Kumar, Santoshmurti Daptardar, Shahlo Seidmedova, Jay Stockwell, Shulana Kpabar

Abstract—This paper is the final product of a Topological Data Discovery project conducted in collaboration with a client sponsor to discover a new and insightful process to visualize supply and demand volatility network data. Initially, our research centered around topological methods, but the data that was made available to us led us to examine other potential methods that could answer our client sponsor's questions. In this paper we intend to examine and discuss our new and creative methods that deviate from our initial topological path. We will discuss in more detail the data that we leveraged and the challenges that we faced initially. We will cover the methodology that our team created and utilized to produce the final visualizations. We will also discuss and display our results, key insights, as well as issues and challenges that arose as we progressed through each project phase. Finally, this was a challenging, engrossing, and thought-provoking project for our team in that our basic assumptions surrounding topological visualizations were tested. Our research and collaborative efforts with our sponsor presented some new avenues for future research, one area in particular was using the client's data in a geospatial format. Our team learned a great deal during the course of this project, and we're excited to share our research and our visualizations.

Introduction

The abundance of data and information we have today has been one of the greatest contributors to what's called the Fourth Industrial Revolution. However, there is a downside to it - a lot of data remains uncovered, which deprives businesses of their opportunity to gain valuable insights about their data. Second Sight is a data accounting company, which started out as a research project, by Reuben Vandeventer. Second Sight has "created network listeners that monitor in real-time how data is created, changed and consumed inside of a corporate ecosystem at a very granular level." This means that "listeners are streaming thousands of data points a minute on data usage." The purpose of this project is to build upon the existing listeners and create a better way to analyze data without sabotaging analytical integrity. So far, the client has focused on topological data analysis and visualization (Fig. 1), but we have detoured from that method.

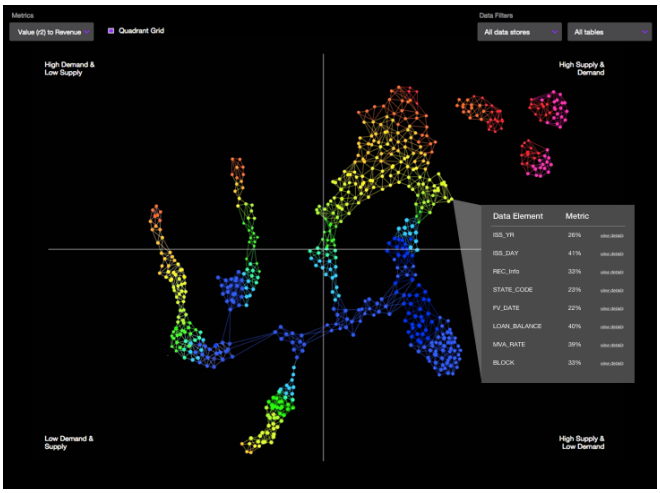

Fig. 1. Topological data analysis by the client

To protect highly sensitive data Second Sight tries to understand which data is in high supply and demand by asking the following questions:

- What data do we have right now?
- Which data is valuable?
- Which data is expensive?
- Which data is worthless?
- Which data is risky?

1 Sample Data and Statistics

Listeners create a supply and demand volatility index for each column of data within an organization or data cloud storage. Supply and demand volatility index are marked as "SVI" and "DVI". So, what does the basic structure of data look like? The table shown below (Fig. 2) contains 7 columns and has short descriptions to accompany them.

Column	Short Description
ID	Unique record ID
DataElementName	Name of the data element (column) from the source datastore being studied
DataStore_Name	Name of the datastore being studied
Table	Name of the table within the datastore that is being studied, and that the Data Element came from.
SVI	6-month trailing supply volatility
DVI	6-month trailing demand volatility
Adj R2	Correlation to monthly revenue

Fig. 2. Dataset Description

SVI indicates how often a given data element is modified, removed or changed. DVI indicates how often a given data element is viewed, moved, exported, copied or selected.

Our team received a small sample data set of 56 rows across 7 quarters from 2017 Q2 to 2018 Q4 by the client due to sensitivity and confidentiality of the data. In order to maximize the insight and results, we created a larger randomized dataset for the final visualization.

We have also generated random data using a python script to test the visualization dashboard. This data contains around 1500 Data Elements. The hierarchy of this data is explained below:

- 5 Data Stores with random number of tables between 1 to 15.
- Within each table, we have created random number of data elements between 1 to 50.

2 METHODOLOGY

For the primary dashboard, we have used Dash by Plotly which a python framework for building server-based analytics application. Other visualizations have been created in Tableau.

Since the number of points in the actual data with client are huge, we used K-Means clustering to group similar points together. We then found out the centroid of these points and represented those on the visualization as nodes. Each node can contain multiple Data Elements belonging to that cluster. This reduced the cluttering of plotted data points while not losing any information about the Data Elements themselves. Output primary visualization is a scatter plot which plots these nodes on SVI-DVI axes (X-axis – SVI, Y-axis – DVI). Adjusted R2, which later, we can replace with any required business indicator, colors the nodes in a divergent color scheme based on their values, red representing high correlation to revenue (higher Adj R2 value) and blue representing low correlation to revenue (lower Adj R2 value).

Another functionality we have added are two parallel coordinate graphs that show the data elements within the selected node across the time period. The values are represented as points which are again color-coded using the business indicator, in this case, adjusted R2. These are connected by lines to show the trends in data elements. Hovering over the nodes can display the data elements that are represented by the nodes on the SVI and DVI parallel coordinate plots. We also have also used a slider bar to display the nodes at different points of time.

For the visualizations in Tableau, we have used the data set that was provided by the client to create dashboards.

3 KEY INSIGHTS

Second Sight's main goal is to protect a company's highly sensitive data, as well as improve the profitability of various data points. As mentioned previously, we strived to ask ourselves the below questions and develop additional insights as we analyzed the results through visualization.

- What data do we have right now?
- Which data is valuable?
- Which data is expensive?
- Which data is worthless?
- Which data is risky?

3.1 Supply and Demand

What data do we have right now? These insights were discovered from the distribution of SVI and DVI. Although this a basic question, it is important for a company to know and understand the types of data being generated in their network. Through creating graphs of distributions of SVI, DVI and Adj R2 within a table, we were able to gain important insights and provided the client with a top down view of their data (Fig. 3).

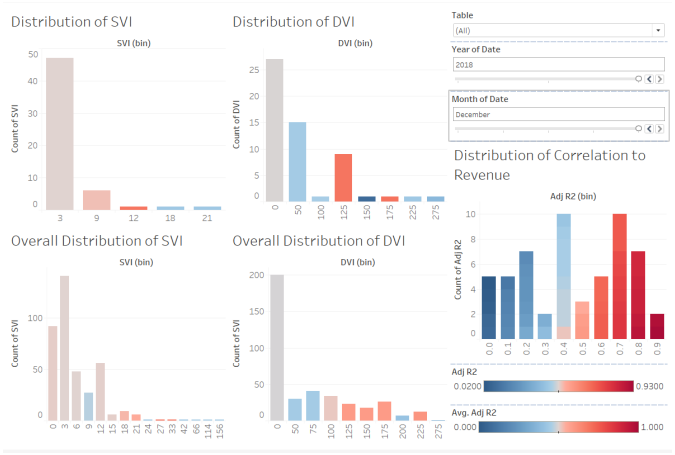

Fig. 3. Distribution of SVI, DVI and Adj R2

The top graphs represent the distribution of SVI and DVI for each quarter and the bottom graphs represent the overall distribution. Looking at this, we were able to quickly gather that the sample data had lots of data elements that were in low supply and low demand area. However, these were moderately related to the revenue. So, by increasing the supply and improving the accessibility to these data elements, the client will be able to generate more revenue.

Another insight we gained from this visualization is that, for certain data elements, increasing demand to very high levels also does not help in generation of revenue. As we can see that from Fig. 3, the overall demand for data elements that currently have 0 DVI can be increased to 125-175 DVI range, by improving methods of data accessibility. This range seems to be the most profitable for this data table.

Adding the time series to our evaluation, we can also see the fluctuations in Supply and Demand over a period. How is the evaluation of time information useful? Monitoring fluctuations can be very useful as it may give a company warning about potential source or destination issues. If a specific data source that was normally in high supply and demand, has a consistently low supply it would warrant an investigation as to why the supply is no longer there. Investigating this information could help to deter potential profit loss for those data elements with a high correlation to revenue.

3.2 Revenue Correlations

How can we find which data is worthless or valuable? Since a business's main goals are to make profits, we decided to focus our analysis on areas that could possibly hinder profits. Second Sight's implementation of a customized regression model gave us the R^2 correlation values to consider during the team's research. Considering the supply and demand of data elements, the Adjusted R^2 correlation to monthly revenue was an area where insight was gained.

In Fig. 3, we found that considering each data element's SVI and DVI against changes in the most important R^2 correlation values would give a company a more detail about a datastore's profitability. Looking at this data granularly over time, a company can gather information about a data element or datastore's Profit and Loss (P&L). They would also be able to model "what if" scenarios to determine future P&L for specific datastores. This information would be highly useful in segments of a company where determining the acquisition of resources must be validated. This information from several of the visualizations can provide proof that could not otherwise be gained from older statistical approaches.

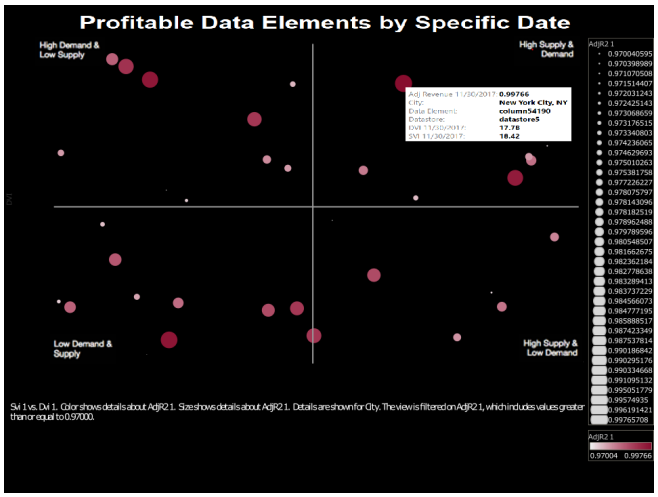

Fig. 4. Scatter plot of profitable data elements by specific date

3.3 Data Categories

Which data is expensive or risky for a company? In the below visualization (Fig. 5), we were able to add some additional categorizations to the client's quadrant grid from Fig. 1. Not only are we able to see the data that is in low and high supply/demand, but we were able translate this information into action steps for the client. In the below chart, we can gather far more insight as to how to handle these types of data when the R-squared correlation is considered. The company will be able to evaluate the expense and riskiness of specific datatypes and act toward their best interest.

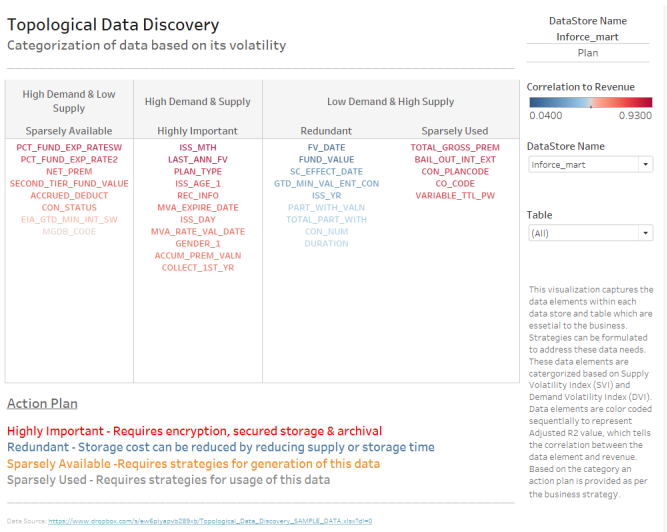

Fig. 5. Tabular categorization of data

Adding this type of categorization in summary format, also makes discovery for the client more simplified. The various insights a client could gain are summarized as follows:

- Highly Important - Requires encryption and secured storage & archival.
- Sparsely Used – Requires strategies for using this data
- Redundant – Storage cost can be reduced by reducing storage time/supply.
- Sparsely Available – Requires strategies to generate this data.
- Sparsely Used – Requires strategies for generation and usage of this data.

For the final visualization, we used the centroids as nodes, as discussed in the methodology section to represent points on a scatter plot of DVI-SVI axes as shown in the Fig. 6.

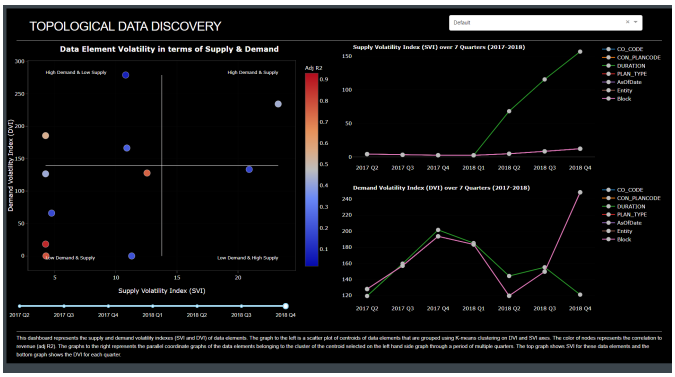

Fig. 6. Final visualization dashboard

With this synthesized and clustered topology of nodes representing data elements with color overlay of correlation to revenue, we can gain a lot of information about the data. The plot quadrants tell us what the demand and supply level of each node is. Based on this and its correlation to revenue, we can say how redundant or how valuable those elements are in terms of profitability for the company. Drilling down each node into its component data elements on the parallel coordinate graph tells us how the data point has been changing over time.

4 Discussion

There were a few opportunities discovered during the project, however, time and resources did not allow for these elements to be added to this project. The client already had a well-constructed algorithm for discovering the revenue correlation and wanted to keep this as their base. After working with our sample data, we soon began to see other areas that could be improved or added to their 2nd generation application.

4.1 Use of Geospatial Data

Second Sight could find it useful to add geographic coordinates to its listeners in order to see where the source data is coming from, as well as who is utilizing the datastore. With most companies having the ability to network across multiple locations, this could prove to be insightful. Knowing where the 'Highly Important' data is coming from and going to will allow the company to determine what locations need to be secured. Some companies have multiple datacenters, and know which servers hold the most important data is invaluable to a business. Putting the necessary security protocols in place for these specific servers can help protect against loss of revenue and data.

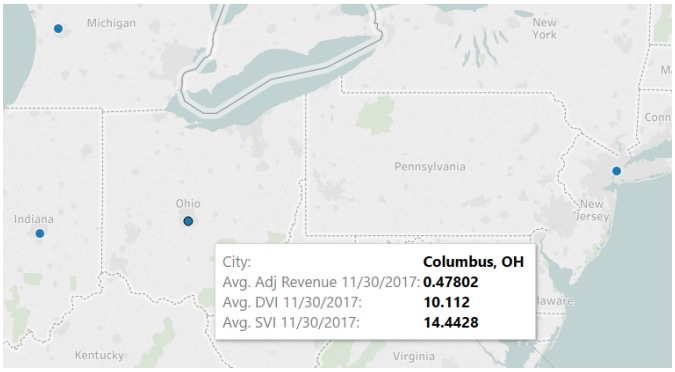

Fig. 7. Geospatial map of data source

4.2 Additional Correlations

With the availability and amount of historical data, Second Sight could also run other predictive models on the data. We can include other correlations to specific data elements and its importance in the network. With predictive models, our client can gain future insights into the volatility of data elements. Also, expanding on business indicators can help our client move data elements more robustly and accurately into desirable domains to help the business grow.

5 Conclusion

In conclusion, the project presented some challenges at the onset, but by shifting our approach, and employing some "out of the box" thinking, we were able to create a solid methodology and a series of visualizations to assist our client sponsor in meeting their visualization goals. Supply and Demand Volatility Index data is complex, and due to that complexity, we made the final determination that a topological approach to visualizing this data was not feasible. We were able to come up with a solution to effectively display fluctuations over a 6-month time period. Additionally, by leveraging R2 correlation data provided by the client, we were able to visualize the profitability of data elements over time.

Our success in this project was due in no small part to the strong partnership we developed working directly with our client sponsors at Second Sight. Their insights really influenced the direction of our work and research, and led to the development of the final visualizations that are presented in this paper. As mentioned earlier, the data that was provided to us presented some challenges in our original plan, but by working and consulting closely with Second Sight, we were able to devise an alternative route that proved to be successful in providing Second Sight with some valuable tools that they can leverage in their visualization journey going forward.

Topological visualizations can be powerful in answering the "what" questions in your data, and that was the initial plan at the beginning of this project. However, it was to our advantage to leverage a different approach to accomplish our final goal; to deliver a final product that proved to be valuable and beneficial to the client.

Acknowledgments

We would like to thank our clients - **Reuben Vandeventer** and **Robert Bob Stanton** for giving the opportunity to work on this industry related project and getting us acquainted with latest mathematical area of research – Topological Data Analysis.

We would also like to express deep gratitude towards the entire IVMOOC team – **Prof. Katy Borner, Andreas Bueckle, Michael Ginda** and **Yingnan Ju** for their constant support, and creating this amazing course by including these client projects.

References

[1] Link to the client presentation about the project https://zoom.us/recording/share/1zQj6XqpI7EmVk_PkmQI6z4fzC0_R4L9Ycs8NmM8NJWwIumekTziMw?startTime=1544137083000

[2] https://www.dropbox.com/s/6gk9wa7e44tgt1p/Imposing_Order_on_Data_Chaos.pdf?dl=0 (White paper given by the client)

[3] NSF Project III: Medium: Collaborative Research: Topological Data Analysis for Large Network Visualization. http://www.sci.utah.edu/~beiwang/networktdav/networktdav.html

[4] User Guide of Plotly https://dash.plot.ly/getting-started?_ga=2.165112063.1902689245.1555864419-893627516.1555543460