Data Visualization: Assignment 2

Adithya Nangarath
IMT2022024
IIIT-Bangalore
Bangalore, India
Adithya.Nangarath@iiitb.ac.in

Mallikarjun Chakoti
IMT2022116
IIIT-Bangalore
Bangalore, India
Mallikarjun.Chakoti@iiitb.ac.in

Ullas G
IMT2022125
IIIT-Bangalore
Bangalore, India
Ullas.G@iiitb.ac.in

***Abstract*—In this paper, we explore scientific and information visualization techniques applied to climate data and network analysis. Using color maps, contour, and quiver plots, we visualize climate patterns in the U.S. (August-October 2015) based on the GridMET dataset[3]. Additionally, we employ node-link diagrams, treemaps, and parallel coordinate plots to reveal insights from the arXiv Physics Paper Citation[2] and Customer Personality Analysis datasets[1].**

I. Introduction

Data visualization techniques are critical in interpreting complex datasets and extracting valuable insights across various domains. This study utilizes scientific visualization (SciViz) to analyze U.S. climate data from the GridMET dataset[3], specifically for August to October 2015. By applying color maps, contour plots, and quiver plots, we capture spatial and temporal variations in atmospheric variables, helping to highlight significant weather patterns and trends that may influence climate research and policy.

For information visualization (InfoViz), we examine two datasets: the arXiv Physics Paper Citation Network[2] and a Customer Personality Analysis dataset[1]. Nodelink diagrams reveal citation networks and influential papers in physics, while treemaps and parallel coordinate plots help identify patterns in customer purchasing behavior across demographics. These visualizations aid in understanding relationships and correlations within multi-dimensional data, supporting data-driven decisions in both academic and business contexts.

Together, these SciViz and InfoViz techniques demonstrate the power of visualization in making complex data accessible and insightful, with applications in research, industry, and policy-making.

II. Scientific Visualization

A. Quiver Plot

A quiver plot is a type of 2D plot used to represent vector fields in a visually intuitive way, particularly useful for displaying directional data in scientific and engineering contexts. Each arrow in a quiver plot represents a vector, with its direction and length indicating the vector's direction and magnitude, respectively.

1) Dataset: For this study, we use climate data from the GridMET dataset[3], focusing on the variables `vs` (wind speed) and `th` (wind direction) for August to October 2015 across the United States. The dataset is available in `.nc` (NetCDF) format and can be accessed from the GridMET website.

2) Data Processing: To handle the NetCDF files, we used Python's `xarray` library, which provides direct support for processing `.nc` files. To achieve a balanced plot density, we sampled the data every 2 degrees of latitude and longitude, creating a grid block for both `vs` and `th` variables. Each grid block contains the average values of `vs` and `th`, which we then used to create the quiver plot.

By reducing the data density through this averaging approach, we produced a quiver plot with a clear, interpretable representation of the vector field—neither too dense nor too sparse—allowing for effective visualization of spatial patterns in wind speed and direction across the region.

3) Objective: Here, we use quiver plots to display wind speed and direction across the U.S. for two sets of dates in September and October 2015. We used Plasma colormap to represent wind strength visually through arrow coloring. Plasma provides a perceptually uniform gradient, making it effective for distinguishing variations in wind strength. Its vibrant color range, from dark purple to bright yellow, highlights intensity differences clearly, enhancing readability and visual appeal in the plot.

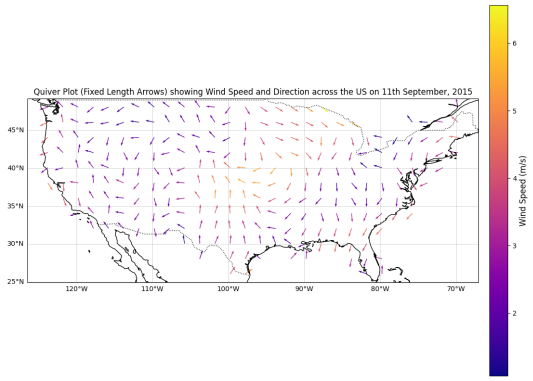

Fig. 1. Quiver plot(Fixed Length Arrow) describing wind speed and direction across the US on September 11, 2015.

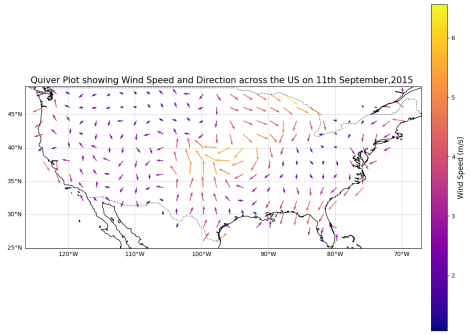

Fig. 2. Quiver plot describing wind speed and direction across the US on September 11, 2015.

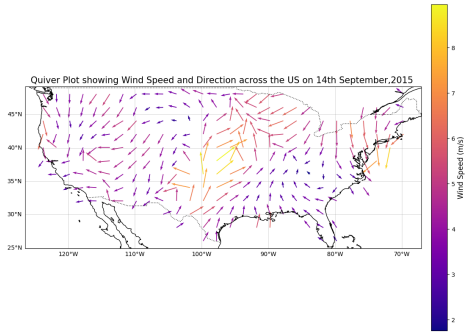

Fig. 3. Quiver plot describing wind speed and direction across the US on September 14, 2015.

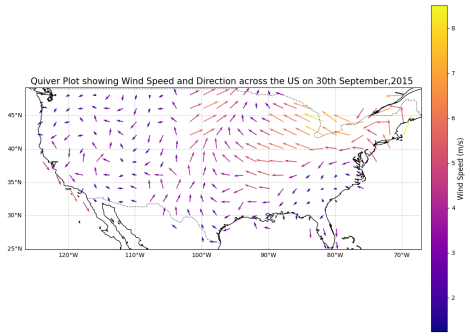

Fig. 4. Quiver plot describing wind speed and direction across the US on September 30, 2015.

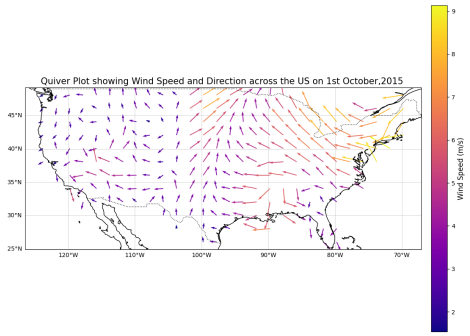

Fig. 5. Quiver plot describing wind speed and direction across the US on October 1, 2015.

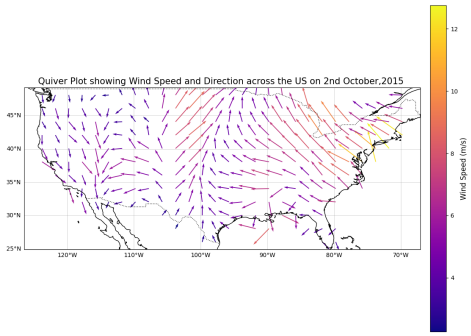

Fig. 6. Quiver plot describing wind speed and direction across the US on October 2, 2015.

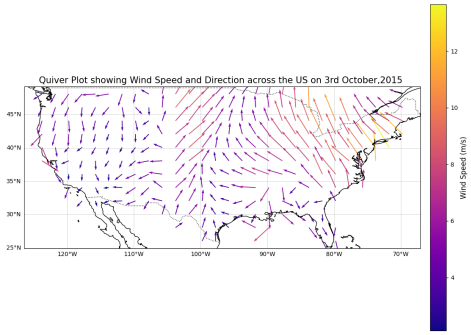

Fig. 7. Quiver plot describing wind speed and direction across the US on October 3, 2015.

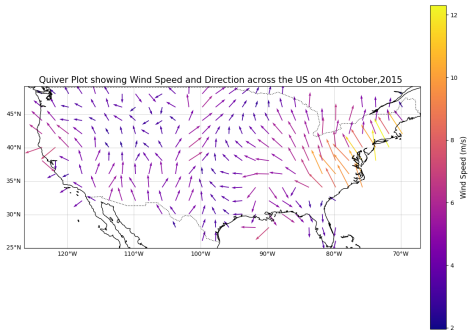

Fig. 8. Quiver plot describing wind speed and direction across the US on October 4, 2015.

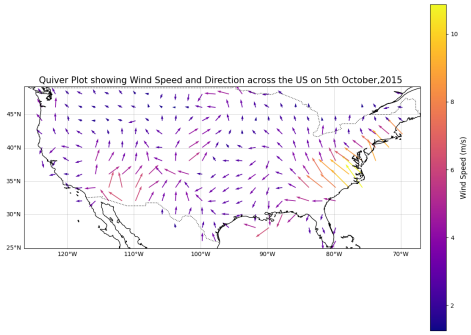

Fig. 9. Quiver plot describing wind speed and direction across the US on October 5, 2015.

a) First set of Dates:September 11 and September 14, 2015: Reason for Selection: On September 14, 2015, heavy rains caused flash flooding in southwestern Utah, tragically resulting in the deaths of 20 people. This event remains the deadliest natural disaster in Utah's history.

Analysis:

- September 11, 2015: I first examined the wind patterns just before the disaster. The quiver plot for this date reveals moderate wind speeds around the central U.S., where Utah lies.For 11th September 2015, I experimented with two quiver plots using the US GridMet dataset. The first plot (Fig. 1) uses fixed-size arrows, with color representing wind strength, making it neat but less intuitive. The second plot (Fig. 2) adjusts arrow size and color to represent wind speed, providing a better visual representation of the wind speed variations.Therefore, for subsequent plots of other dates, I opted to use magnitude-proportional vectors in the quiver plots, rather than fixed-size arrows, to enhance interpretability.
- September 14, 2015: The quiver plot for this date(Fig 3) shows increased wind speeds in the central part of the U.S., especially around Utah. This buildup may indicate shifts in atmospheric conditions preceding the flooding.

Observation:

- By comparing both dates, we notice a marked increase in wind activity near Utah, visible in the center of the map. This could suggest atmospheric pressure changes linked to the impending heavy rains.

b) Second Set of Dates: September 30 to October 5, 2015: Reason for Selection: In October 2015, an exceptional flooding event occurred in the Carolinas. From October 1-5, a stalled offshore front, combined with tropical moisture, led to historic rainfall levels, especially around Charleston. Flooding caused extensive property and road damage, requiring many emergency rescues.

Analysis:

- September 30 - October 5, 2015: Wind activity progressively intensified, peaking on October 3. The quiver plots show increased wind speeds particularly along the southeastern coast, where the Carolinas are located.

Observation:

- Wind speed and direction during this period highlight the conditions that likely contributed to the sustained, heavy rainfall, with the most intense winds around the affected coastal areas.

4) Inference from Quiver Plots: By observing these quiver plot animations (available in our codebase), we can see the shifts in wind flow patterns on selected dates associated with each natural event. These visualizations provide insight into how wind conditions correlate with weather events across different U.S. regions.

The changes in wind speed and direction, as depicted in the quiver plots, can help us identify potential atmospheric conditions leading to extreme weather events. For instance, shifts in wind patterns often indicate pressure changes that may precede severe storms or flooding, as seen in the case of southwestern Utah and the Carolinas.

B. Contour plot

A contour plot is a 2D representation that uses contour lines to show levels or regions of constant values within a dataset. It is particularly useful for visualizing spatial distributions and variations in scalar fields, such as temperature, humidity, or other atmospheric data. Each line or region represents a particular value, providing an easy-to-interpret visualization of gradients and patterns within the data.

1) Dataset: For this analysis, we utilize climate data from the GridMET dataset[3], specifically focusing on three variables: `srad` (solar radiation), `sph` (specific humidity), and `pet` (potential evapotranspiration) for the year 2015 across the United States. Like the previous dataset, this data is available in `.nc` (NetCDF) format, and we accessed it from the GridMET website.

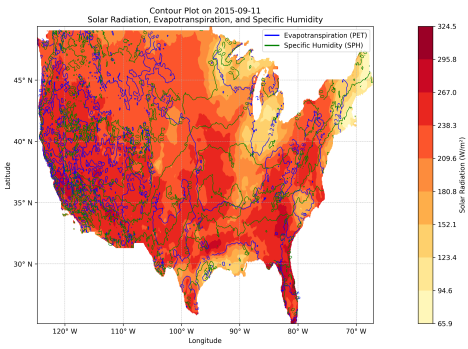

Fig. 10. Contour plot describing solar radiation, specific humidity, and potential evapotranspiration across the US on September 11, 2015.

2) Data Processing: Similar to the quiver plot, we used Python's `xarray` library to handle and process the NetCDF files. After extracting the `srad`, `sph`, and `pet` variables from the dataset[3], we performed downsampling to achieve an appropriate plot density.

To ensure that the contour plot remains clear and visually accessible, we created a 2-degree grid over the latitude and longitude dimensions. For each grid block, we computed the average values of `srad`, `sph`, and `pet`. This averaging approach provides a smoothed representation of the data, ensuring that the contours are neither too dense nor too sparse. Using these averaged values, we then generated the contour plot, which illustrates the spatial distribution of solar

radiation, humidity, and evapotranspiration across the study area.

This contour plot highlights regional climate characteristics, enabling the identification of areas with high or low solar radiation, humidity levels, and evapotranspiration rates.

3) Purpose of Contour Plots: Contour plots are typically used to represent continuous data on a 2D plane, especially when we need to visualize variations across a geographic region. In our case, we created contour plots for solar radiation and other related variables on the same dates as our quiver plots, enabling us to analyze these instances together for easier comparison and interpretation.

4) Observations from Specific Dates: September 11 and September 14, 2015

- On both dates, we observed a consistent pattern of solar radiation across the U.S.
- PET and SPH values around Utah showed increased wind and atmospheric moisture just before the flash flood on September 14, indicating heightened moisture demand in drier conditions.

October 1 to October 5, 2015:

- The PET contour density was significantly lower near South Carolina, corresponding with heavy rainfall reported in that region. The lower density suggests reduced moisture demand due to high rainfall.
- SPH contours showed elevated humidity, confirming the moist conditions in the region.

5) Plot Details: Primary Variable - SRAD (Solar Radiation):

- We used the "Contour Fill" method to plot SRAD (Surface Downwelling Solar Radiation), which provides a smooth and visually appealing representation of continuous data across regions.
- Solar radiation (SRAD) represents the amount of solar energy reaching the Earth's surface, impacting crop growth and weather patterns.
- By using contour fill for SRAD, we can clearly observe regions with varying solar radiation levels.

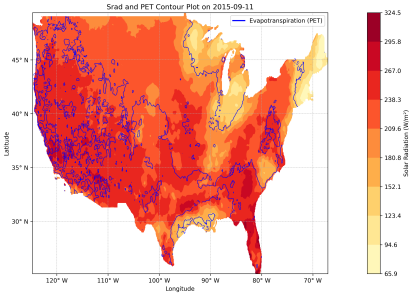

Fig. 11. Contour plot describing solar radiation and potential evapotranspiration across the US on September 11, 2015.

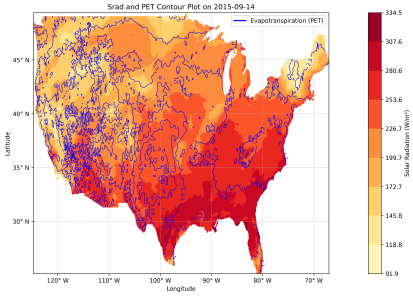

Fig. 12. Contour plot describing solar radiation and potential evapotranspiration across the US on September 14, 2015.

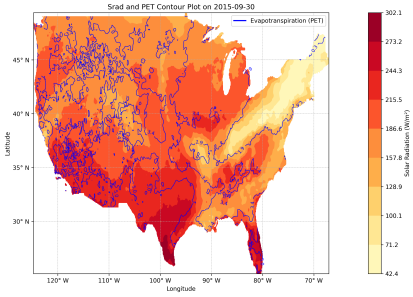

Fig. 13. Contour plot describing solar radiation and potential evapotranspiration across the US on September 30, 2015.

Reason for Contour Fill:

- Contour fill is ideal for SRAD because it shows gradual transitions between values, highlighting changes in solar radiation in a more intuitive way.
- Since SRAD is a continuous variable, contour fill effectively illustrates these smooth transitions without abrupt breaks, making it easier to visually interpret varying intensity levels across the region.

Overlay Variables - PET and SPH: To provide additional context on environmental conditions, we overlaid two more variables:

- **PET (Potential Evapotranspiration):** This measures the atmospheric demand for moisture based on evaporation and plant transpiration. Higher PET indicates areas where moisture demand is high, often due to higher temperatures.
- **SPH (Specific Humidity):** This variable reflects the actual amount of water vapor in the air, with higher SPH values representing more humid regions.

Reason for Using Marching Squares for PET and SPH:

- As we had already used contour fill for SRAD, adding contour fill again for PET and SPH would result in overlapping areas, reducing visual clarity and masking the SRAD contours.
- Thus, we opted to display PET and SPH using the marching squares algorithm, which outlines the

contour levels as distinct lines rather than filled areas, ensuring SRAD contours remain visible in the background.

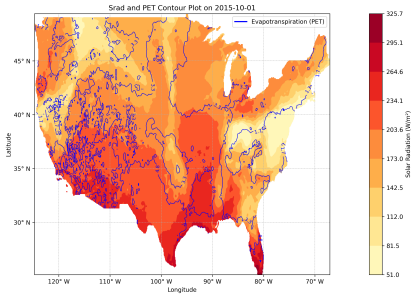

Fig. 14. Contour plot describing solar radiation and potential evapotranspiration across the US on October 1, 2015.

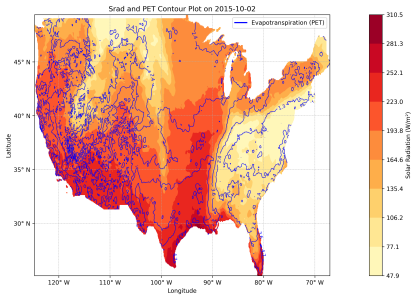

Fig. 15. Contour plot describing solar radiation and potential evapotranspiration across the US on October 2, 2015.

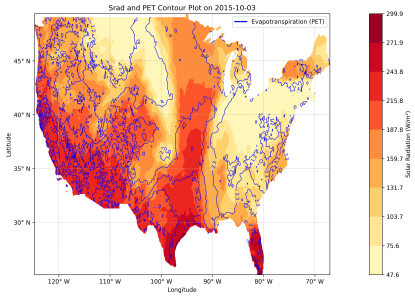

Fig. 16. Contour plot describing solar radiation and potential evapotranspiration across the US on October 3, 2015.

Levels Parameter Adjustment for PET and SPH:

- For PET and SPH, we adjusted the levels parameter to a lower value than usual.
- A higher levels parameter increases contour detail but also adds visual clutter, making it harder to interpret the plot.
- By choosing an optimal level, we maintained visual clarity without sacrificing much information, striking a balance between readability and accuracy.

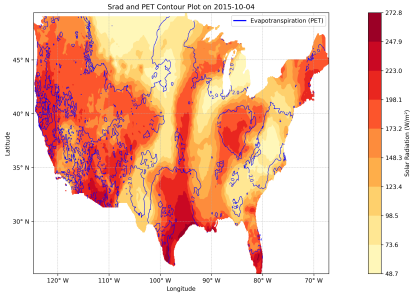

Fig. 17. Contour plot describing solar radiation and Potential evapotranspiration across the US on October 4, 2015.

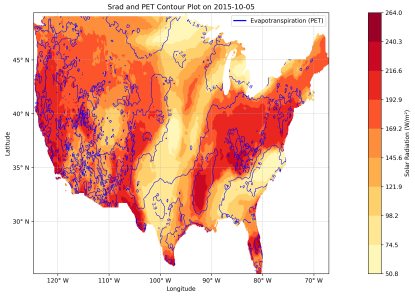

Fig. 18. Contour plot describing solar radiation and Potential evapotranspiration across the US on October 5, 2015.

Overlaying SRAD with PET and SPH Separately: As initially seen in (fig 10), we plotted both PET and SPH contour lines together on the same plot. However, this approach introduced too much visual clutter. To reduce this, we decided to plot SRAD with PET on one set of plots and SRAD with SPH on another.

- By overlaying SRAD with PET and SPH on separate plots, we can directly compare their spatial patterns.
- This approach provides a clearer and more focused visualization, helping us understand how these variables interact and influence each other.
- By carefully selecting contour levels, we ensure that both variables are clearly represented without sacrificing visual clarity.

6) Inference: The contour plots reveal that areas with high solar radiation and high PET may face crop stress due to increased water demand, particularly in drier regions. Conversely, areas with lower solar radiation and higher specific humidity could have more favorable conditions for crop growth, supported by higher atmospheric moisture. Integrating these observations with precipitation data could help identify regions at risk for drought stress and those with optimal moisture levels for agriculture.

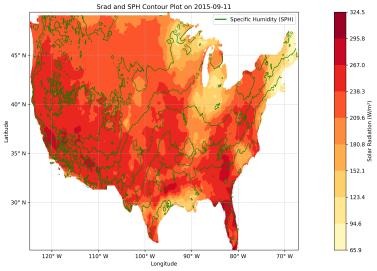

Fig. 19. Contour plot describing solar radiation and Specific Humidity across the US on September 11, 2015.

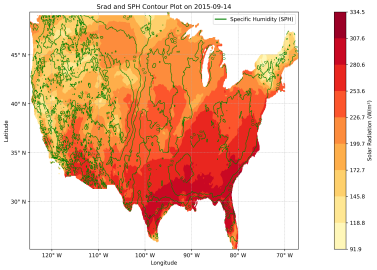

Fig. 20. Contour plot describing solar radiation and Specific Humidity across the US on September 14, 2015.

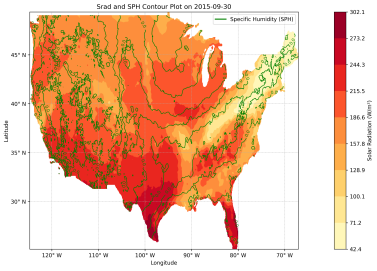

Fig. 21. Contour plot describing solar radiation and Specific Humidity across the US on September 30, 2015.

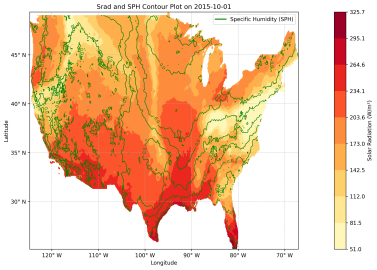

Fig. 22. Contour plot describing solar radiation and Specific Humidity across the US on October 1, 2015.

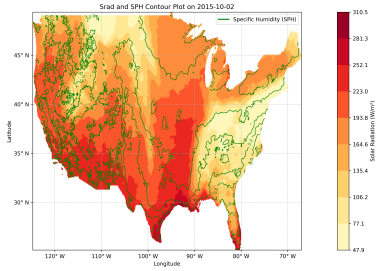

Fig. 23. Contour plot describing solar radiation and Specific Humidity across the US on October 2, 2015.

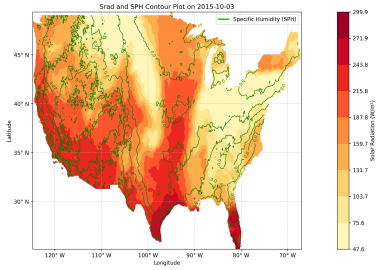

Fig. 24. Contour plot describing solar radiation and Specific Humidity across the US on October 3, 2015.

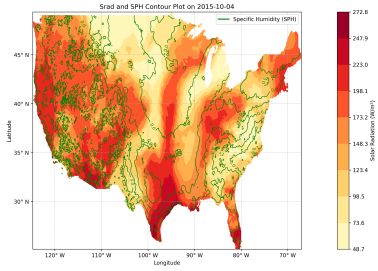

Fig. 25. Contour plot describing solar radiation and Specific Humidity across the US on October 4, 2015.

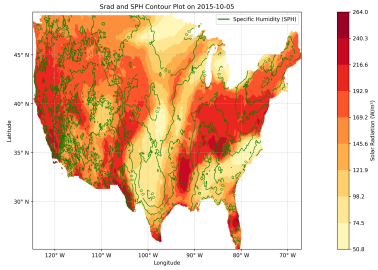

Fig. 26. Contour plot describing solar radiation and Specific Humidity across the US on October 5, 2015.

C. Color Mapping

Color mapping, also known as a colormap or color scale, is a visualization technique for displaying variations in scalar quantities—like temperature, pressure, or density—across a spatial domain. By assigning specific colors to scalar values, it visually represents data patterns and distributions within the field. Effective color choice and arrangement are crucial for clear and accurate interpretation of the scalar data.

1) Dataset: For this study, we use climate data from the GridMET dataset[3], focusing on the variables `bi` (burning index), `sph` (specific humidity), `fm100` (fuel moisture 100hr), and `fm1000` (fuel moisture 1000hr) for August to October 2015 across the United States. The dataset is available in `.nc` (NetCDF) format and can be accessed from the GridMET website.

2) Data Preprocessing: The data processing is performed by sampling the dataset[3] for a duration of every 10 days for each month, covering the period from August to September 2015. This approach allows for a detailed analysis of the climate variables over short time intervals, providing insights into the variability and trends of `bi` (burning index), `sph` (specific humidity), `fm100` (fuel moisture 100hr), and `fm1000` (fuel moisture 1000hr) across the United States. The dataset, available in `.nc` (NetCDF) format, is accessed from the GridMET website for further analysis.

3) Objective: In this study, we analyze wildfire-related climate variables in California and Washington for August and September 2015, focusing on conditions leading up to and during peak wildfire events. Specifically, we examine the following variables:

- **Burning Index (BI)**
- **Specific Humidity (SPH)**
- **Dead Fuel Moisture (100hr)**
- **Dead Fuel Moisture (1000hr)**
- **Composite Fire Index (CFI)**: A derived variable that combines the above variables using weighted contributions to represent an overall fire risk indicator.

To capture these dynamics, we utilize spatial maps with both global and local scaling approaches. The analysis for each variable is supported by sample visualizations from August 21, one of the days with the highest wildfire activity, providing a snapshot of conditions during peak fire events.

4) Visualization Approach: To effectively convey the spatial distribution of each variable, we apply distinct colormaps and scaling methods, chosen based on the variable's characteristics and expected influence on wildfire dynamics:

- **Burning Index (BI) fig27 and fig28:**
  - **Colormap:** Diverging coolwarm colormap.
  - **Scaling:** Linear scaling.
  - **Explanation:** The burning index represents potential fire intensity, making it critical to distinguish between high and low values clearly. The diverging coolwarm colormap aids in this by using contrasting colors at each end of the scale, with a smooth transition between. Linear scaling ensures a direct interpretation of values, where each unit increase reflects a proportional change in fire intensity.

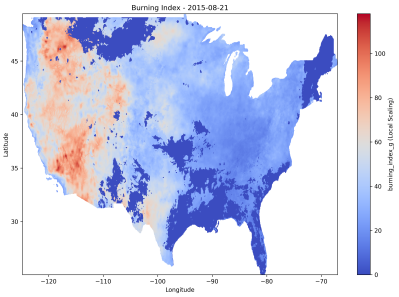

Fig. 27. Color Map describing Burning index across the US on August 21, 2015 Local scale

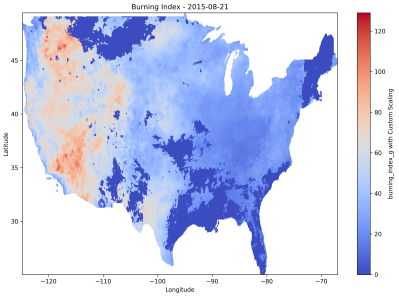

Fig. 28. Color Map describing Burning index across the US on August 21, 2015 Global scale.

- **Specific Humidity (SPH) fig29 and fig30:**
  - **Colormap:** Continuous Blues colormap.
  - **Scaling:** Logarithmic scaling.
  - **Explanation:** Specific humidity, which affects fire spread, has a wide dynamic range, with values spanning several orders of magnitude. Logarithmic scaling helps compress this range, making lower values more distinguishable on the map. This is particularly useful for observing subtle humidity changes that could influence fire risk. A continuous colormap ensures smooth visualization of these variations.
- **Dead Fuel Moisture 100hr (FM100) fig31 and fig32:**
  - **Colormap:** Diverging PiYG colormap.
  - **Scaling:** Linear scaling.
  - **Explanation:** Dead fuel moisture indicates the dryness of combustible material, affecting fire susceptibility. The continuous colormap represents moisture levels accurately, while

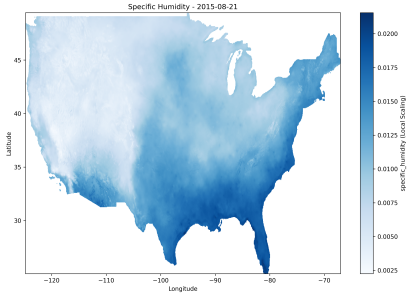

Fig. 29. Color Map describing Specific Humidity across the US on August 21, 2015 Local scale.

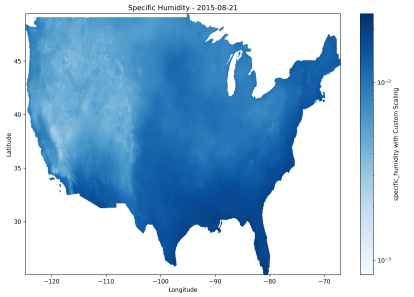

Fig. 30. Color Map describing Specific Humidity across the US on August 21, 2015 Global scale.

linear scaling allows direct interpretation of moisture content, where each increase in value reflects a proportional increase in fuel moisture, showing areas less susceptible to fire.

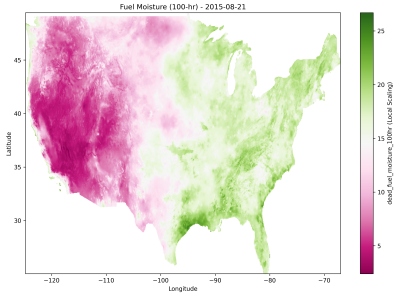

Fig. 31. Color Map describing Fuel Moisture(100-hr) across the US on August 21, 2015 Local scale.

- **Dead Fuel Moisture 1000hr (FM1000) fig33 and fig34:**
  - **Colormap:** Diverging PiYG colormap.
  - **Scaling:** Linear scaling.
  - **Explanation:** Similar to FM100, FM1000 represents moisture content over a longer time lag and responds to prolonged dry conditions. The continuous colormap with linear scaling enables direct comparison with

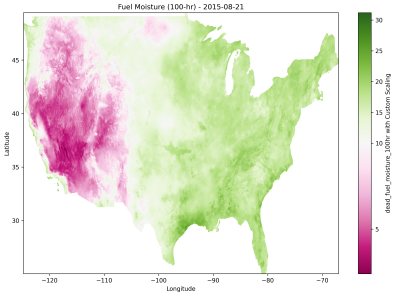

Fig. 32. Color Map describing Fuel Moisture(100-hr) across the US on August 21, 2015 Global scale.

FM100 maps, highlighting areas where fuel is critically dry.

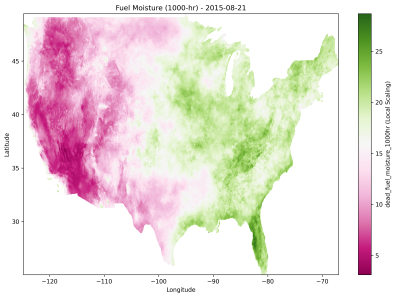

Fig. 33. Color Map describing Fuel Moisture(1000-hr) across the US on August 21, 2015 Local scale

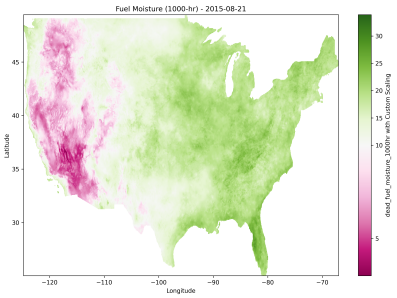

Fig. 34. Color Map describing Fuel Moisture(1000-hr) across the US on August 21, 2015 Global scale.

- **Composite Fire Index (CFI) fig35 and fig36:**
  - **Colormap:** Continuous viridis colormap.
  - **Scaling:** Logarithmic scaling.
  - **Explanation:** The Composite Fire Index is a derived metric that combines weighted contributions from BI, SPH, FM100, and FM1000, providing an integrated fire risk assessment. Logarithmic scaling is applied to manage the broad range of values, enhancing visualization of both lower and higher fire risk zones. A continuous colormap aids in

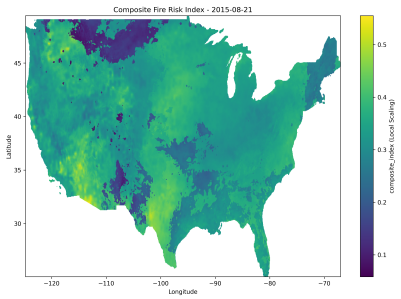

Fig. 35. Color Map describing Composite Fire Risk Index across the US on August 21, 2015 Local scale.

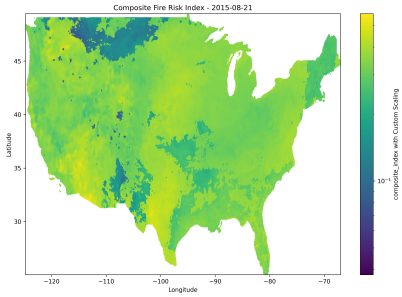

Fig. 36. Color Map describing Composite Fire Risk Index across the US on August 21, 2015 Global scale.

smooth representation across the range, helping to clearly identify areas of concern.

5) Inference: The analysis of wildfire-related variables in California and Washington provides insight into the conditions contributing to wildfire susceptibility in these regions. Across all plots, the following patterns emerge:

- **High Burning Index (BI):** The California and Washington regions display consistently high burning index values, indicating a significant potential for fire intensity. This high BI suggests that vegetation in these areas is primed for ignition, further increasing the likelihood of wildfires.
- **Low Specific Humidity (SPH):** Specific humidity values are notably low in these regions, contributing to dry atmospheric conditions. Low humidity reduces moisture in vegetation and fuels, creating an environment where fires can ignite and spread more easily.
- **Low Dead Fuel Moisture (FM100 and FM1000):** The 100-hour and 1000-hour dead fuel moisture levels are also low in California and Washington. These low values indicate that larger dead fuel components, which respond more slowly to changes in moisture, are dry and highly flammable, sustaining fires for longer durations and intensifying fire behavior.
- **High Composite Fire Index (CFI):** The Composite Fire Index, calculated using weighted contributions of the above variables, is elevated across these regions, providing a combined measure of high wildfire risk. This elevated CFI highlights how the interaction of high burning index, low specific humidity, and low fuel moisture levels contributes to the advent and spread of wildfires in California and Washington.

Overall, these observations suggest that the combination of high burning potential, dry atmospheric conditions, and low fuel moisture content creates a critical environment for wildfire outbreaks in these regions. The consistent patterns across all variables reinforce the susceptibility of California and Washington to wildfires during the analyzed period.

III. Information Visualization

A. Tree Map

1) Dataset: For this study, we chose the **Customer Personality Analysis** dataset from Kaggle[1], which provides detailed customer data for analyzing demographic characteristics, purchasing behavior, and interactions with marketing campaigns. The dataset consists of 2240 entries and 29 variables, grouped as follows:

- **Demographics:** Information on age, education, marital status, income, and family composition.
- **Purchasing Behavior:** Total spending on various product categories and usage of purchase channels.
- **Campaign Response:** Indicators of customer engagement with marketing campaigns.

2) Data Processing: In this study, the data processing phase required minimal handling, as the dataset contained very few missing values. These were easily addressed through imputation. Additionally, some features were split or restructured to simplify the analysis, particularly where multiple variables were combined or cluttered. This preprocessing step helped ensure a clean dataset, facilitating a more accurate analysis.

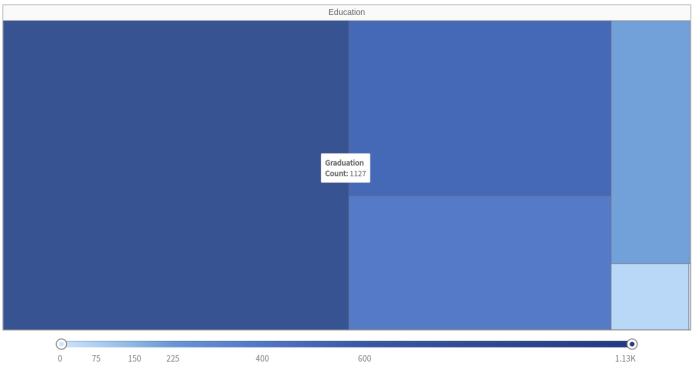

Fig. 37. Squarified plot Treemap of Education Levels, Showing Distribution in Non-Hierarchical way .

3) Implementation and Experiment:

a) Objective:: The objective of this experiment was to analyze the distribution of customer education levels and their marital statuses using hierarchical and non-hierarchical treemap visualizations. We aimed to determine the effectiveness of various treemap layouts (Squarified, Slice and Dice, and Hierarchical) in representing customer data in a visually interpretable format.

b) Methodology:: The experiment consisted of generating both hierarchical and non-hierarchical treemaps with the following steps:

- **Data Processing:** We used customer data to compute the frequency of each education level and marital status combination.
- **Treemap Layouts:** Multiple treemap configurations were implemented using FusionCharts:
  - **Squarified Treemap:** A square layout that minimizes visual clutter and allows easy comparison of education level frequencies.
  - **Slice and Dice (Horizontal, Vertical, and Alternate):** Linear and alternating layouts presenting education levels in a sequence to emphasize ordered categories.
  - **Hierarchical Treemap:** A nested layout that represents education levels as primary categories, with marital status as subcategories.
- **Color Gradient:** A color range from light to dark blue was applied to all layouts, where darker shades represent higher frequencies, making high-count categories instantly recognizable.

c) Observations::

- **Non-Hierarchical Treemaps:**
  - The **Squarified Treemap** provided a balanced representation of education levels, allowing efficient comparisons across categories with minimal visual complexity.
  - **Slice and Dice Horizontal/Vertical** layouts presented education levels in a structured format but required additional inspection to compare minor frequency differences.
  - The **Slice and Dice Alternate** layout enhanced visual interest but was less straightforward for quickly interpreting relative values.
- **Hierarchical Treemap:** This layout effectively visualized both education and marital status data. The nesting of marital statuses within each education level allowed for a clear view of subcategory distributions, making it useful for analyzing secondary relationships. However, the hierarchy added visual complexity that may not be necessary for quick overviews.
- **Color Gradient Analysis:** Across all layouts, the color gradient immediately highlighted categories with high counts. This feature was especially beneficial in the hierarchical layout, where nested groups could still be identified by intensity.

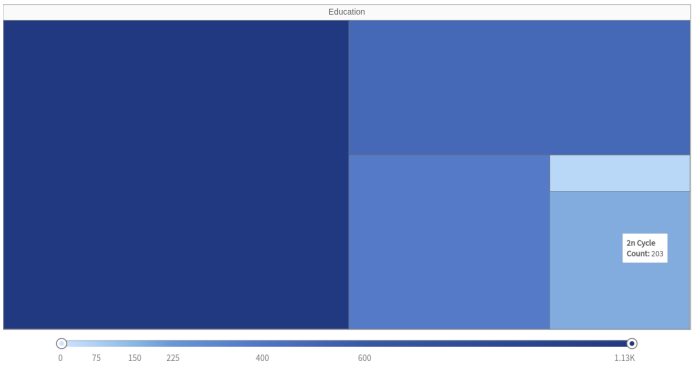

Fig. 38. Slice and Dice Alternate Treemap of Education Levels, Showing Distribution in Non-Hierarchical way

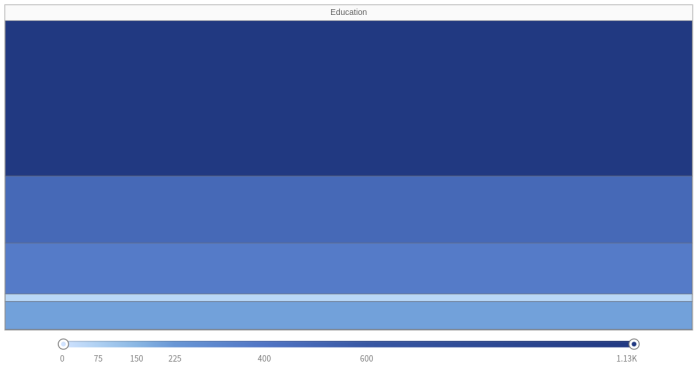

Fig. 39. Slice and Dice Horizontal Treemap of Education Levels, Showing Distribution in Non-Hierarchical way

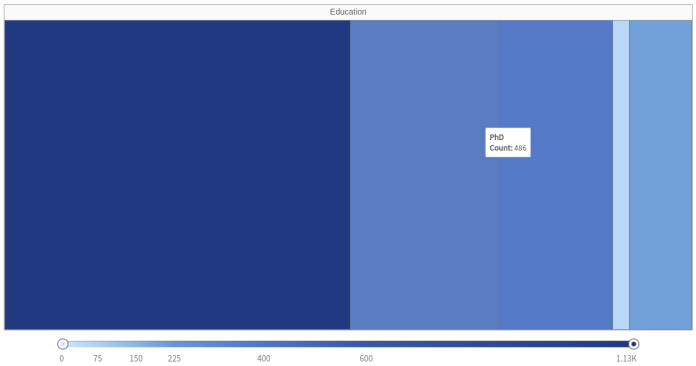

Fig. 40. Slice and Dice Vertical Treemap of Education Levels, Showing Distribution in Non-Hierarchical way

4) Inferences:

- **Overall Distribution:** The treemaps reveal the distribution of customer education levels and marital statuses, highlighting predominant categories and assisting in identifying potential target groups.
- **Preferred Layout:** For a straightforward view of education distribution, the Squarified layout is preferable. However, the hierarchical layout offers additional insights into marital status within each education group, making it suitable for in-depth analysis.

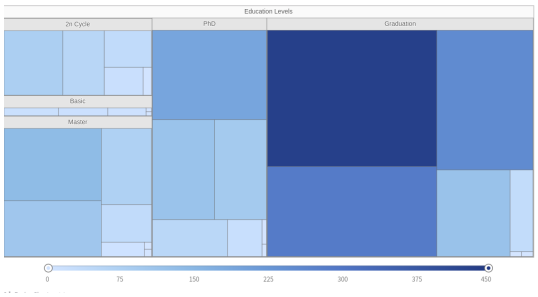

Fig. 41. Squarified Treemap Representing the Hierarchical Distribution of Education Levels and Marital Status

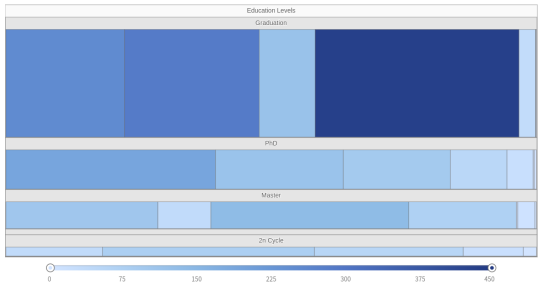

Fig. 42. Slice and Dice Horizontal Treemap Representing the Hierarchical Distribution of Education Levels and Marital Status

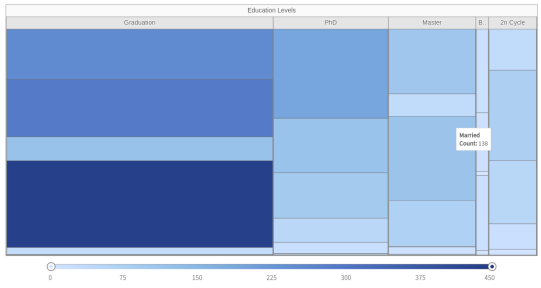

Fig. 43. Slice and Dice Vertical Treemap Representing the Hierarchical Distribution of Education Levels and Marital Status

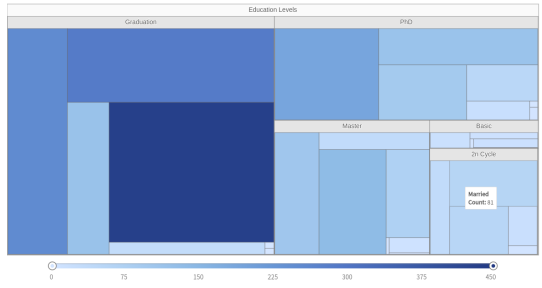

Fig. 44. Slice and Dice Alternate Treemap Representing the Hierarchical Distribution of Education Levels and Marital Status

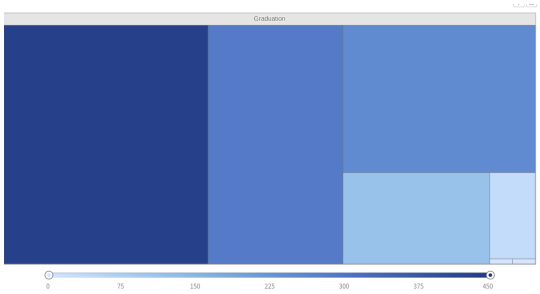

Fig. 45. Squarified plot Treemap User interaction on clicking Graduation

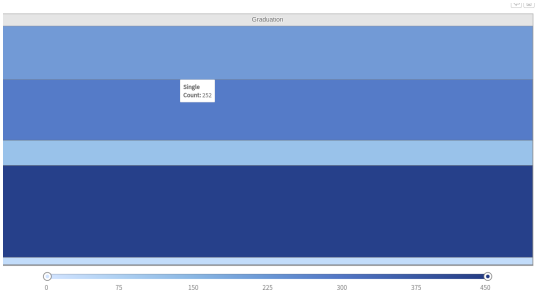

Fig. 46. Slice and Dice Horizontal Treemap User interaction on clicking Graduation

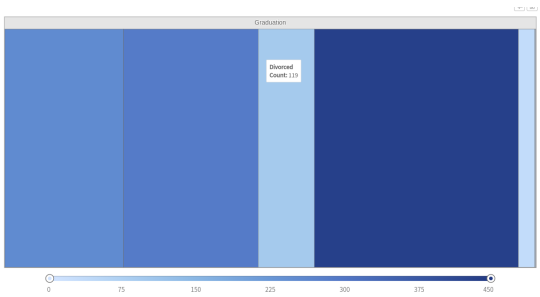

Fig. 47. Slice and Dice Vertical Treemap User interaction on clicking Graduation

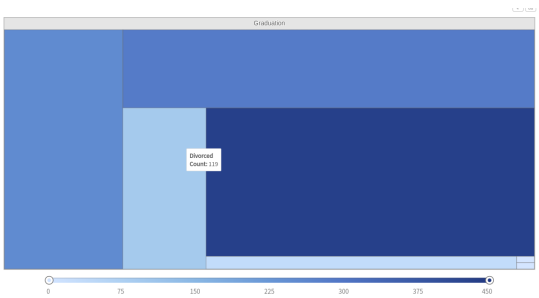

Fig. 48. Slice and Dice Alternate Hierarchical way

B. Parallel Coordinate Plot

1) Dataset: For this study, we chose the **Customer Personality Analysis** dataset from Kaggle[1], which provides detailed customer data for analyzing demographic characteristics, purchasing behavior, and interactions with marketing campaigns. The dataset consists of 2240 entries and 29 variables, grouped as follows:

- **Demographics:** Information on age, education, marital status, income, and family composition.
- **Purchasing Behavior:** Total spending on various product categories and usage of purchase channels.
- **Campaign Response:** Indicators of customer engagement with marketing campaigns.

2) Methodology: This project explores Parallel Coordinates Plots (PCP) to visualize customer spending patterns across different variables in a dataset. I created three distinct PCP plots using subsets of variables, experimenting with factors like sample size and sampling methods. By adjusting sample sizes, I aimed to assess whether larger samples with more lines could still provide clear insights. I also experimented with sampling types—such as selecting equal samples from each education level—to see if different sample compositions improved perceptibility. Additional techniques included brushing, to highlight specific data ranges, and axis reordering, to reduce visual clutter.

3) Experiments: This section presents three key experiments conducted on customer data.

a) Customer Product Preferences by Income, Spending, and Education

Preprocessing:

The key variables in this section include:

- **Income**: Customer's annual household income.
- **Education**: Customer's level of education.
- **Product Spending**: Total amounts spent in the last two years on different categories, including wine, fruits, meat, fish, sweets, and gold products.

Initially, plotting these variables directly led to heavy visual clutter. To manage this, I preprocessed the data using Python's Pandas library, calculating statistical values like the median, quantiles, minimum, and maximum for numerical variables (Income and product spending). This allowed me to group values into bins(in Javascript), reducing clutter and making patterns more visible.

Observations:

- **Sample Selection and Coloring by Education:** I selected a sample of 100 data points, consisting of 20 samples from each education category, with each education level represented by a unique color. While a legend was omitted, the education axis provides a clear reference for identifying each category's color. The code allows flexibility in sample size, so the user can adjust it to explore different sample sizes and their impact on readability.(Fig 49) represents Pcp for this experiment.

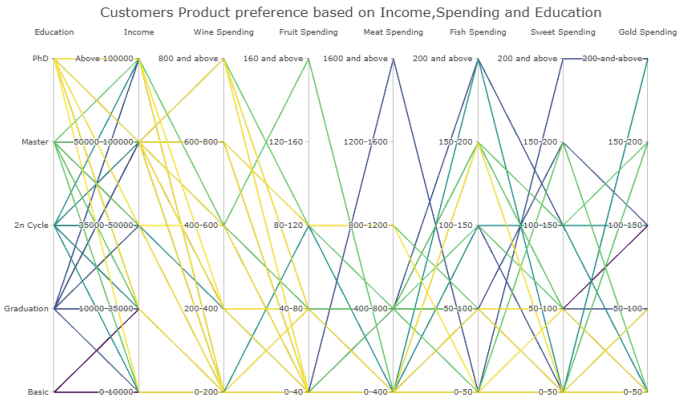

Fig. 49. Parallel Coordinate Plot for Experiment 1.

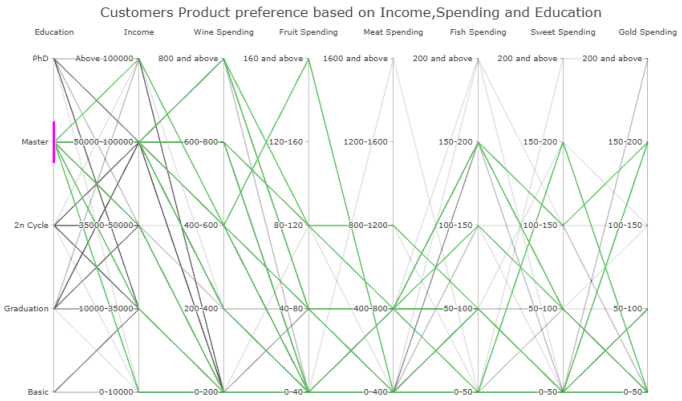

Fig. 50. Parallel Coordinate Plot for Experiment 1, Highlighting Brushing on Education axis.

- **Brushing:** By brushing on the education axis(Fig 50), we can isolate a specific category—such as 'Masters'—and observe spending patterns across various product types. Increasing the sample size and brushing on the income axis(Fig 51) reveals spending tendencies within a specific income range, helping us see how particular income groups allocate spending across product categories.

Inference: The PCP plots are a helpful way to see how different customer groups spend on products. By organizing data into bins and adjusting sample sizes, we reduce clutter and make patterns clearer. Using color for education levels and brushing for specific income or education groups helps us see how these groups spend on different products. This makes it eas-

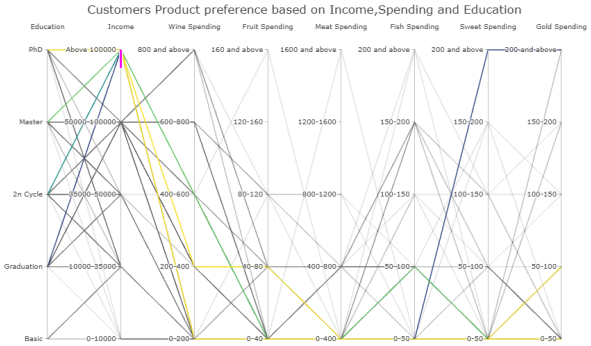

Fig. 51. Parallel Coordinate Plot for Experiment 1, Highlighting Brushing on Income axis.

Fig. 52. Parallel Coordinate Plot for Experiment 2.

ier to spot trends and understand customer preferences based on income and education.

b) Customer Loyalty and Recency of Purchases

Preprocessing: The variables used include:

- **Recency**: Number of days since the customer's last purchase.
- **NumDealsPurchases**: Number of purchases made with a discount.
- **AcceptedCmp1 - AcceptedCmp5**: Whether a customer accepted the offer in each of the 5 campaigns (1 = accepted, 0 = not accepted).

Additionally, I created a new feature **Total Accepted** to represent the total number of campaigns each customer accepted (sum of all accepted offers).

Observations:

- **Color-Coding:** The lines were color-coded based on the total number of accepted campaigns(Fig 52), using 6 colors (from the Viridis color map). This helps identify trends, such as: High recency and low deal purchases may indicate inactive or low-interest customers. Low recency and high deal purchases suggest interested and engaged customers.
- **Brushing:** Brushing on recency(Fig 53) allows us to filter customers based on recent activity. By doing this, we can explore which campaigns were more successful for customers with higher or lower recency. This was plotted with 40 lines, but the number of samples can be adjusted based on the analysis needs.

Inference: This PCP plot helps identify customer loyalty by showing how recency and deal purchases relate to campaign responses. It reveals active vs. inactive customers and highlights which campaigns were more successful. This makes it useful for targeting future campaigns and re-engaging less active customers.

c) Customer's Preferred Purchase Method

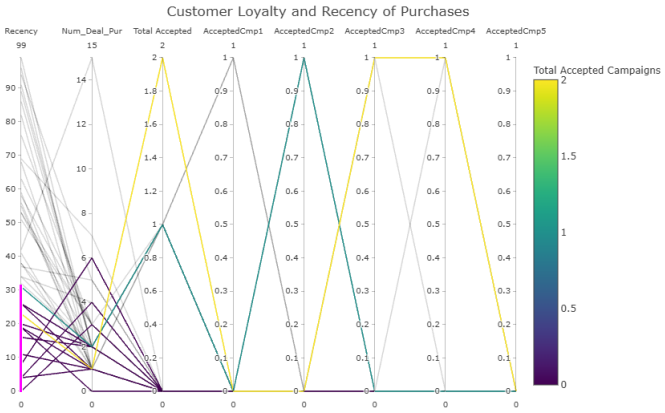

Fig. 53. Parallel Coordinate Plot for Experiment 2, Highlighting Brushing on Recency axis.

Preprocessing: Variables used: **Year of Birth**, **Marital Status**, **Income Category**, **Kids at Home**, **Teens at Home**, **Store Purchases**, **Web Purchases**, and **Catalog Purchases**. For the income variable, bins were created similar to the first PCP plot to reduce visual clutter.

Observations:

- Initially, I color-coded the lines by Year of Birth(Fig 54) using the Viridis color map, where lighter colors represented younger customers and darker ones older customers. However, this wasn't very informative. So, I switched to color coding based on Marital Status(Fig 55 and Fig 56), using five distinct colors, which provided more clarity. I also reordered the income axis to reduce clutter.
- **Brushing:** Brushing on specific Year of Birth ranges helps us identify preferences of age groups. For example, brushing on 1950(Fig 57) shows that people from this range prefer store purchases over web or catalog purchases. Similarly, brushing on specific marital status cate-

gories(Fig 58) highlights how different marital groups prefer different purchase methods (store, web, or catalog).

Inference: This PCP plot reveals how customers' preferred purchase methods vary by factors like age, marital status, and income. By color-coding lines based on marital status and brushing on specific categories like year of birth or marital status, we can easily identify patterns, such as which groups prefer store purchases over web or catalog purchases.

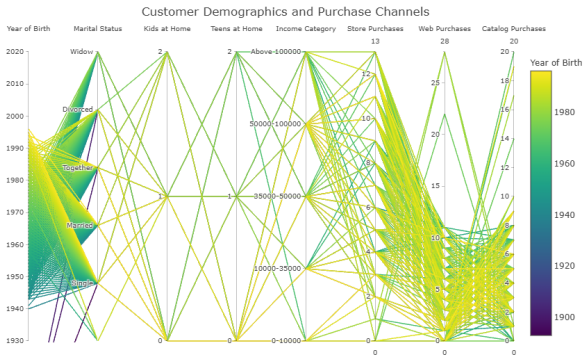

Fig. 54. Parallel Coordinate Plot for Experiment 3, Color-Coded by Year of Birth.

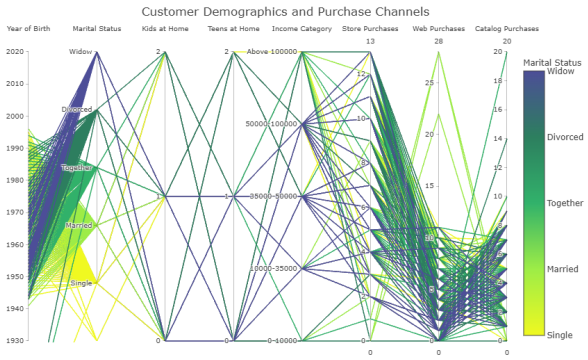

Fig. 55. Parallel Coordinate Plot for Experiment 3, Color-Coded by Marital Status with more samples(lines).

4) Final Inference:: Parallel Coordinate Plots (PCP) enabled effective visualization of customer data across multiple dimensions, providing valuable insights in each experiment. PCP illustrated spending preferences by income, education, and product types, highlighting the impact of education on spending patterns. It was also useful for analyzing customer loyalty, where recency and deal purchases helped differentiate active from inactive customers. Lastly, PCP allowed us to explore purchase preferences by age, marital status, and income level. PCP's flexibility in adjusting sample sizes proved beneficial: larger samples provided a broad overview, while smaller samples focused on specific groups. However, PCP has limitations—plots can become cluttered with too many lines, and overlapping lines can obscure details. So we have to pre-process data carefully based on our specific requirements and use interactive brushing to improve clarity.

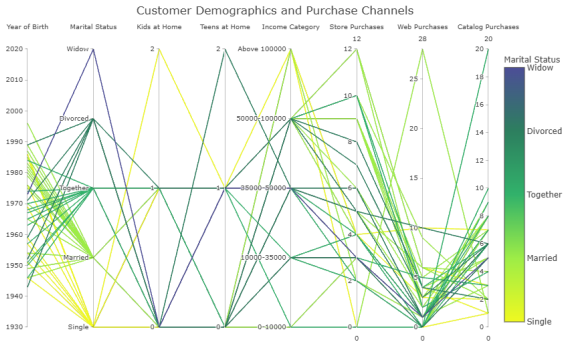

Fig. 56. Parallel Coordinate Plot for Experiment 3, Color-Coded by Marital Status with less samples(lines)

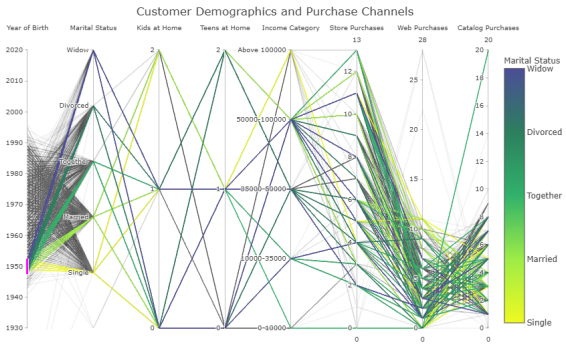

Fig. 57. Parallel Coordinate Plot for Experiment 3, Color-Coded by Marital Status, Highlighting Brushing on Year of birth.

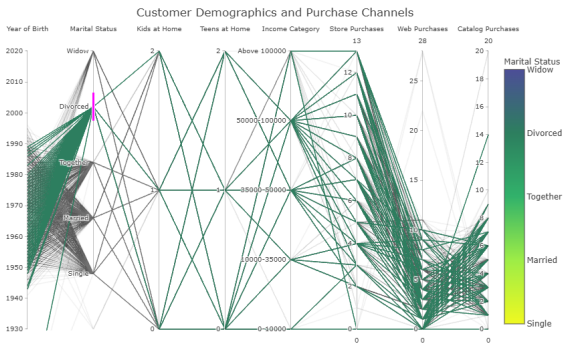

Fig. 58. Parallel Coordinate Plot for Experiment 3,Color-Coded by Marital Status, Highlighting Brushing on Marital Status.

C. Node Link Diagram

1) Dataset: The dataset used in this study is the arXiv High Energy Physics (HEP) citation network[32], which consists of two main text files: a citation network file and an abstract file. The citation network file represents citation relationships between papers, where each line records a source-target relationship, indicating that one paper cites another. In this directed network, nodes correspond to research papers, and directed edges denote citations, capturing the structure of influence within the scientific community. The abstract file contains detailed metadata for each paper, including a unique identifier, title, authors, abstract, and additional descriptors. This metadata enriches the dataset by providing context for each paper, allowing for content-based analyses such as topic modeling and clustering. Together, these files enable the study of citation dynamics in high-energy physics, offering insights into research trends, influential works, and collaboration patterns within the field.

2) Data Processing: The preprocessing phase begins with converting the raw text files of the arXiv High Energy Physics (HEP) citation network dataset[2] into a structured CSV format for easier handling and analysis. This conversion involves parsing the citation network file, where each source-target citation relationship is reformatted into a tabular structure with two columns, `Source` and `Target`, and then organized into a CSV file. During this conversion, unnecessary content, such as extraneous metadata or irrelevant formatting characters, is removed to ensure that only pertinent information is retained. This cleaning step reduces noise and prepares the dataset for downstream tasks, such as network analysis and natural language processing on the abstracts. Overall, these preprocessing steps ensure that the data is in a consistent and accessible format, facilitating efficient analysis and model training.

3) Analysis 1: To facilitate community detection within the arXiv High Energy Physics (HEP) citation network, we first filtered the dataset to include only nodes with a degree range from 206 to 2468. This degree threshold allowed us to focus on highly connected papers, which are central to understanding citation structures and influential research clusters.

a) Modularity Class for Community Detection: For community detection and grouping, we used the modularity class, a method that maximizes the density of edges within communities compared to edges between different communities. The modularity class assigns each node to a community based on its connections, effectively grouping papers with stronger internal connections. This allows for the identification of distinct research communities within the citation network, each represented by a different color.

b) Size Ranking with Betweenness Centrality: To rank nodes by size, we applied *betweenness centrality*, which measures the frequency with which a node appears on the shortest path between other nodes. Nodes with higher betweenness centrality are considered more influential as they bridge multiple communities or play a key role in connecting different parts of the network. This metric was instrumental in identifying significant nodes within each community.

Fig. 59. Node-link diagram using the Forceatlas 2 algorithm

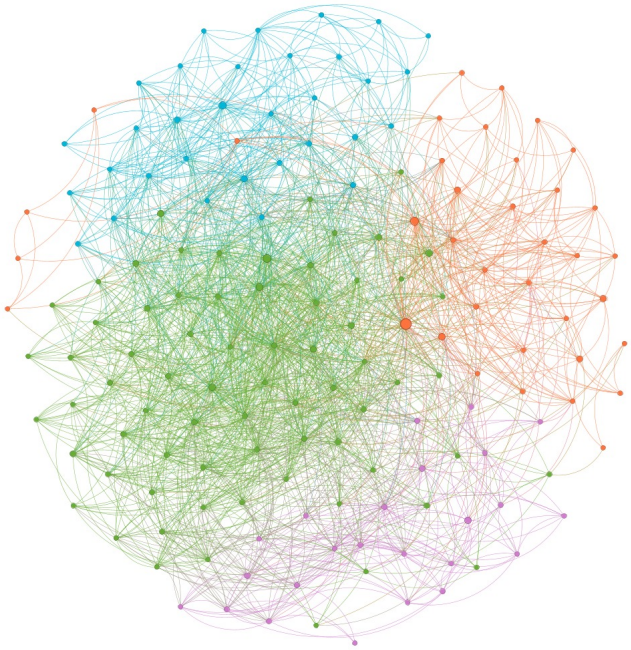

Fig. 60. Node-link diagram using the Fruchterman-Reingold algorithm.

c) Layout Algorithms: For visualizing the citation network, we experimented with the following layout algorithms:

- **Fruchterman-Reingold Layout(fig 60):** This algorithm provided a clear structure with minimal clutter, making it easier to distinguish between communities.
- **Force Atlas2 Layout(fig 59):** Although slightly more cluttered than Fruchterman-Reingold, this layout produced a well-organized representation of the network.
- **Yifan Hu Proportional Layout(fig 61):** While effective for certain configurations, this layout resulted in more overlap and was less effective at visually distinguishing communities than the other layouts.

Based on these comparisons, the Fruchterman-Reingold layout was found to offer the best visualization quality, followed by Force Atlas2 and Yifan Hu Proportional.

d) Research Communities and Group Descriptions: Through community detection, we identified four main research groups in the network, each representing a distinct area of focus in high-energy physics. The groups were assigned the following color codes and thematic descriptions:

- **Group 1 (Blue): Noncommutative Dynamics and Tachyon Physics in String Theory**
  - This group investigates the role of noncommutative geometry, tachyon condensation, and brane dynamics in string theory.
  - Topics include mathematical structures such as K-theory and physical phenomena like non-BPS states and tachyon condensation, particularly in the context of D-branes.
  - Matrix models are utilized to describe M-theory and noncommutative field dynamics, contributing to understanding string and brane interactions.
- **Group 2 (Orange): Holography, AdS/CFT Correspondence, and Large N Gauge Theories**
  - Papers in this group focus on the AdS/CFT correspondence, a duality linking anti-de Sitter (AdS) space with conformal field theories (CFT).
  - Key themes include holography, black hole thermodynamics, renormalization group flow, and large N gauge theories.
  - This research has implications for quantum gravity and gauge duality, exploring supergravity in AdS space, holographic bounds, and the confinement/deconfinement transition in gauge theories.
- **Group 3 (Green): Brane Dynamics, M-Theory, and Compactifications in String Theory**
  - This group focuses on brane dynamics within string and M-theory, covering topics like the Born-Infeld action, black hole microstates, and compactification methods such as F-theory on Calabi-Yau spaces.
  - Topics include D-brane physics, supergravity solutions, and U-duality, all essential for understanding low-energy gauge theories.
  - Research in this group contributes to the theoretical understanding of how branes impact the structure of space-time.
- **Group 4 (Purple): Supersymmetric Gauge Theories, Integrable Systems, and Non-Perturbative QFT**
  - This group centers on supersymmetric gauge theories, integrable systems, and non-perturbative quantum field theory.
  - Key topics include $N = 2$ supersymmetric Yang-Mills theories, Seiberg-Witten theory, and electric-magnetic duality.
  - Papers in this group explore exact solutions in gauge theories, connecting field theory results to string theory through mathematical structures like BPS states and dualities.

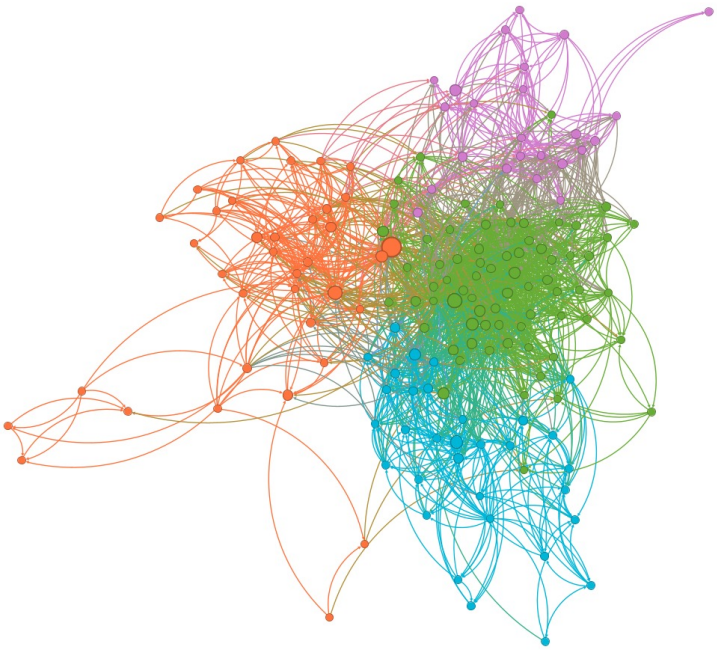

Fig. 61. Node-link diagram using the Yifan Hu proportional algorithm.

4) Analysis 2:

a) Inferred Class for Community Detection: An *inferred class* represents a partition of nodes within the network based on statistically detected communities. By applying community detection algorithms, we assign each node to a specific class, often revealing groups of papers with shared research themes or overlapping references. These inferred classes are then visualized using distinct colors, allowing for an intuitive visual partitioning of thematic clusters within the

network. In our case, we observed a total of 12 distinct classes. Among these, three classes were particularly prominent, representing major research themes in the network.

b) Size Ranking with Harmonic Closeness Centrality: Harmonic closeness centrality is used to measure the importance of a node based on its average distance to other nodes in the network. For node v, the harmonic closeness centrality $C_H(v)$ is defined as:

$$C_H(v) = \sum_{u \neq v} \frac{1}{d(u,v)}$$

where $d(u,v)$ is the shortest path distance between nodes u and v. This metric is particularly useful in large, sparse networks where traditional closeness centrality may yield high values for nodes in disconnected subgraphs. By ranking nodes according to their harmonic closeness centrality, we effectively highlight the most influential or accessible nodes within the network. In our analysis, node sizes are scaled according to their harmonic closeness centrality, making prominent nodes more visually discernible in the diagrams.

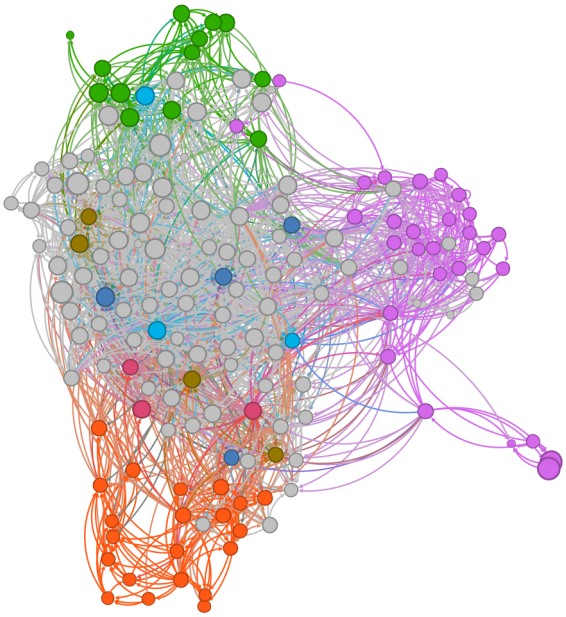

Fig. 63. Node-link diagram using the Forceatlas 2 algorithm.

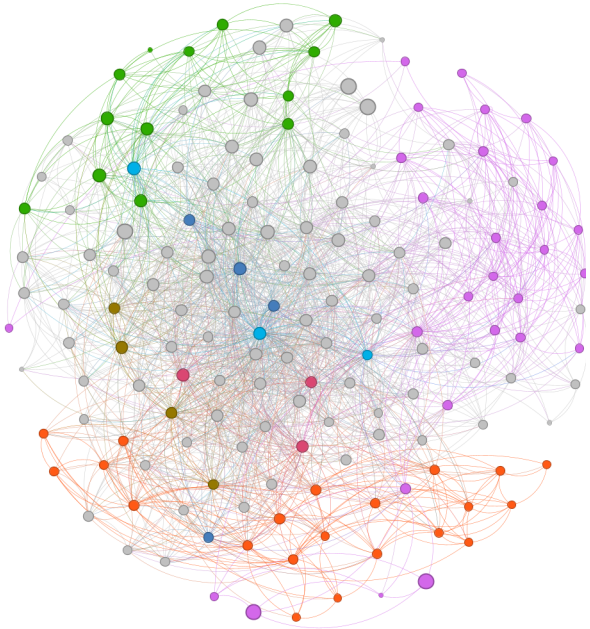

Fig. 62. Node-link diagram using the Fruchterman-Reingold algorithm.

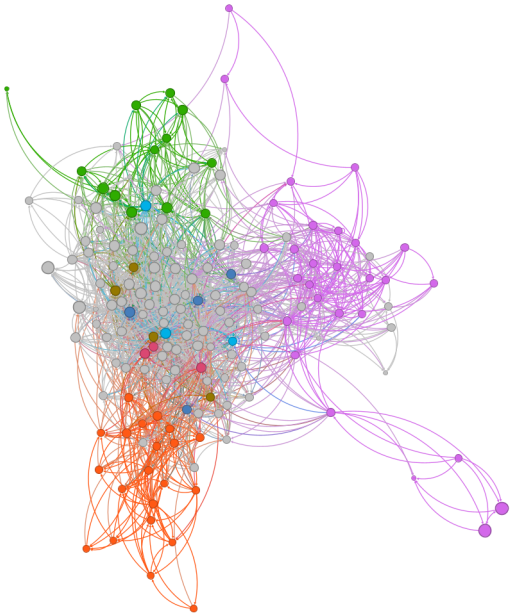

Fig. 64. Node-link diagram using the Yifan Hu algorithm.

c) Layout Algorithms: We employ three distinct layout algorithms to generate the node-link diagrams: Fruchterman-Reingold, ForceAtlas2, and Yifan Hu. These algorithms arrange nodes in a way that minimizes edge crossing and clustering, providing visually appealing layouts. We rank the algorithms based on the resulting layout quality, with an emphasis on minimizing clustering and enhancing overall readability:

1) **Fruchterman-Reingold(fig 62)**: Offers the least clustering and overall best layout.
2) **Yifan Hu(fig 64)**: Provides a balanced layout with moderate clustering.
3) **ForceAtlas2(fig 63)**: Results in more clustering, making it less optimal for our purposes.

d) Research Communities and Group Descriptions: After partitioning by color and observing the inferred classes, we identified 12 distinct groups. Among these, three groups were notably more prominent. The thematic descriptions for these groups are as follows:

- **Group 1(Green): AdS/CFT Correspondence**

and Cosmological Applications of Brane Theory
This group centers on the AdS/CFT (anti-de Sitter/conformal field theory) correspondence and its applications in brane cosmology, large N gauge theories, and gravitational theories. These papers explore how AdS/CFT duality aids in understanding renormalization group flow, supersymmetry, and the holographic principle in cosmological settings. Key topics include brane cosmological evolution, holographic bounds, and phase transitions in AdS space, revealing insights into non-compactified dimensions, bulk cosmology, and large N field theories.

- **Group 2(Orange): Supersymmetric Gauge Theories, Integrability, and Quantum Moduli Spaces**
  This group focuses on $N = 2$ supersymmetric gauge theories, integrable systems, and exact solutions for quantum moduli spaces. Topics such as Seiberg-Witten theory, mirror symmetry, and non-perturbative solutions are covered, along with the vacuum structure of $N = 2$ Yang-Mills theories and heterotic string compactifications. This group's emphasis on integrable systems, moduli spaces, and dualities is essential for understanding the quantum structure of supersymmetric field theories.
- **Group 3(Purple): Noncommutative Dynamics, Tachyon Condensation, and Non-BPS Brane Physics**
  This group explores noncommutative geometry in string theory, focusing on non-BPS branes, tachyon condensation, and noncommutative solitons. Topics include noncommutative field theories, K-theory in D-brane physics, and tachyon potentials describing instability phenomena in brane-antibrane pairs. Research on large-N models, thermal instabilities, and deformation quantization provides insights into string field theories, noncommutative solitons, and the dynamics of non-BPS D-branes.

IV. Author's Contribution

Ullas G:(Quiver Plot and Parallel Coordinates Plot)
Adithya Nangarath:(Color Map and Node-Link Diagrams)
Mallikarjun Chakoti:(Contour Map and Tree Map)
Each team member worked independently on their assigned task, including all aspects of data analysis, visualization creation, interpretation, and description, unless specified otherwise. The collaborative effort of all team members contributed to the comprehensive analysis presented in this report.

References

1) Kaggle. (n.d.). Customer Personality Analysis Dataset. Retrieved from https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis
2) Stanford Network Analysis Project (SNAP). (n.d.). High-Energy Physics Theory Citation Network. Retrieved from https://snap.stanford.edu/data/cit-HepTh.html
3) Climatology Lab. (n.d.). GridMET: Gridded Meteorological Data. Retrieved from https://www.climatologylab.org/gridmet.html