Data Visualization on Crimes in India Dataset

1st Rishav Chandel
MT2023058
Data Visualization
International Institute of Information Technology
Bangalore, India
rishav.chandel@iiitb.ac.in

2nd Subhodip Rudra
MT2023103
Data Visualization
International Institute of Information Technology
Bangalore, India
subhodip.rudra@iiitb.ac.in

I. Introduction

In this paper, we will perform data visualization on crimes in India, spanning various years and crime categories. The visualizations will include trends across different states, crime types, and justice system parameters, such as arrests and convictions. Through charts, graphs, and comparisons, we aim to uncover patterns, disparities, and potential correlations in the data. The insights gained will help better understand crime dynamics and support data-driven policy recommendations.

II. Task - I : Crime Trends and Distribution Visualization

II-1 Dataset Description and Structure: Crimes in India

The Crime and Anti-Corruption dataset contains district-level crime data and anti-corruption case statistics across India from 2001 to 2014. It includes several tables with the following key components:

- **District-wise Crimes (IPC)**: Records crime data by district, including categories such as murder, rape, kidnapping, dacoity, robbery, theft, and total IPC crimes, aggregated by year.
- **Crimes Against SCs and STs**: Tracks crimes against Scheduled Castes (SCs) and Scheduled Tribes (STs), including murder, rape, kidnapping, dacoity, and crimes under the POA and PCR Acts.
- **Crimes Against Children**: Covers crimes such as child murder, rape, abduction, foeticide, and violations of the Prohibition of Child Marriage Act.
- **Crime by Place**: Classifies crimes by location (e.g., residential, highways, railways, etc.), including dacoity, robbery, burglary, and theft.
- **Anti-Corruption Cases**: Provides details on anti-corruption investigations, including the number of cases registered, investigated, convicted, acquitted, and recoveries made.

Each dataset includes columns for `STATE/UT`, `DISTRICT`, `YEAR`, and various crime-specific categories. The anti-corruption dataset also includes metrics on case outcomes such as convictions, acquittals, and recoveries.

II-2 Justification for Using this PCP and TreeMap

Parallel Coordinates Plots (PCP) are ideal for visualizing multivariate datasets, such as the total crimes per year categorized by crimes against SC, ST, women, and children. PCP allows us to represent each variable on a parallel axis and show how these variables correlate across states and years. For instance, we can observe trends, relationships, and patterns in crime distribution over time. PCP is particularly effective for comparing states and detecting outliers or high-crime regions where specific categories of crime are significantly higher. By using PCP, we gain a holistic view of the temporal and categorical interplay of crime data across states.
Treemaps are well-suited for visualizing hierarchical and proportional data, making them ideal for analyzing state-wise crime types, corruption cases, and crimes by place of occurrence. By organizing data into nested rectangles, treemaps provide an intuitive representation of the magnitude and contribution of each category. For example, state-wise crime types can highlight the dominant crimes in each state, while state-wise corruption data can show how various corruption metrics contribute to the total cases. Similarly, crimes by place of occurrence (e.g., residential premises, highways) can be visualized proportionally to identify hotspots. The hierarchical structure and proportional sizing of treemaps allow for quick comparisons and efficient insight extraction.

II-3 Task Overview and Objectives

This task focuses on analyzing crime data across different districts, with a particular emphasis on the location of occurrence. Two key areas of analysis will be:

- This task involves analyzing the trends in various crimes such as murder, rape, robbery, and theft, broken down by district and year. The goal is to understand regional crime patterns and their evolution over time.
- This task focuses on categorizing crimes based on the place where they occurred, such as residential premises, highways, railways, banks, and other commercial establishments. The aim is to provide a spatial distribution of crime types, helping to identify hotspots and crime trends specific to locations.
- The second task focuses on anti-corruption data for the years 2001 to 2010, with visualizations representing yearly trends, case processing, and property seizure. This task highlights the number of cases registered, the number

of cases investigated, and the amount of property seized during the specified years.

- These tasks aim to provide a comprehensive understanding of crime patterns and anti-corruption efforts, enabling insights for regional and temporal variations in crime rates and anti-corruption actions for data-driven decision-making.

II-4 Visualization using Parallel Coordinates Plot

The Parallel Coordinates Plot (PCP) is utilized to visualize year-wise and state-wise crime data trends from 2001 to 2013 across multiple crime categories, including total IPC crimes, crimes against Scheduled Castes (SC), Scheduled Tribes (ST), women, and children. Each vertical axis represents one of these crime categories, while each line represents the data for a specific year, capturing both the overall and state-wise crime distribution.

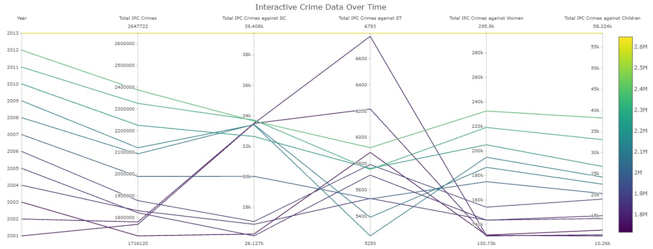

Fig. 1. Parallel Coordinates Plot visualizing year-wise and state-wise trends in crime data from 2001 to 2013. The plot highlights total IPC crimes and crimes against SC, ST, women, and children, with each line representing a specific year. The color gradient indicates the magnitude of crimes, aiding in the identification of significant trends and patterns across different crime categories.

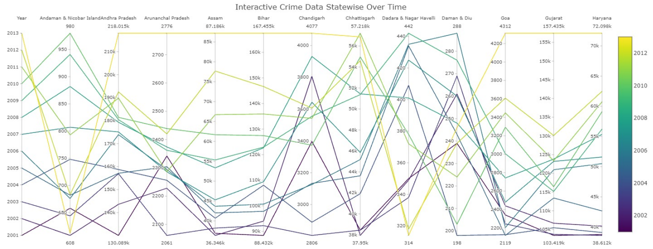

Fig. 2. Parallel Coordinates Plot showing state-wise and year-wise crime data across all states from 2001 to 2013. Each line represents crime trends for a specific state over the years, encompassing total IPC crimes and crimes against SC, ST, women, and children. The color gradient denotes the severity of crime rates, enabling a comparative analysis of crime patterns across states and over time.

Key insights from the plot:

- **Trends Over Time:** The plot highlights an increasing trend in total IPC crimes over the years, with significant variations in certain categories, such as crimes against women and children.
- **State-wise Comparison:** The PCP enables year-wise state-level comparison by displaying each state's crime data across multiple categories, making it easier to observe which states have higher crime rates in specific categories over time.
- **Correlations Between Categories:** By examining line proximities and intersections across different axes, potential correlations between various crime categories can be identified, such as the link between crimes against SC/ST and total IPC crimes.
- **Magnitude of Crime Types:** The color gradient serves as a visual aid to represent the scale of crimes, where lighter colors indicate higher values, allowing for a quick identification of years and states with peak crime rates.

This PCP effectively captures temporal, categorical, and regional variations in crime data, providing a comprehensive overview of trends and enabling quick identification of significant patterns and anomalies.

II-5 Visualization using TreeMap

The treemap visualization categorizes crime data hierarchically, displaying the distribution of crime incidents across different states and categories (such as crimes against SC, ST, women, and children). Each rectangle represents a state, with the size indicating the proportion of crimes relative to other states. Within each state section, sub-categories represent specific types of crimes, allowing for a comparative analysis of crime proportions across regions and types. The color gradient in the treemap provides an additional dimension to gauge the intensity or volume of crimes, highlighting areas with higher or lower crime rates.

This visual approach aids in identifying states with higher crime proportions and in observing the distribution across various crime categories, contributing to a clearer understanding of crime patterns at a state and category level.

III. Task - II : Visualization of Sea Surface Dynamics: Color Mapping, Contour Mapping, and Quiver Plots with Temporal Animation

III-1 Dataset Description and Structure: GridMet

The GridMET climate dataset provides comprehensive meteorological data, capturing high-resolution weather variables over the continental United States. This analysis utilizes GridMET data from August 2023 to October 2023, specifically focusing on climate parameters relevant to assessing daily weather patterns, humidity, solar radiation, precipitation, and wind dynamics.

Dataset Structure:

The dataset comprises netCDF files, with each file representing a specific climate parameter over a spatial grid. Each entry includes essential elements, allowing detailed spatiotemporal analysis:

- **Latitude and Longitude:** Specifies the geographic grid for each data point.
- **Time:** Each file contains daily observations over the analysis period, capturing temporal variations.

- **Parameter Variables:** The dataset includes several key meteorological variables, each stored in a separate file:
  - **Burning Index (bi):** Assesses fire potential based on atmospheric and fuel conditions.
  - **Energy Release Component (erc):** Measures the potential energy release in wildfire conditions.
  - **Evapotranspiration (etr, pet):** Includes reference evapotranspiration for grass and alfalfa.
  - **Fuel Moisture (fm1, fm10, fm100, fm1000):** Reflects dead fuel moisture content over various time scales.
  - **Precipitation (pr):** Records daily precipitation levels across the spatial grid.
  - **Relative Humidity (rmin, rmax):** Minimum and maximum near-surface relative humidity measurements.
  - **Specific Humidity (sph):** Indicates near-surface humidity at different points on the grid.
  - **Solar Radiation (srad):** Surface-level solar radiation measurements.
  - **Temperature (tmmn, tmmx):** Daily minimum and maximum near-surface air temperatures.
  - **Vapor Pressure Deficit (vpd):** Measures atmospheric water demand.
  - **Wind Speed and Direction (vs, th):** Captures 10-meter wind speed and direction data.

The dataset structure supports both quantitative and spatial analyses of climate variables, offering a rich basis for exploring climate patterns, studying seasonal shifts, and understanding the dynamics of weather events. With each parameter captured at daily intervals and across a structured geographic grid, this dataset facilitates in-depth studies on regional climate trends and meteorological variations over time.

III-2 Justification for Using Contour Plot, Heatmap and Quiver Plot

Contour plots and heatmaps are ideal for visualizing spatially continuous variables like those in this dataset, such as temperature (tmmn, tmmx), humidity (rmin, rmax), and solar radiation (srad). These plots provide a clear and intuitive representation of gradients, patterns, and anomalies across a geographical area. Contour plots highlight regions of similar values with isolines, making them effective for identifying zones of extreme conditions, such as droughts (pdsi) or high vapor pressure deficits (vpd).

Heatmaps, on the other hand, use color intensity to convey the magnitude of variables, enabling quick comparisons between regions or time periods. Both visualizations allow researchers to identify correlations or trends among multiple environmental factors, making them powerful tools for climate and environmental analysis.

Quiver plots are well-suited for visualizing wind data, specifically wind speed (vs) and direction (th). Each arrow in the quiver plot conveys both the magnitude (via length) and direction (via orientation) of the wind, offering an intuitive way to analyze spatial wind patterns. This visualization is particularly effective for understanding dynamic weather systems, detecting prevailing wind directions, and identifying regions of high wind activity. By overlaying quiver plots on maps or alongside other environmental variables, researchers can assess how wind interacts with humidity, temperature, or fire risk indicators, adding depth to meteorological and ecological studies.

III-3 Task Overview and Objectives

We will focus on two key tasks: Climate Parameter Visualization and Wind Dynamics Representation.

- The first task involves creating visualizations to analyze climate parameters, including temporal trends, spatial distribution, and intensity variations. These are represented through contour plots, heatmaps, and contour fill animations to highlight changes over time and geography.
- The second task focuses on wind dynamics, visualized using quiver plots to illustrate wind speed and direction patterns across different regions and timeframes.
- These tasks aim to provide a comprehensive understanding of climate variability and wind behavior, enabling insights for data-driven climate analysis and decision-making.

III-4 Data Preprocessing

The following preprocessing steps were conducted on the GridMET dataset to prepare it for visualization and analysis, focusing on the period from August 2023 to October 2023:

- **Data Extraction and Filtering:** Each climate parameter file (in netCDF format) was opened and converted to a DataFrame. The data was filtered to include only dates between August 1, 2023, and October 31, 2023, focusing the analysis on this timeframe.
- **Handling Missing Values:** Missing or NaN values were removed from each climate variable to ensure that only complete records were used in further analysis and visualization.
- **Pivoting Data for Spatial Grid Structure:** For each selected date, data for the respective climate variable was pivoted into a 2D grid format with latitude and longitude as axes, allowing for spatial mapping.
- **Contour Level Definition:** For each climate parameter, contour levels were defined using a linear scale from the minimum to the maximum values within the dataset. This ensured consistent scaling across visualizations.
- **Vector Calculation for Wind Data:** For wind speed and direction data, the horizontal (u) and vertical (v) components were calculated using trigonometric functions to enable vector visualization in quiver plots.
- **Sampling for Wind Vector Clarity:** A sampling factor was applied to reduce the density of wind vectors, facilitating clearer visual representation.

These preprocessing steps structured the data appropriately for the creation of contour plots, heatmaps, and vector field animations, ensuring clarity and consistency in visualizations.

III-5 Visualization and Analysis of Climate Variables

A: ContourPlot

Contour plots were generated to visualize the spatial distribution of climate parameters from August 1, 2023, to October 31, 2023. The process involved the following steps:

- **Contour Level Definition:** Contour levels were defined using a linear scale between the minimum and maximum values of each parameter.
- **Plotting Contours:** The ‘matplotlib‘ library was used to create the contour plots, representing the climate variables against latitude and longitude.
- **Color Mapping:** A color map (e.g., ’inferno’) was applied, with a color bar indicating the value range.
- **Customization:** Titles, axis labels, and a color bar were added to the plots for clarity.
- **Saving and Animation:** The plots were saved as images and compiled into an animated GIF to show temporal variations.

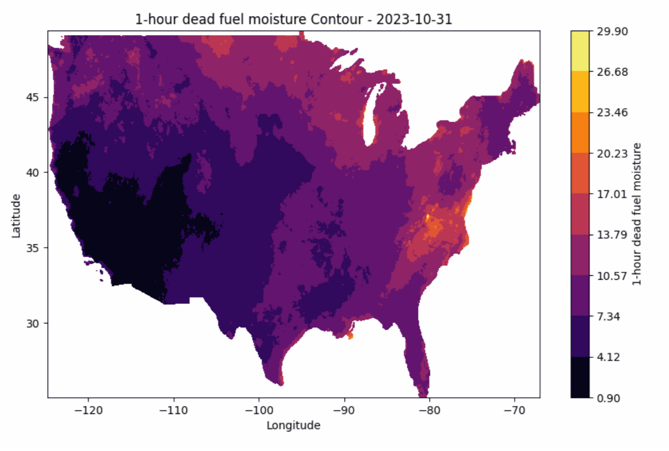

Fig. 3. Contour map of 1-hour dead fuel moisture across the United States. This visualization highlights spatial variations in moisture levels, indicating potential fire risk areas with lower moisture values. Data captured for the period from August to October 2023.

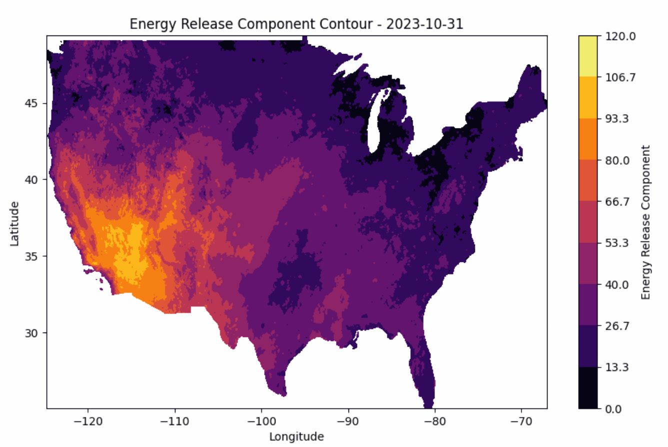

Fig. 4. Contour map of the Energy Release Component (ERC) across the United States. This map illustrates spatial patterns in ERC values, providing insights into regions with higher energy release potential, which may indicate areas at elevated fire risk. Data captured from August to October 2023.

These contour plots effectively displayed the geographical trends and variations of the climate parameters over time.

B: HeatMap

Heatmaps were generated to visualize the spatial distribution of climate parameters from August 1, 2023, to October 31, 2023. The process involved the following steps:

- **Heatmap Generation:** The ‘matplotlib‘ library was used to create heatmaps by plotting the climate parameters on a 2D grid with latitude and longitude.
- **Color Mapping:** A color map (e.g., ’inferno’) was applied to represent the intensity of the parameter values, with a color bar indicating the value range.
- **Customization:** Titles, axis labels, and a color bar were added for better clarity and context.
- **Saving and Animation:** The heatmaps were saved as images and compiled into an animated GIF to visualize changes over time.

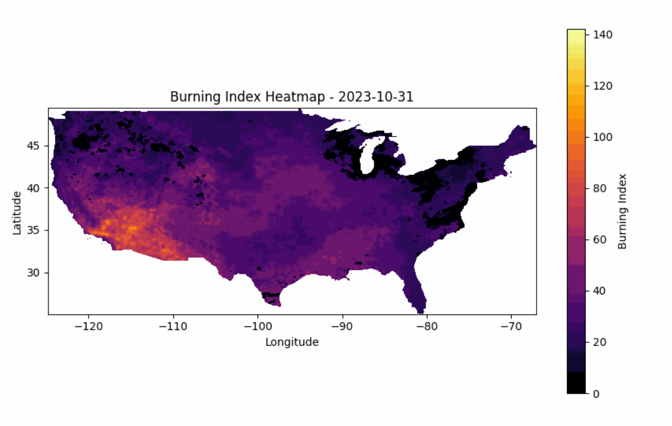

Fig. 5. Heatmap of the Burning Index across the United States. This visualization displays the spatial distribution of the Burning Index, which estimates potential fire intensity and spread. Higher index values indicate regions with elevated fire risk. Data captured from August to October 2023.

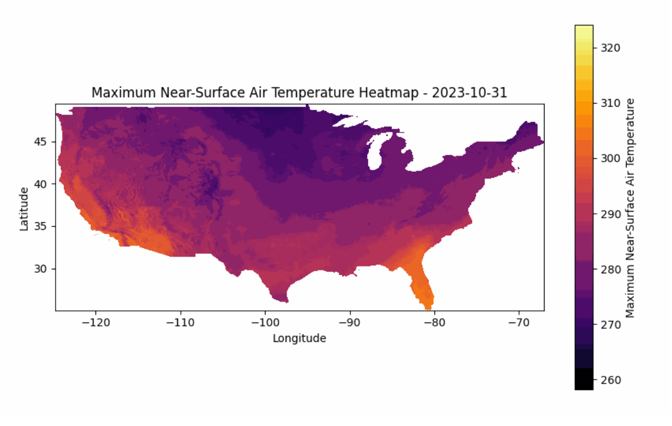
Maximum Near-Surface Air Temperature Heatmap - 2023-10-31
Latitude
45
40
35
30
−120
−110
−100
−90
−80
−70
Longitude
Maximum Near-Surface Air Temperature
320
310
300
290
280
270
260

Fig. 6. Heatmap of Maximum Near-Surface Air Temperature across the United States. This visualization highlights temperature variations across regions, indicating areas with higher near-surface temperatures that may contribute to fire risk. Data captured from August to October 2023.

The heatmap visualizations provided an effective way to understand the intensity and spatial variations of climate parameters across the study period.

C: Quiver

Quiver visualizations were created to represent wind patterns, including speed and direction, over the study area from August 1, 2023, to October 31, 2023. The process involved the following steps:

- **Wind Vector Components:** Wind speed and direction data were converted into u and v components using trigonometric transformations.
- **Grid Sampling:** The dataset was sampled at regular intervals to reduce the density of vectors, ensuring clarity in visualization.

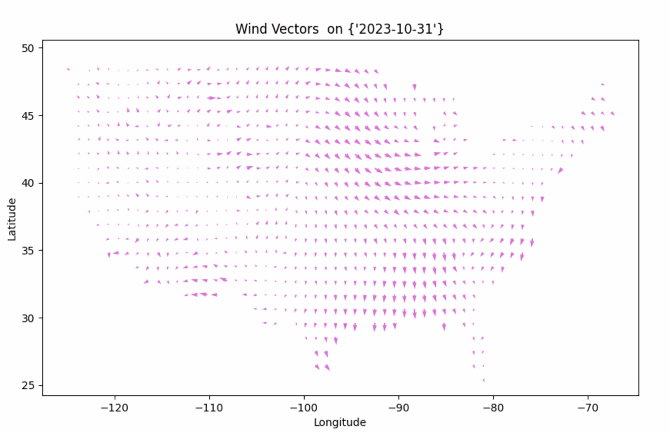

Fig. 7. Quiver plot showing vector field representation with uniform arrow lengths, indicating direction and magnitude across the domain.

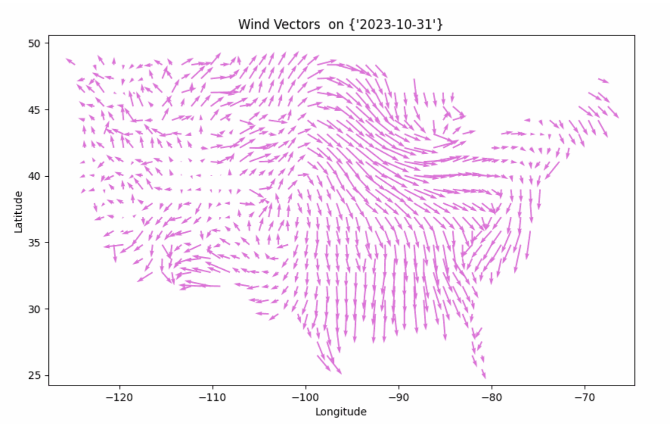

Fig. 8. Quiver plot showing wind vectors with magnitudes proportional to wind speed across the United States. The arrows represent wind direction, while arrow length and color intensity indicate wind speed. This plot provides insights into wind patterns affecting fire risk by highlighting areas of stronger winds that could drive fire spread. Data spans from August to October 2023.

- **Plotting:** Quiver plots were generated using the sampled wind vector components, with arrows indicating both direction and magnitude of the wind.
- **Customization:** Titles, axis labels, and a consistent color scheme were applied to enhance readability.
- **Animation Creation:** Individual quiver plots for each date in the period were compiled into animated GIFs to visualize temporal changes in wind patterns.

These visualizations provided insights into spatial and temporal variations in wind dynamics, aiding the analysis of meteorological conditions.

Insights from Visualizations

Contour Plot Insights

The contour plot visualization provides valuable insights into the spatial distribution of weather variables across the United States:

- **Spatial Patterns:** The contour plot highlights geographic patterns in variables such as temperature, moisture, or pressure. Areas with densely packed contour lines represent regions with rapid changes, indicating a sharp gradient in the variable, while smoother gradients suggest more uniform conditions.
- **Influence of Topography:** The contours reveal how topography impacts climatic variables, particularly in mountainous regions. The presence of high-density contour lines in these areas suggests that elevation influences weather conditions significantly.
- **Temporal Variability:** By observing contour changes over time, one can identify seasonal shifts or weather fronts moving through various regions. This temporal insight helps in understanding how weather patterns evolve and affect different regions throughout the observation period.

Heatmap Insights

The heatmap visualization for the 1-hour dead fuel moisture variable, from August to October 2023, provides insights into moisture distribution across the United States:

- **Moisture Distribution:** The color gradient in the heatmap represents moisture levels, with lighter colors indicating higher moisture and darker shades signifying lower moisture. The western U.S. exhibits lower moisture levels, particularly in desert and arid regions, suggesting a higher risk of wildfires in these areas.
- **Temporal Fluctuations:** The animation reveals changes in moisture over time, with some regions experiencing increased moisture levels, possibly due to rainfall events or seasonal shifts, especially in the Midwest and the Northeast.
- **Regional Variability:** The Southeastern U.S. shows relatively high moisture levels, which might reduce fire risk, whereas the Southwest and parts of California consistently display low moisture levels, indicating areas of elevated wildfire risk.

Quiver Plot Insights

The quiver plot visualization depicts wind vectors, showing both the direction and speed of winds over the observation period:

- **Wind Patterns:** The quiver plot demonstrates the predominant wind directions and speeds, with vector lengths corresponding to wind speed. This helps in identifying consistent wind patterns, such as prevailing westerlies or easterlies in certain regions.
- **Weather Systems:** Variations in wind strength and direction over time suggest the movement of weather systems, such as the approach of a storm front or a high-pressure zone. Regions with longer vectors indicate areas experiencing stronger winds, which could relate to low-pressure systems bringing weather changes.
- **Implications for Fire Risk:** Strong winds, especially in regions with low fuel moisture, increase the risk of wildfire spread. The quiver plot highlights areas with elevated wind activity, offering insight into potential high-risk zones where wind-driven fire spread may be more likely.

Combined Insights

When analyzed together, these visualizations provide a comprehensive understanding of the environmental conditions:

- **Interactions Between Variables:** The contour, heatmap, and quiver plots reveal how variables such as moisture, temperature, and wind interact. For instance, low fuel moisture and high wind speeds can create ideal conditions for fire spread, particularly in dry areas of the western U.S.
- **Temporal Patterns:** The animations enable tracking of changes over time, providing insights into seasonal trends and weather system movements that impact climate and environmental conditions.
- **Applications in Risk Management:** These visualizations are crucial for understanding regional fire risks, guiding resource allocation, and informing preventive measures in fire-prone areas.

Together, the contour plot, heatmap, and quiver plot visualizations offer a holistic view of spatial and temporal weather patterns across the United States, aiding in identifying trends that may impact wildfire risk and environmental management efforts.

IV. Task - III : Visualization and Comparison of Graph Layout Algorithms in Complex Network Structures Using Gephi

IV-1 Dataset Description and Structure: Wikipedia Request for Adminship

The Wikipedia Request for Adminship (RfA) dataset provides detailed information on the support and opposition votes for users nominated for admin roles on Wikipedia. It consists of 11,382 nodes and 176,584 edges, representing the complex interactions between voters and candidates. Each entry in the dataset includes key elements such as the source (SRC), which indicates the username of the voter; the target (TGT), referring to the candidate's username; the vote (VOT), which records a numerical representation of the voter's stance (1 for support, 0 for neutral, and -1 for oppose); and the result (RES), which captures the outcome of the vote (1 for successful, -1 for unsuccessful). Additionally, the dataset records the year (YEA) of the vote, a timestamp (DAT) specifying the exact date and time of voting, and a text field (TXT) that provides a brief comment or rationale for the voter's decision. This rich dataset offers a comprehensive view of the dynamics of voting behavior and the outcomes of adminship requests on Wikipedia.

IV-2 Justification for Using this Visualization

A node-link or graph layout visualization is particularly well-suited for analyzing the Wikipedia adminship election dataset due to its inherent network structure. Each node represents a Wikipedia member, and directed edges represent the votes cast, capturing the relational dynamics between voters and candidates. This visualization enables us to clearly illustrate the complex voting patterns, highlighting clusters of strongly connected users and revealing influential individuals or groups. The directed edges with signed values allow for easy differentiation between positive and negative votes, providing a visual insight into community polarization.

The visualization also facilitates the identification of voting behavior trends over time and across different elections. By integrating edge thickness or color to denote vote frequency or type, the graph enhances the interpretability of repetitive voter-votee interactions. Furthermore, graph layouts are excellent for detecting structural properties such as communities, bridges, or key players, which are vital for understanding the social and collaborative aspects of Wikipedia's governance. This approach combines clarity with the ability to uncover non-obvious patterns, making it an ideal tool for exploring this dataset's richness.

IV-3 Task Overview and Objectives

This task aimed to analyze the Wikipedia Request for Adminship (RfA) dataset and derive insights into voting patterns and user influence within the Wikipedia community. The primary focus was on creating a network of voters (SRC) and candidates (TGT) based on support and opposition votes, as well as identifying the most influential users each year through network analysis. Using this information, we constructed directed graphs to visualize voting dynamics, detect patterns in user interactions, and identify top candidates for adminship.

Objectives

1) **Data Processing and Filtering**: Clean and preprocess the dataset to ensure valid and complete information, specifically focusing on the source, target, and vote values.
2) **Network Construction**: Build directed networks for each year based on the support and opposition votes, creating edges between voters (SRC) and candidates (TGT).
3) **Node and Edge Analysis**: Analyze nodes (voters and candidates) based on in-degree and other centrality measures, identifying the top 10 most influential users for each year.
4) **Gephi Visualization**: Export the network data to GEXF format, allowing for visualization in Gephi to better understand voting behavior, user influence, and community structure over time.
5) **Yearly Trends**: Examine yearly changes in the voting patterns and the influence of different users, providing insights into the dynamics of Wikipedia's adminship process.

Through this analysis, we aimed to reveal key patterns in Wikipedia's community governance and decision-making process, highlighting influential users and the evolution of voting behavior across different years.

IV-4 Data Preprocessing

Prior to using Gephi for network analysis, several preprocessing steps were applied to prepare the dataset effectively:

- **Year-wise Data Split:** The dataset was split based on each year to facilitate a time-series analysis of network changes over time. This step allowed for a more granular examination of node influences and interactions as they evolved annually.
- **Standardization of Nodes and Edges:** Node and edge identifiers were standardized across the dataset to maintain consistency. This included verifying unique node IDs and ensuring that edges only connected valid nodes.
- **Export for Gephi Compatibility:** The preprocessed data was saved in a format compatible with Gephi (e.g., CSV or GEXF), ensuring smooth import and analysis within the software.

This preprocessing pipeline allowed for a robust, year-by-year network analysis, facilitating insights into evolving patterns of influence and connectivity within the network.

A: Fruchterman Reingold Algorithm

The **Fruchterman-Reingold** algorithm is a force-directed layout that simulates a physical system where nodes repel each other like charged particles, and edges act as springs that attract connected nodes. It uses two forces:

- *Repulsive Force:* Prevents nodes from clustering too tightly.
- *Attractive Force:* Pulls connected nodes closer together.

The algorithm iteratively adjusts node positions until equilibrium is reached, making it useful for visualizing networks and revealing structure.

Pros and Cons

Pros:

- Produces clear, aesthetically pleasing layouts.
- Reveals clusters or communities in the network.

Cons:

- High computational cost with a quadratic time complexity $O(n^2)$.
- Inefficient for large networks.

Application to Wikipedia RfA Dataset

Given the large size of the Wikipedia Request for Adminship (RfA) dataset (11,382 nodes, 176,584 edges), the Fruchterman-Reingold algorithm is computationally expensive. It is better suited for smaller or filtered subsets of the data.

Finding Top 25 Most Influential Nodes Using Gephi

To find the top 25 most influential nodes based on degree, you can use the degree filter in Gephi. This filter allows you to:

- Sort nodes by degree.
- Select the top 25 nodes with the highest degree.

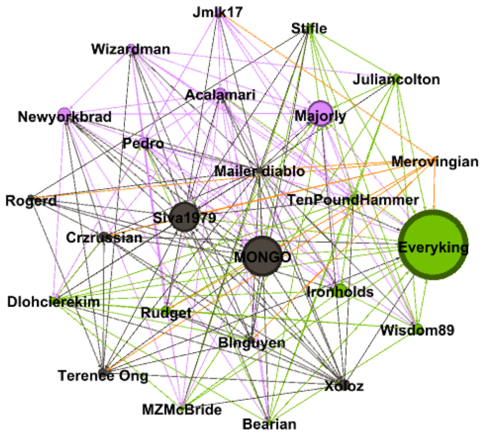

Fig. 9. The layout highlights clusters of influential nodes, with the top 25 nodes identified based on degree. Nodes are colored on Modularity class and sized on degree.

Inference

- Nodes with higher degrees (more connections) are likely the most central and influential in the Wikipedia admin nomination process, suggesting a stronger impact on the outcome of nominations.
- Clusters formed by the algorithm indicate communities or groups of users with similar voting patterns, which can be analyzed further to identify trends in support or opposition for candidates.
- The combination of the Fruchterman-Reingold layout and the degree filter provides deeper insights into the network dynamics, revealing patterns of user engagement and influence within the Wikipedia administrative process.

B: ForceAtlas2 Algorithm

ForceAtlas2 is a force-directed layout algorithm used for visualizing networks. It simulates nodes as repelling particles and edges as attractive springs.

- *Repulsive Force:* Nodes repel to prevent clustering.
- *Attractive Force:* Edges pull connected nodes together.
- *Gravitational Force:* Centers the network, useful for disconnected graphs.
- *Parameters:* Adjustable *scaling* (node distance) and *speed* (convergence rate).

Pros:

- Organizes clusters naturally, making it easy to visually identify communities or clusters.
- Works well with modularity-based clustering, as nodes within the same community tend to be grouped closely together.

Cons:

- Computationally intensive, especially for large networks.
- May not reveal the underlying network structure as clearly for very large and sparse networks.

Application to Wikipedia RFA Dataset

The dataset used in this analysis spans from 2008 to 2013, selected based on the following reasoning:

- The dataset shows a substantial reduction in the number of interactions (edges) starting from 2008. The earlier years (2003-2007) had much higher levels of voting activity, which may skew the network structure.
- The period from 2008 to 2013 provides a more stable and relevant subset of data, reflecting more recent and consistent voting behavior on Wikipedia RfA nominations.
- Using only data from 2008 to 2013 helps focus on the evolving dynamics of the community's engagement with the RfA process, excluding the influence of higher voting activity in the earlier years.

By focusing on this range of data, we aim to provide a clearer representation of the network's structure and dynamics, with more meaningful insights into the influential nodes and their connections.

Filters Applied in Gephi

The graph was generated using ForceAtlas2 with the following filters applied:

- **Giant Component: True**, which ensures only the largest connected component of the graph is considered.
- **K-Core: 5**, which retains nodes with a minimum degree of 5.
- **In-Degree Range: 87 to 566**, focusing on nodes with significant in-degree values to highlight the most influential nodes in the network.

The importance of in-degree is significant in identifying influential nodes as it directly correlates to the number of incoming connections or votes a user has received in the RfA process, representing the level of support or opposition they have garnered.

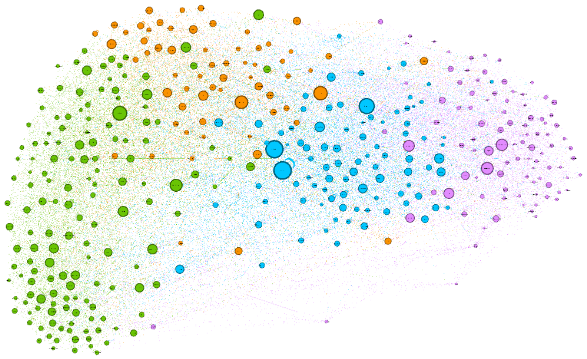

Fig. 10. Network visualization of the Wikipedia Request for Adminship dataset using the ForceAtlas2 algorithm with a 5-core filter and in-degree range of 87 to 566. The graph shows the largest connected component, highlighting influential nodes based on their in-degree centrality.

Inference

- Clusters of voters can be observed, suggesting that certain groups or factions of users may share similar opinions on the candidate's suitability for adminship.
- Candidates who are more widely supported by different voter clusters appear at the center, with dense connections from various voter nodes.
- The algorithm's ability to reduce edge crossings and minimize overlap between nodes enhances the readability of the graph, making the relationships between voters and candidates clearer.
- The layout highlights the connectedness and centrality of nodes, offering insights into the influence and popularity of candidates within the Wikipedia community.

C: Yifan Hu Algorithm

Yifan Hu is another force-directed layout algorithm known for its ability to handle large graphs efficiently. It uses a hierarchical, multi-scale approach to achieve a well-organized layout with minimal edge crossings and node overlap. The algorithm optimizes the layout by combining the concepts of force-directed layout with techniques from spectral graph theory.

Algorithm Description

- Yifan Hu algorithm applies forces between nodes, where attractive forces pull connected nodes together, and repulsive forces push them apart.
- Unlike simpler force-directed algorithms, Yifan Hu leverages a multilevel approach, optimizing the layout progressively across multiple levels of graph coarseness.
- It is computationally efficient, enabling it to work well with large-scale graphs while minimizing layout time.

Pros:

- Handles large-scale graphs efficiently with minimal edge crossings and good node distribution.
- Computationally faster than other force-directed layouts like Fruchterman-Reingold, especially for large networks.

Cons:

- Less effective at revealing fine-grained community structures compared to some other algorithms.
- Can sometimes produce suboptimal layouts for highly irregular graphs.

Application to Wikipedia RFA Dataset

The dataset used in this analysis spans from 2008 to 2013, selected for the following reasons:

- The voting activity in the earlier years (2003-2007) was considerably higher, which may distort the visualization, especially when focusing on the most influential nodes.
- By restricting the data to 2008-2013, we focus on a more stable and representative portion of the dataset, providing a clearer understanding of the evolving network dynamics.
- This range also eliminates the influence of early spikes in voting activity, offering a more balanced view of the RfA process.

Focusing on this subset of data allows for a more meaningful analysis of network structure, where the centrality of nodes plays a more significant role in determining influence.

Eigenvector Centrality Threshold

The **Eigenvector Centrality** of nodes was used as a measure of influence, with a threshold value of 0.466 applied to filter the most influential nodes in the network. Eigenvector centrality measures the influence of a node in a network based on the centrality of its neighbors. A high eigenvector centrality indicates that a node is connected to other highly influential nodes, making it a key player in the network.

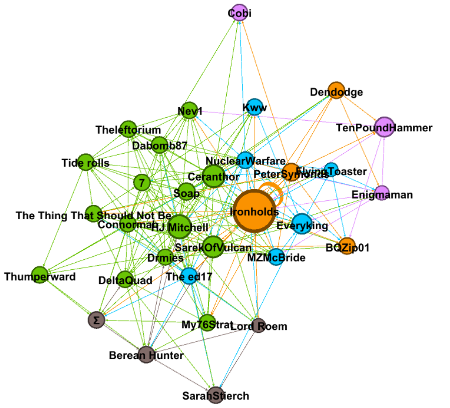

Fig. 11. The graph demonstrates an Yifan Hu layout with minimal edge crossings and node overlap, showcasing the relationships between voters and candidates in the adminship process colored on modularity class and sized on in degree

Inference

- The nodes in the graph are positioned in such a way that groups of closely related participants (voters with similar opinions or candidates with mutual support) are placed near each other, highlighting communities within the RFA process.
- The centrality of candidates who receive broad support can be easily identified, with these nodes positioned in more central locations of the graph, surrounded by many edges from other nodes.
- The algorithm's effectiveness in minimizing node overlap and edge crossings allows for better visibility of individual relationships, which is crucial in understanding the dynamics of the voting process in RFA.
- The Yifan Hu layout reduces visual clutter by spreading out the graph in a balanced manner, making it easier to explore the various levels of support and the structure of the voter-candidate interactions.
- Larger clusters of supporters and opposing voters are visually separated, providing insight into potential factions or coalitions in the decision-making process.

Conclusion

In this task, we have explored and compared various graph layout algorithms for visualizing the Wikipedia Request for Adminship (RfA) dataset using Gephi. The dataset, with its 11,382 nodes and 176,584 edges, represents a complex network of interactions between voters and candidates, where the goal was to uncover voting patterns and identify influential users.

Fruchterman-Reingold Algorithm, a force-directed layout, successfully revealed clusters within the network, highlighting communities of voters with similar stances. It generated aesthetically pleasing and intuitive layouts, making it easy to identify highly influential nodes based on degree centrality. However, the algorithm's quadratic time complexity limited its scalability, making it computationally expensive for large networks like the Wikipedia RfA dataset. Despite this, it was effective for smaller or filtered subsets of the data, where it provided valuable insights into user engagement and influence.

ForceAtlas2 Algorithm, another force-directed layout, demonstrated better organization of the network with natural clustering. By applying a gravitational force, it centered the network and reduced edge crossings, offering clearer visualizations of the connections between voters and candidates. The algorithm excelled in revealing clusters of supporters and opposition, showcasing the dynamics of Wikipedia's community governance. However, like the Fruchterman-Reingold algorithm, ForceAtlas2 also faced challenges with large networks in terms of computational complexity. Its ability to group voters by similar opinions helped identify factions within the community, while the inclusion of a 5-core filter and in-degree centrality allowed for a more focused analysis of the most influential nodes.

Yifan Hu Algorithm, known for its efficiency with large-scale graphs, provided a well-organized layout with minimal edge crossings and node overlap. This algorithm was particularly effective in handling the computational challenges posed by large datasets like Wikipedia RfA. Its hierarchical approach allowed for faster layout generation compared to other force-directed algorithms, making it suitable for visualizing large networks. By focusing on the eigenvector centrality threshold, we were able to filter out the most influential nodes, providing a clearer view of user influence in the network. The Yifan Hu layout also reduced visual clutter, making it easier to identify communities, factions, and the centrality of candidates in the decision-making process.

Overall Insights

Each algorithm presented unique advantages and limitations, with the choice of layout depending on the specific objectives and characteristics of the dataset. While Fruchterman-Reingold and ForceAtlas2 excelled in generating aesthetically pleasing visualizations with clear community structures, they were computationally intensive for large datasets. On the other hand, the Yifan Hu algorithm provided a more efficient solution for handling large-scale graphs, maintaining clarity and minimizing visual clutter.

In conclusion, a combination of these algorithms can offer a comprehensive view of the Wikipedia RfA network. The Fruchterman-Reingold and ForceAtlas2 algorithms are useful for revealing community structures and influential nodes in smaller subsets of data, while the Yifan Hu algorithm is more suitable for efficiently handling larger datasets and providing a scalable solution for visualizing network dynamics. By applying these layout algorithms, we have gained deeper insights into voting patterns, user influence, and the overall structure of the Wikipedia adminship process.

Acknowledgment

We would like to express my sincere gratitude to **Prof. Jaya Sreevalsan Nair**, whose guidance and support were invaluable throughout this project. Our appreciation also goes to **the Teaching Assistants of Data Visualization course**, whose collaboration and insights greatly enhanced the quality of this work. Special thanks to **International Institute of Information Technology, Bangalore** for providing the resources and data necessary for this analysis.

References

[1] Kaggle - Crime in India Dataset
[2] Kaggle - India GIS Data
[3] India Map Files were collected from Igismap