Chronic Disease Indicators Community Analysis

Devanshi Pratiher[1], Dhatri Patel[2], Mayank Tiwari[3], Brandon Winn[4]
[#]*San Jose State University*
California, USA

***Abstract*—This paper presents a comprehensive analysis of the Chronic Disease Indicator (CDI) dataset, obtained from the Center for Disease Control (CDC), focus on cancer-related factors, geographical information, mortality rates, and gender demographics. The team performed community detection on the Chronic Disease Indicator (CDI) dataset, published by the Center for Disease Control (CDC). The te explored the CDI dataset with respect to Cancer, location data, mortality, and gende provided by data source). The data was analyzed using several community detection algorithms, including Affiliation Graph Model, BigClam, Spectral Graph Partitioni Girvan Newman, and Louvain to identify factors closely related to a high mortality incidence for persons diagnosed with cancer. Ultimately, the project aimed to identif most critical communities to target with resources in order to improve the health outcomes of cancer patients.**

I. INTRODUCTION

Cancer and other chronic diseases have emerged as the leading ca of mortality and disability, exerting a significant burden on healt systems worldwide. The understanding of mortality trends and dispa associated with chronic diseases, particularly cancer, is crucial for p health initiatives and resource allocation. Therefore, collecting, analy interpreting, and disseminating data on chronic diseases like canc vital to understanding and raising awareness about mortality disparities. The Division of Population Health within the Cente Disease Control and Prevention (CDC) has developed a comprehe set of 124 indicators within the Chronic Disease Indicator (CDI) da that was developed by consensus which allows states and territories large metropolitan areas to uniformly define, collect, and report ch disease data that are important to public health practice and availabl states, territories and large metropolitan areas. Considering these fa and its impact on individuals as well as healthcare industries, this pr research focuses on community analysis for one of the chronic disea Cancer. This research uses data published by the CDC for det analysis.

In light of the substantial impact of cancer on individuals the healthcare industry, this research focuses on community ana within the realm of cancer research. By leveraging the rich da provided by the CDC, we aim to delve into the intricate detai cancer-related factors, such as location, disease duration, and mor outcomes, to gain insights into the dynamics of cancer communities. CDC's Chronic Disease Indicator dataset encompasses 43 columns, representing specific information pertaining to cancer incidence outcomes. These columns offer a nuanced view of the dis considering different factors like location, duration of the dis whether the disease caused mortality or not, etc. Considering different parameters affecting the cancer, this research focused detecting the different communities using different community dete algorithms.

By applying state-of-the-art community detection techni including Affiliation Graph Model, BigClam, Spectral Graph Partitio Girvan Newman, and Louvain, we seek to identify distinct communities within the cancer population. This analysis aims to shed light on factors closely associated with higher mortality rates among cancer patients. Understanding these communities and their unique characteristics will enable targeted resource allocation strategies, ultimately leading to improved health outcomes for individuals battling cancer. By employing advanced community detection algorithms, we aim to uncover meaningful insights that will inform strategic interventions and resource allocation efforts. Ultimately, this research seeks to contribute to the field by enhancing our understanding of cancer communities and facilitating more effective support systems for cancer patients.

We hypothesize that by applying community detection algorithms to analyze the social networks of individuals diagnosed with cancer, distinct communities will emerge, each exhibiting different levels of mortality incidence. Through this research, we aim to identify key factors that are closely associated with high mortality rates among cancer patients. Additionally, we seek to determine the critical communities within the cancer population that require targeted resource allocation in order to improve health outcomes.

II. RELATED WORK

Researchers M.E. Newman and M. Girvan developed a technique for community detection which is still used today and has been iterated upon by several preceding algorithms. Betchel used community detection on data for diseases such as cancer and tumor types as well. This included a proposed community-based lung cancer detection approach. In their study, Haq and Wang examine genomic datasets to identify subgroups within twelve types of cancers. They investigate the survival rates of these communities and analyze the distribution of tumor types across these subgroups. Taya and their team compared community detection algorithms on neuroimaging data to identify regions of the brain responsible for certain behaviors. A paper by Yang and Sun introduces a novel measure that combines closed walks and clustering coefficients to replace the edge betweenness in the Girvan and Newman method for hierarchical clustering. The experimental results demonstrate that this method achieves a better balance between accuracy and runtime, and reveals the significance of nontrivial closed walks in constructing community structures and analyzing network structures. Additionally, the proposed method offers a new perspective for addressing the double peak structure problem in complex networks. “Community-Affiliation Graph Model for Overlapping Network Community Detection.” This paper introduces the Community-Affiliation Graph Model, a novel model-based community detection method that effectively captures overlapping, non-overlapping, and hierarchically nested communities in various types of networks. The method outperforms existing approaches and challenges

the conventional wisdom that community overlaps are less dei connected than the non-overlapping parts.

III. TECHNICAL APPROACH

A. Data Cleaning and Preprocessing

In order to ensure the quality and reliability of the data, a thor data cleaning process was performed on our dataset. This invo removing any duplicates, missing values, and irrelevant information the dataset obtained from the CDC. Furthermore, the data pre-processed to standardize variables, normalize features, and ad any outliers or inconsistencies. This process was time-consuming as were a lot of inconsistencies with the dataset. To commence our cleaning process, we accounted for NaN values in our dataset. following columns had 100% null values which is why we dropped columns.

Response: 662608
StratificationCategory2: 662608
Stratification2: 662608
StratificationCategory3: 662608
Stratification3: 662608
ResponseID: 662608
StratificationCategoryID2: 662608
StratificationID2: 662608
StratificationCategoryID3: 662608
StratificationID3: 662608

Due to the substantial number of missing values in DataValueFootnoteSymbol and DatavalueFootnote features, accou for 69.22% of the total values (458,647 out of 662,608), these fea were deemed insufficient for comprehending their applicability t obtained results. Furthermore, these columns were found to be irrel to the problem's outcome and inconsequential for compreher predictions derived from the dataset. Consequently, these two col were eliminated. Additionally, following the same rationale, the col LowConfidenceLimit and HighConfidenceLimit, which exhibit missing data rate of 48.52%. Duplicate values within the columns discarded, and only rows containing the 'Cancer' topic within the da were retained. Furthermore, specific column data types were convert numeric to facilitate their utilization within the algorithms.

The next step was to prepare the data for visualization. The da represented by the DataFrame df_cancer_int, was scaled using StandardScaler from the sklearn.preprocessing module. The scaled was then converted into a new DataFrame called X_scaled_df, columns labeled as 'LocationID', 'category_strat', 'category_strat1', 'Mortality'. The scaled data was further normalized using the norm function from sklearn.preprocessing. This ensured that all features on a similar scale, allowing for more accurate analysis. The norma data was stored in the DataFrame X_normalized.

To reduce the dimensionality of the data for visualization employed Principal Component Analysis (PCA) from sklearn.decomposition module. We specified the number of compo as 3 to create a 3D plot. The resulting transformed data was stored i DataFrame X_principal with columns labeled as 'P1', 'P2', and 'P3'.

B. Spectral Clustering Algorithm

To detect communities in the network of persons diagnosed with cancer, the Spectral Clustering algorithm was employed. This algorithm leverages the spectral properties of the graph Laplacian to partition the network into clusters. First, an affinity matrix was constructed based on the similarity between nodes.

The spectral clustering algorithm was applied using the Gaussian kernel affinity matrix. We imported the SpectralClustering class from sklearn.cluster and initialized an instance named spectral_model_rbf with the number of clusters set to 4 and the affinity parameter set to 'rbf'.

To visualize the clustering results, the prep_graph function was defined to extract the relevant columns from the DataFrame X_scaled_df and store them in a list format suitable for plotting.

The clustering results were visualized using a 3D scatter plot. We created a figure with subplots using plt.subplots and iterated over the three data representations in Xs (i.e., X_scaled_df, X_normalized, and X_principal). For each data representation, we retrieved the corresponding graph data using prep_graph and plotted it using the scatter function. The cluster assignments were determined by calling graph_model.fit_predict on the respective data representation, and colors were assigned using the 'winter' colormap. The resulting figures were displayed, with each axis representing a principal component.

The process described above was repeated for the affinity matrix computed using the Euclidean distance. The SpectralClustering instance model_nn was initialized with the number of clusters set to 4 and the affinity parameter set to 'nearest_neighbors'. We visualized the clustering results using the same procedure as in step 6, but now with the affinity matrix calculated using the Euclidean distance.

To analyze specific questions using the clustering results, we performed the following steps.

Cluster Label Assignment: We assigned cluster labels to the data points in X_principal using the fit_predict method of the spectral clustering model model. The labels were stored in the labels variable.

Location Mapping and Data Exploration: We created a mapping of location abbreviations using the columns 'LocationAbbr' and 'LocationID' from the df_cancer DataFrame. This mapping was used to identify the corresponding location abbreviation for each data point.

Data Analysis Based on Labels: We examined specific subsets of the data based on the assigned cluster labels. For example, we selected a sample of data points and printed their corresponding labels, location IDs, location abbreviations, stratification categories, stratification IDs, and mortality indicators. The mappings category_strat_map, category_strat1_map, and Mortality_map were used to interpret the numerical values in the data and provide meaningful labels.

Data Analysis - Southern States: We focused on analyzing whether southern states had a higher level of cancer. To do this, we extracted the data points from the southern states by comparing the location abbreviations with a predefined list of southern states. The relevant data points were stored in the south_data list. We then printed the labels, location IDs, location abbreviations, stratification categories, stratification IDs, and mortality indicators for these data points.

Data Analysis - Men's Mortality: We further explored the clustering results by examining the mortality indicators specifically for men across different states. We extracted the data points corresponding to men's

mortality and stored them in the mens_mort_data list. We printe labels, location IDs, location abbreviations, stratification categ stratification IDs, and mortality indicators for these data points.

These analyses provided insights into the clustering results allowed us to explore specific questions related to cancer comm detection. The visualization and analysis presented in this se contribute to a better understanding of the clustering patterns and implications for different aspects of cancer research and public health

C. Louvain's Algorithm

We started by extracting the 'LocationID' column from the Ca dataset and converting it to the integer data type. This column repre the entities (locations) in the dataset. Additionally, we created a li edges by pairing each entity with its corresponding question fron dataset. Building the Affiliation Graph: We created an example undir graph using nx.Graph() and added nodes representing the entities edges representing the connections between entities and corresponding questions.

Degree Centrality Calculation: We calculated the degree centrali each node in the affiliation graph using the nx.degree_centra function. The results were stored in the degree_centrality variable.

Visualizing the Affiliation Graph: We visualized the affiliation g using the nx.draw() function, setting the node size, color, and font siz better visualization. The resulting plot provided an overview o affiliation relationships between entities and questions.

Community Detection using Louvain's Algorithm: We ap Louvain's algorithm for community detection by calling community_louvain.best_partition() function on the affiliation graph. resulting community assignments were stored in the comms dictio mapping each node to its assigned community ID.

Community Visualization: We visualized the communities usi spring layout for improved spatial representation. The nodes were col based on their assigned community using a color mapping defined b cmap dictionary. The resulting plot provided insights into the comm structure within the affiliation graph.

Location Data: We extracted the unique location abbreviations IDs from the df_cancer dataset and stored them in the df_loc DataFrame. This data can be used for further analysis and interpret of the community assignments based on geographical information.

The implementation of Louvain's algorithm allowed us to identify visualize communities within the cancer dataset, providing valu insights into the underlying structure and relationships between loca and questions. These findings contribute to a better understanding o interconnections and patterns in the dataset, facilitating targ interventions and resource allocation in cancer research and public he

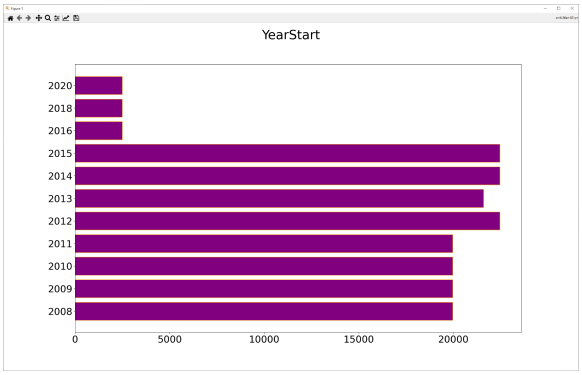
D. BigCLAM

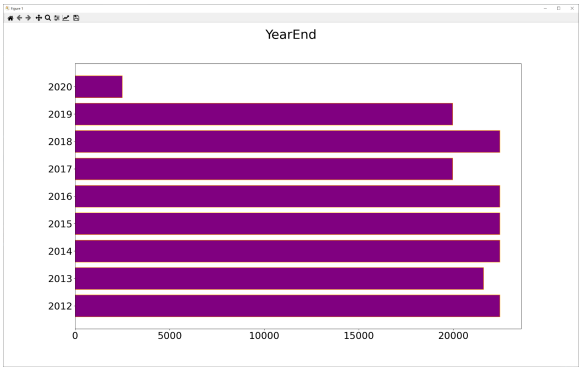
The intuition behind using the BigCLAM algorithm is that there be overlapping communities for different types of cancer within dataset. We created a different subset of data that included DataValu as the frequency of occurrence. Once created, we encoded all the s data into categories via integers, and finalized the rows multidimensional points of a graph. In order to build a g representation, we calculated edge weights as relative distance between rows as a simple subtraction function across all the rows. This way the DataValueAlt would be the biggest influence on the edge weight.

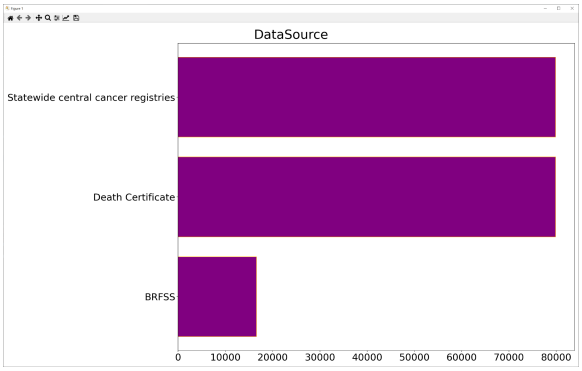
Unfortunately, the algorithm was not able to determine any clear communities, and dumped all the data into one big class. We attempted to run several other out-of-the-box algorithms on the same graph like Clauset-Newman-Moore greedy modularity maximization and LPA. The greedy modularity maximization algorithm was able to detect several communities, but falsely puts a large number of rows into 2 giant communities and only selects a few for the rest.

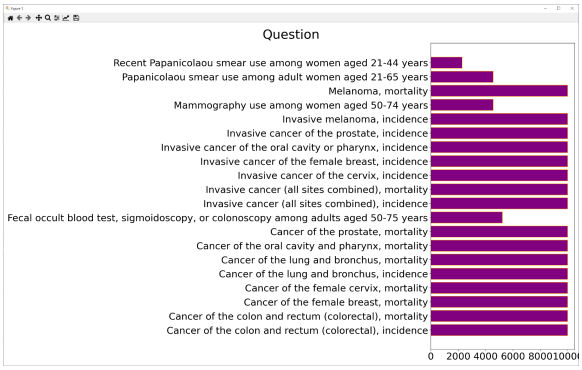
Ultimately, this dataset is not suitable for a BigCLAM-type graph analysis for community detection. Drawing inference is possible only with decreasing specificity in analyzing row by row relationships. Due to the dataset's nature, it is only possible to make extremely high-level inferences without the detailed knowledge provided by missing in-state data.

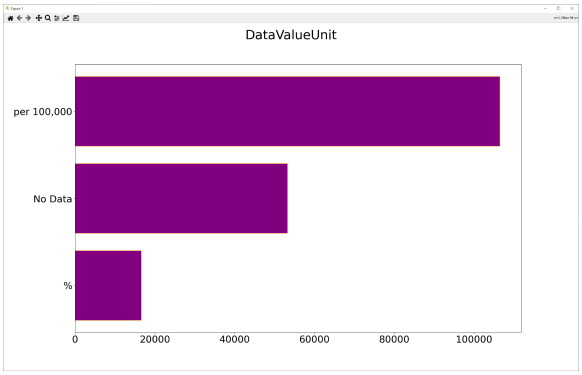
E. Data

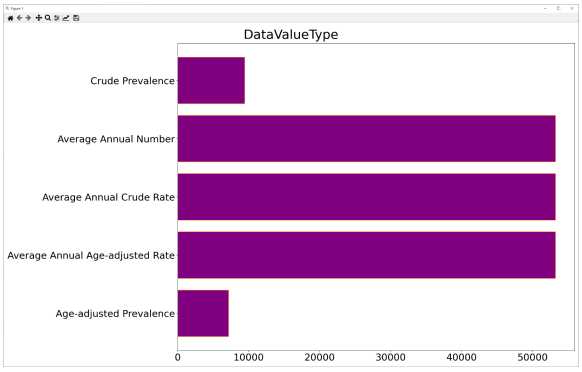


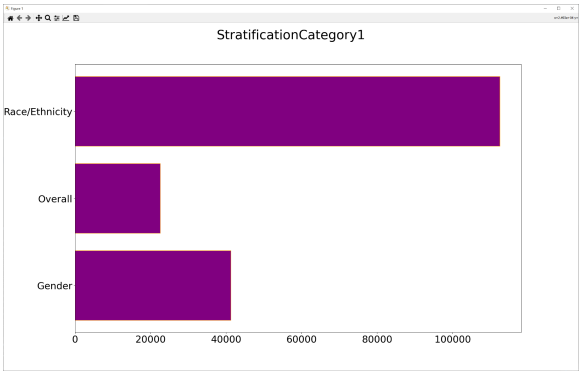


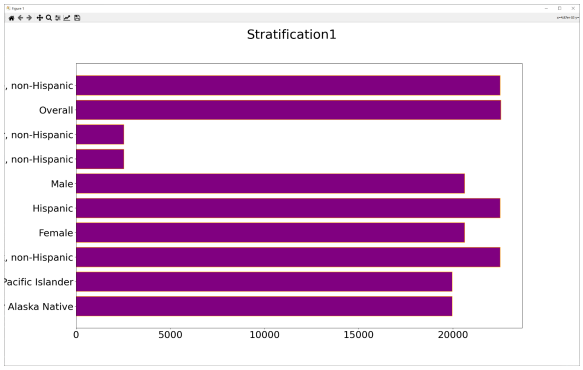


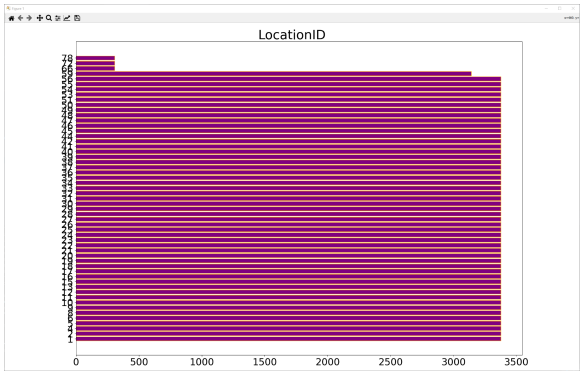












- Regarding YearStart/End, most studies take place >2008 and generally end <2020. Certain states have not made the cut due to lack of data quality and quantity (Notable examples: Texas, Florida).
- Question - The text assigned to DataValue. The real context lies in here and needs to be picked out by filters. Example: our mortality data column is derived from whether ‘mortality’ exists in the text.
- Response - Not filled out in the dataset. We have to use some DataValue as our main data - percentage/number per capita of mortality vs incidence for example. It has to be scaled by DataValueType contextually.
- Data value types will provide context for different scales of numbers. We include “Gender” which is defined as sex assigned at birth under Stratification 1.

It is important to restructure the data into a format that provides context to its values. In order to facilitate community detection, we transformed the data to represent coordinates in multidimensional space, where categorical variables are encoded into discrete values along their dimensional axis. In doing so, we hope that community detection algorithms can determine complex boundaries

Once loaded into a multidimensional matrix, all the values are normalized to facilitate speed and minimize the effect of outliers. In order to visualize some boundaries, we performed PCA on a sample set to reduce the number of dimensions to a visualizable number, 3.

IV. METHODOLOGY

Once we had the relevant data after the selection of the cancer specifically in the anonymized data, we proceeded to the first step o data processing, which was selection of Relevant Features. A s selected features related to patient demographics, cancer types, location of patients were chosen for community detection. These fea were carefully curated to capture the key factors influencing resc requirements.

Next, we implemented our Community Detection Algorithms. Affiliation Graph Model, BigClam, Spectral Graph Partitioning, G Newman, and Louvain algorithms were implemented and applied t chronic diseases cancer dataset. Each algorithm utilized diff approaches to identify communities or clusters within the dataset.

VI. CONCLUSION

The analysis of the CDC data revealed several communities inclu a close grouping of high mortality from cancer diagnosis southern/Appalachian states, a clustering of high mortality for men cancer diagnoses, and a decrease in cancer diagnoses and cancer dis duration over the years.

In this study, our team applied various community dete algorithms, including Affiliation Graph Model, Spectral Cluste Louvain's Algorithm, and BigCLAM, to analyze the Cancer dataset goal was to identify distinct communities within the dataset and insights into factors associated with high mortality rates among ca patients. Through our analysis, we discovered the community struc exist within the Cancer dataset, indicating the presence of group individuals with similar characteristics and outcomes. The comm detection algorithms allowed us to uncover these communities and g deeper understanding of the underlying patterns and relationships ar the variables.

One of the first algorithms we implemented was the Spe Clustering algorithm which revealed four clusters within the dataset b on location, stratification category, stratification ID, and mort Louvain's Algorithm further provided insights into community struc within the affiliation graph of the dataset, helping identify commur and their corresponding nodes. Unfortunately, BigCLAM underfi dataset and classified every point into 1 community.

However, our analysis also had limitations. The dataset conta extensive missing values and duplicates, which required data cleaning preprocessing steps. Additionally, certain features in the dataset found to have little relevance to the research question and were ther dropped.

In terms of future research, there are several avenues to exp Firstly, further investigation into the identified communities could pr valuable insights into the factors contributing to high mortality among cancer patients. Understanding these factors can help in desig targeted interventions and resources for improving health outcomes.

Moreover, the application of advanced community dete algorithms opens up possibilities for uncovering more intr community structures and gaining a deeper understanding of underlying dynamics. Exploring other community detection method comparing their performance on the Cancer dataset could also contr to a comprehensive analysis.

Overall, this study demonstrates the effectiveness of certain community detection algorithms in analyzing complex datasets such as the Cancer dataset. The identified communities and their characteristics can aid policymakers, healthcare professionals, and researchers in developing targeted strategies to improve cancer patients' health outcomes. Further research in this area has the potential to enhance our understanding of chronic diseases and facilitate the development of more effective interventions.

ACKNOWLEDGEMENT

The team would like to thanks the contributors to Cdlib (https://cdlib.readthedocs.io/en/latest/index.html) and Networkx (https://networkx.org/documentation/stable/index.html) whose libraries we utilized in this project.

REFERENCES

[1] (2023, Jan.) Centers for Disease Control and Prevention website. [Online]. Available: chronicdata.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-CDI-/g4ie-h725.

[2] Price JD. Amerson NL, Barbour KE, Emuze DV. (2020, Jul.) American Journal of Health Promotion. [Online]. Available: Prevalence of Frequent Mental Distress Among Illinois Adults with Chronic Conditions: Estimates From the Behavioral Risk Factor Surveillance System, 2011-2017.

[3] Calanan RM, Sandoval-Rosario M, Price JD, Samanic CM, Lu H, Barbour KE. (3028. Nov.) [Online]. Available: Achieving Excellence in the Practice of Chronic Disease Epidemiology.

[4] M. E. Newman and M. Girvan. "Finding and evaluating community structure in networks". (2004) Physical Review E, vol. 69, no. 2, pp. 026113, 2004.

[5] J. J. Bechtel, W. A. Kelley, T. A. Coons, M. G. Klein, D. D. Slagel and T. L. Petty, "Lung cancer detection in patients with airflow obstruction identified in a primary care outpatient practice", Chest, vol. 127, no. 4, pp. 1140-1145, 2005.

[6] N. Haq and Z. J. Wang, "Community detection from genomic datasets across human cancers", 2016 IEEE Global Conf. on Signal and Infor. Process., pp. 1147-1150, 2016.

[7] F. Taya, J. De Souza, N. V. Thakor and A. Bezerianos, "Comparison method for community detection on brain networks from neuroimaging data", Appl. Network Sci., vol. 1, no. 1, pp. 8, 2016.

[8] Y. Yang, P. G. Sun, X. Hu and Z. J. Li, "Closed walks for community detection", Physica A: Statistical Mechanics and its Applications, vol. 397, pp. 129-143, 2014.