CoWiz: Interactive Covid-19 Visualization Based On Multilayer Network Analysis

Kunal Samant*, Endrit Memeti†, Abhishek Santra‡, Enamul Karim§ and Sharma Chakravarthy¶
IT Lab and CSE Department, University of Texas at Arlington, Arlington, Texas
Email: *kunalnitin.samant@mavs.uta.edu, †endrit.memeti@mavs.uta.edu,
‡abhishek.santra@mavs.uta.edu, §enamul.karim@mavs.uta.edu, ¶sharmac@cse.uta.edu

***Abstract*— `Covid Wizard or CoWiz` is a Covid-19 visualization dashboard based on Multilayer Network (MLN) analysis underneath[1]. Online dashboards typically plot/visualize statistical information gleaned from raw data, such as daily cases, deaths, recoveries, tests, etc. However, for a better understanding, we need aggregate analysis (e.g., community, centrality) and its visualization which is the purpose of CoWiz. As an example, grouping counties across a country/region based on similarity of increase/decrease in cases, deaths, hospitalizations over intervals is not possible without aggregate analysis. This is where CoWiz utilizes community and other concepts over MLNs that are inferred from Covid and other relevant data sets for visualization.**

This demo presents a flexible, interactive dashboard which is capable of visualizing various aspects of Covid-19 data, including composition of Covid data with demographics (population density, education level, average earning, vehicle movements, and change in purchase patterns) at the granularity of county for USA. This paper elaborates on the types of analysis, underlying model, and how a flexible visualization dashboard has been developed using open source software and data sets. As new data becomes available, they can be incorporated into the visualization with no manual intervention.

***Index Terms*—Community Detection, Modeling Using MLNs, Composition using Decoupling Approach, Covid-19 data Analysis**

I. Motivation

On Jan 20, 2020, USA reported its first COVID-19 positive case. Since then the virus has spread to all (3141) US counties at different rates. As of mid-November 2020, USA has more than 11M cases and 240K deaths! A number of features are associated with every confirmed case, such as hospitalization, death, or recovery making this data set complex with diverse entity (person, county), feature (case, hospitalization, death, ...), and relationship (similarity in cases, hospitalizations, deaths, ...) types. Currently, many statistical data models are used to plot the *'peak', 'dip', and 'moving averages' or 'colored maps'* in the number of cases along with the daily reporting of new cases, deaths, recoveries and tests ([4], [6], [1], [7], [10], [9]). However, for a comprehensive understanding of the spread of the pandemic, there is a need to *analyse the effect of different events (mask requirement, social distancing, gathering restrictions), demographics, and protocol enforcement-related features in multiple geographical regions across different time periods*.

Geographical proximity of counties, demographics (population densities, traffic movements across boundaries, per-capita income, educational qualifications), and economics (spending, occupancy) are directly or indirectly influenced by the spread or curb of Covid-19. Broadly, the analysis and visualization of this complex Covid-19 data set can be broken down into following categories with increasing difficulty.

I. Statistical Visualization of Raw Data

(A1) Which counties have reported the maximum/minimum number of new cases, deaths, recoveries and tests?

(A2) Visualize the daily cases on world map

(A3) When will US go past the peak in the number of new cases? Is a next wave predicted?

II. Temporal Aggregate Analysis & Visualization

(A4) In which regions, was the lockdown most effective? That is, how have geographical regions with maximum (and minimum) rise in cases shifted between the periods around the lockdown?

(A5) How has been the spread in the regions where election rallies were held? Are the regions with an uptick caused by the rallies with mass gatherings?

III. Event-Based Aggregate Analysis & Visualization

(A6) Which geographical regions showed a downward trend similar to the lockdown period during long weekends?

(A7) Are weekdays or weekends leading to a larger spread?

(A8) Which regions got significantly affected due to major events like July 4^{th}, Labor Day, Memorial Day, Halloween? What precautions need to be taken for future events like Thanksgiving and Christmas? The inverse can be computed which may be more helpful.

IV. Parameter-Based Aggregate Analysis & Visualization

(A9) What is the effect of traffic movement on new cases across centrally connected counties? How to choose a county (centrality) for lockdown or vaccine administration for maximum impact?

(A10) Compare increase/decrease in number of a feature (e.g., new cases) with respect to avg. education, per-capita income, mask usage, types of expenditure done by people and population density. For instance, are counties with higher population density witnessing a surge in cases in periods where the brick-and-mortar spending was more?

(A11) Which dense regions of the country are more susceptible to the virus spread due to the presence of institutions with large work force (e.g., meat packing plants with difficulty in maintaining social distancing, clustered traffic

[1]UTA CoWiz dashboard URL: https://itlab.uta.edu/cowiz/

movement in specific time periods?) In turn, it will also be helpful to find which institutions should continue to enforce in-person or remote activity?

Currently available online dashboards address Category I above and focus on either reporting and visualizing daily cases on maps ([10], [1], [9], [2]) or time series plots and statistical modeling ([4], [7], [6]). They are more focused on visualizing the raw daily data. However, for the other categories, there is a need to **model** the *different entities and relationships* involved in the Covid data set (cases, regions or time periods) in order to *analyze* and understand from *multiple perspectives* and then finally *visualization* to maximize understanding. Tools such as Database management systems can be used for long-term storage and analysis. They require schema generation, population, data integration as and when they become available, and writing specific queries. Even then some aggregate analysis queries are difficult to express in SQL. Hence, for the dashboard presented in this paper, we have streamlined the process by developing a **flexible and efficient interactive, web-based portal for Covid-19 analysis and visualization**.

Underneath, this dashboard is supported by **Multilayer Networks** that model [13] the COVID-19 data set due to its ability to handle multiple entities, relationships and features. Informally, MLNs are layers of networks (also called multiplexes) where *each layer is a simple graph and captures the semantics of a (or a subset of) feature of an entity type*. The layers can also be connected. Moreover, the **divide and conquer based decoupling approach** [14], [15] is used to accomplish different analysis objectives making the process sound, efficient, and scalable.

The primary contributions of this demo are:

- An **automated, interactive web-based demo** for analyzing and visualizing the spread of Covid-19.
- Use of **multilayer network for modeling** Covid and related data set collected from different sources.
- Use of **decoupling-based analysis** to integrate related information for study of Covid-19 and visualization.
- A **flexible and automated architecture** that can be repurposed easily for other data sets.

II. Role of Multilayer Networks

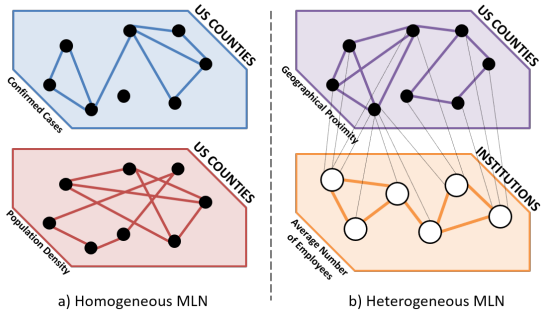


Fig. 1: Homogeneous and Heterogeneous MLNs

Multilayer networks (or MLNs) have been proposed in the literature to be an important alternative for *flexible, expressive, and structure-preserving* modeling as well as efficient for analyzing complex data sets based on different combinations of features [12], [17], [13]. MLN model is a *network of networks*. In this case, every layer represents a distinct relationship among entities with respect to a single feature. The sets of entities across layers, which may or may not be of the same type, can be related to each other as well.

Based on the type of relationships and entities, multilayer network can be of different types. Layers of a **homogeneous MLN (or HoMLN)** are used to model the diverse relationships that exist among the **same type of entities**. **(A4)** to **(A10)** need to analyze the same set of US counties thus connecting them based on similar number of confirmed covid cases, population density and so on (Fig.1(a)). Objectives, such as **(A11)** that need to analyze the relationships among **different types of entities** like US counties (connected based on geographical proximity) and Institutions like universities and MNCs (connected if have similar average number of employees (or students and staff) across branches) can be modeled using **heterogeneous MLNs (or HeMLNs).** The inter-layer edges represent the relationship across layers as shown in Fig. 1(b).

A. Decoupling-based Analysis Approach

Recently, a novel decoupling approach has been proposed for detecting communities and centralities in an efficient manner. This uses the equivalent of "divide and conquer" for MLNs ([13], [18], [14], [16]). Decoupling-based analysis uses individual layers as partitions. The partial (layer-wise or intermediate) analysis results for community/centrality detection of MLNs are iteratively composed to produce the final results. Substantial work has been done to identify the composition function that is appropriate for efficient community/centrality detection (referred to as Ψ) for MLNs.

III. `CoWiz:` UTA Covid-19 Dashboard

Our major focus here is analyzing the spread in geographical regions across different time periods. And to relate it to one or more demographic and other features. The notion of a region can be of different granularity - country, state, county, zip code etc. Here, we have used the Covid-19 data reported by New York Times [3], census data [8] and interesting data from other sources to compile a data set for the **3141 counties** in the US starting from February 2020. The data set includes features like number of new cases and deaths, lat-long of the county seat, the median household income, population density by land area, total land area, educational qualifications (like %adults with a high school diploma), traffic movement and so on.

In this visualization, we have focused on analyzing *all US counties* based on individual or combination of features. This results in a HoMLN, where the 3141 nodes in each layer correspond to counties. The analysis of categories II, III, and IV (of Section I) require *understanding the impact of an event or preventive measure*. For this, either we consider **(i)** the two time periods centered around the event or **(ii)** a **base period**, when the event did not occur and a **target period**, when the event occurred. Moreover, we quantify the **extent of Covid-19 spread between two periods** by finding out the **percentage change in the number of new cases between the periods**. Each chosen time period pair leads to the generation of a layer where the nodes (counties) are connected if there is a *similar* percentage change (for a given

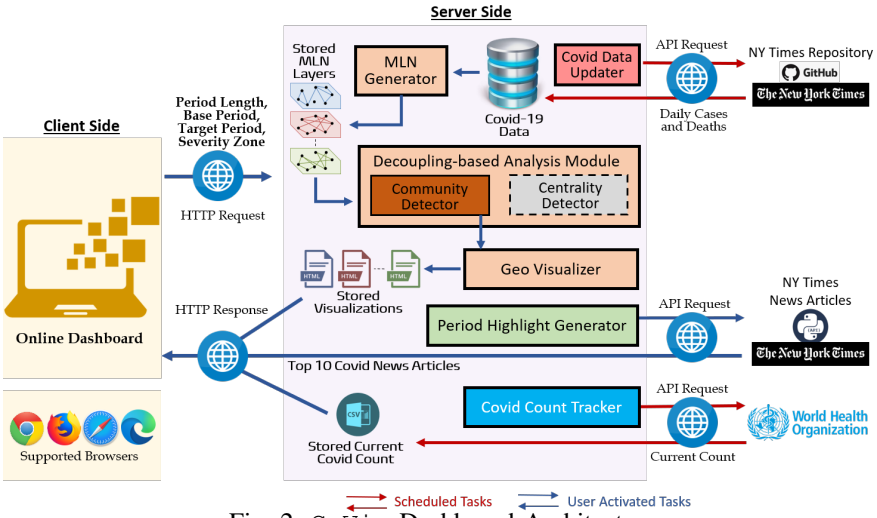


Fig. 2: CoWiz Dashboard Architecture

range) in the number of new cases. Also, the similarity among the counties based on the *static demographic information* like geographical proximity, population density, median household income and educational qualifications leads to the generation of other layers. For finding out the regions with similar effect in each layer, we have detected groups of similar counties using the community detection algorithm [11].

A. The Flexible Interactive Dashboard

The visualization system is completely automated – data collection, latest data from WHO, period highlights, and inputs. The system architecture is shown in Figure 2. Detailed description of the various components is provided below.

1) Intuitive User Interaction

The dashboard navigates the user for choosing the periods and the extent of spread that he/she wants to analyze and visualize without allowing the user to make mistakes.

- *First,* the user chooses the **length of the time periods** in terms of number of days - (1 to 30 days).
- *Then,* the user **selects the start dates of the base and the target periods**. The base period is the one whose new cases form the basis of comparison against the new cases in the target period. The periods (start to end date) are highlighted automatically based on the start dates chosen.
- *Finally*, the user chooses the desired **level of severity** for visualization. Regions are classified into **7 Severity levels/rates: Big Dip, DownTick, Decrease, Flat/Plateau, Increase, Uptick and Spike**, based on percentage increase in number of new cases between periods.

2) Map-Based Visualization with Zoom/Hover

Based on the parameters received, the task of the server side is to return geographical visualization on the US map.

- *First*, `Geo Visualizer` module checks whether required visualization already exists in server storage.
- *If visualization is not present*, then we check for the stored MLN layers that match the parameters received. If the layer is present, visualization is created.
- *If the layer is not present*, then the `MLN Generator` produces the required layer from the stored Covid-19 raw data set. This layer is **analyzed for communities** by the `Decoupling-based Analysis Module`. *Every community corresponds to one of the severity zones*. This result is finally fed to the `Geo Visualizer` that generates the final visualization.

The visualization generated is enabled with the **mouse hover, zoom-in, pan, and map download capability** through which detailed study of the regions is possible. On hover, the user can view county details like demographics and exact value of the percentage change in the number of new cases.

3) Period Highlights/Live updates from WHO

The `Period Highlight Generator` uses the New York Times article search API [5] to fetch **top 10 most relevant news headlines related to Covid in the target period**.

`Covid Data Tracker` is scheduled to execute every 6 hrs to fetch the *latest official* number of confirmed cases and deaths for the World and the US reported by the World Health Organization and stored on the server, which is displayed on the top bar.

4) Automated Daily Data Update

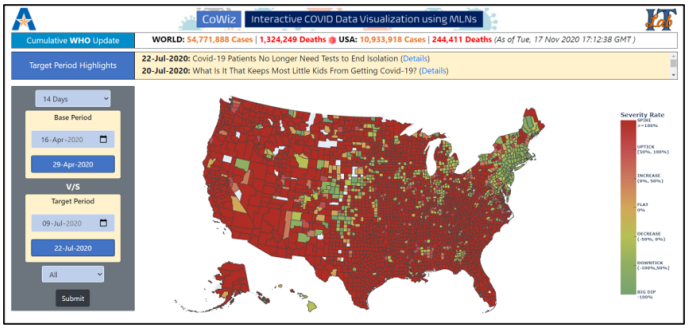
The `Covid Count Updater` is scheduled to execute every night at 1230am HST to fetch the day's number of cases and deaths for each US county from the ongoing NY Times repository [3] and updates the back-end Covid-19 data set. This keeps the dashboard **up-to-date** for analysis and visualization.

5) Hashing and Caching For Response Time

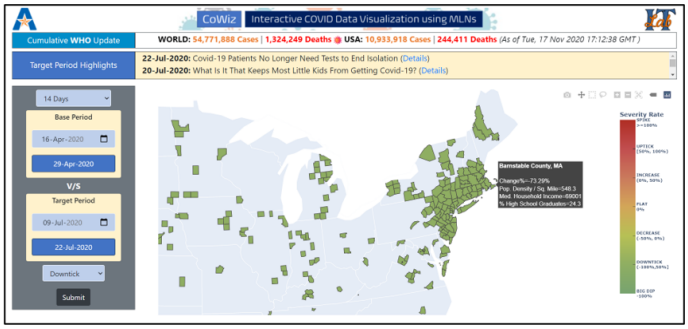
Based on user selections, the layers and map visualizations generated are **persisted** on the server with a handle to access from a main memory hash table. Cached files are associated with unique key constructed from user input which are looked up in the hash table. Multi-level lookup is used to avoid recomputation and to improve response time.

6) Implementation Summary

The `MLN Generator` uses C++. Louvain method is used for `Community Detection` in the `Analysis Module`. Python script extracts the `Period Highlights`. For `Covid Data Updater` and `Covid Count Tracker` modules, cron jobs periodically run shell scripts to retrieve the necessary .csv files from the web.



(a) **Zoomed Out** Viz.: Majority of US adversely effected; Experienced a surge in cases



(b) **Zoomed-In and Flexible** Visualization: Downtick in Northeastern US

Fig. 3: **(A6)**: Impact of July 4 long weekend (Jul 9 - Jul 22) compared to the Lockdown period (Apr 16 - Apr 29)

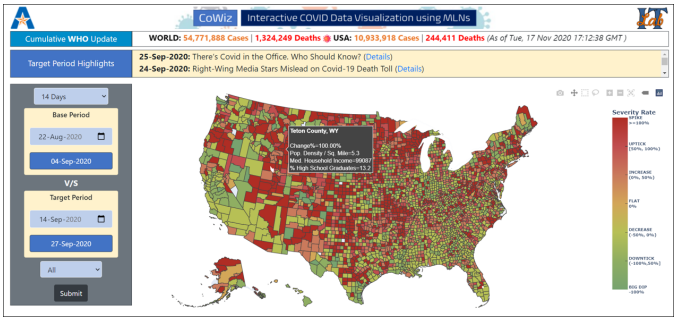


Fig. 4: **(A8)**: *Upward trend* in cases around Midwest US in period post Labor day weekend (Sep 14 - Sep 27)

`Geo Visualizer` uses *plotly* python library to generate US map plots and its *express* module gives the foundation to store relevant information and display on the map through the hover option. Flask web framework has been used as skeleton for the dashboard. Bootstrap front-end web framework has been used to build the various interactive components of the interface. JavaScript is used to enhance the user experience as they explore the dashboard and its features. The dashboard is hosted on a Apache HTTP Server 2.4.7 on Linux machine. It is supported on all major web browsers. For best user experience, screen sizes above 1200 pixels are recommended.

B. Including Complementary Data For Better Understanding

Currently, we are concentrating on *period-based MLN layers*. For addressing the analysis objectives of IV (see Section I), the *static demographic MLN layers* are generated where the counties are linked based on their similarity with respect to the feature specified. Moreover, availability of additional daily county-wise data like *number of hospitalizations*, *number of miles covered by vehicles*, *expenditure categories* and *average hourly density of people visiting public areas (movie theatres, grocery stores, recreation parks, train or bus stations, airports, schools)* will lead to other period-based layers. Analysis and composition algorithms are available in the `Decoupling-based Analysis Module` [14], [15].

IV. Use Cases

Based on client-side period selections, we are able to address **(A4)** to **(A8)**. For **(A8)**, figure 4 shows how the **Labor Day weekend led to spike in the cases around the Midwest region**. On the other hand, for **(A6)**, figure 3 (b) shows that through the ***zoom-in, pan and selective*** visualization features we are able to deduce that counties from **MA, NY and NJ showed a downward trend post the 4th of July long weekend**, even lesser than the April lockdown period!

V. Conclusions and Future Extensions

Our dashboard goes beyond basic statistical analysis of Covid-19 and is extensible using the underlying theory. Other data can be used interchangeably with minimal effort (e.g., accident data) As new data becomes available, we can visualize more complex analysis using the same dashboard. On the server-side, integration only requires increasing layers for analysis.

Acknowledgments

This work was partly supported by NSF Grant 1955798.

References

[1] The centre for disease control covid dashboard. https://covid.cdc.gov/covid-data-tracker/.
[2] Covid-19 surveillance dashboard by univ. of virginia. https://nssac.bii.virginia.edu/covid-19/dashboard/.
[3] Data from the new york times, based on reports from state and local health agencies. https://github.com/nytimes/covid-19-data.
[4] Johns hopkins university covid dashboard. https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6.
[5] The new york times article search api. https://developer.nytimes.com/docs/articlesearch-product/1/overview.
[6] The new york times covid dashboard. https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html.
[7] University of washington covid dashboard. https://hgis.uw.edu/virus/.
[8] Us census data. https://data.census.gov/cedsci/.
[9] The world health organization covid dashboard. https://covid19.who.int/.
[10] Worldometer covid statistics. https://www.worldometers.info/coronavirus/country/us/.
[11] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of community hierarchies in large networks. *CoRR*, abs/0803.0476, 2008.
[12] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter. Multilayer networks. *CoRR*, abs/1309.7233, 2013.
[13] Kanthi Komar, Abhishek Santra, Sanjukta Bhowmick, and Sharma Chakravarthy. Eer→mln: Eer approach for modeling, mapping, and analyzing complex data using multilayer networks (mlns). In *ER 2020*.
[14] A. Santra, S. Bhowmick, and S. Chakravarthy. Efficient community re-creation in multilayer networks using boolean operations. In *Int. Conf. on Computational Science, Zurich, Switzerland*, pages 58–67, 2017.
[15] A. Santra, S. Bhowmick, and S. Chakravarthy. Hubify: Efficient estimation of central entities across multiplex layer compositions. In *IEEE International Conference on Data Mining Workshops*, 2017.
[16] A. Santra, K. Komar, S. Bhowmick, and S. Chakravarthy. A new community definition for multilayer networks and a novel approach for its efficient computation. *arXiv preprint arXiv:2004.09625*, 2020.
[17] Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and S Yu Philip. A survey of heterogeneous information network analysis. *IEEE Trans. Knowl. Data Eng.*, 29(1):17–37, 2017.
[18] Xuan-Son Vu, Abhishek Santra, Sharma Chakravarthy, and Lili Jiang. Generic multilayer network data analysis with the fusion of content and structure. In *CICLing 2019, La Rochelle, France*, 2019.