PharmaSUG 2023 – Paper RW-060

Exploring the Spread of COVID-19 in the United States Using Unsupervised Graph-Based Machine Learning

Kostiantyn Drach, IST Austria / Intego Group, Klosterneuburg, AUSTRIA

Sergey Glushakov, Intego Group, Maitland, FL, USA

Iryna Kotenko, Intego Group, Kharkiv, UKRAINE

ABSTRACT

Real-world data can be essential for our understanding of clinical data, especially with the emergence of phenomena such as the COVID-19 outbreak. In this paper, we analyze how the spread of the virus has advanced across the U.S. during the initial phase of the pandemic using novel graph-based machine-learning techniques. First, a cloud of graphs is extracted from several publicly available datasets. In these graphs, each node corresponds to a single county (>3000 nodes per graph), whereby two counties are connected with an edge if they have similar patterns in the advance of the pandemic spread over a specific timeframe. A graph (or a subset of graphs) from the cloud with the most robust geometric properties is subsequently revealed. This constitutes a topological model of data. Next, unsupervised machine learning algorithms discover communities of nodes within the chosen graph relying on pure geometric properties of the model. Finally, the highlighted communities are compared to each other based on the real-world data employed by the model to explain dissimilarities between the communities. A variety of publicly available real-world data, including healthcare, social, demographic, economic, and geographic data, was used in the analysis. Our geometric, data-driven approach reveals insights that would otherwise have been difficult to identify through the implementation of standard statistical methods alone. The focus on topological properties helps to identify the underlying geometry of the dataset and to discover a set of unrelated features that may be causing the similarity in the spread of COVID-19 across the U.S.

KEYWORDS: real-world data, graph-based machine learning, topological data analysis, COVID-19, time series, unsupervised machine learning, community detection, open-source, clinical trials.

CONTENTS

INTRODUCTION 2
1. TOPOLOGY-BASED WORKFLOW FOR DATA ANALYSIS 2
1.1. Topology and data mining 3
1.2. Understanding clinical data using topology 3
1.3. A workflow for graph-based data analysis 6
1.4. Selection of the optimal graph 6
1.4.1. Defining the most representative graph using an optimality score 7
1.4.2. Defining the most representative graph using graph aggregation 9
1.4.3. Defining the most representative graph using graph stability 10
1.5. Community detection on graphs 10
1.5.1. The Girvan-Newman algorithm 11
1.5.2. The clique percolation method 12
2. THE COVID-19 EXPERIMENT 13
2.1. Experiment workflow 13
2.2. Sources of data used to define outcomes 13
2.3. Selection of observation time interval 15
2.4. Analysis of epidemic curves 16
2.5. Outcomes selected for the model 18
2.6. Defining the most representative graph 19

2.7. Predictors used for statistical analysis 20
3. FINDINGS EVALUATION AND INTERPRETATION 21
3.1. Girvan-Newman communities 21
3.2. Percolation-detected communities 33
REFERENCES 38
ACKNOWLEDGMENTS 39
CONTACT INFORMATION 39

INTRODUCTION

Topological Data Analysis, or TDA for short, is a modern approach in data analysis for discovering hidden patterns in large and complex datasets. Besides traditional analytics, such as geometric, statistical, and data-driven algorithms, TDA employs a wide range of machine learning (ML) techniques, both supervised and unsupervised by humans. In unsupervised ML, unlike supervised, only input data is used to identify a hidden structure of the dataset. Algorithms do not rely on prior training in order to discover and present previously unknown insights in the data. This is extremely important for real-world analytics when expected results may not be obvious, considering the volume and variety of data that is needed to be processed. Utilizing unsupervised ML such as TDA allows us to extract comprehensive topological data maps represented by graphs without first having to develop a hypothesis.

In this paper, we discuss a TDA-based workflow designed to analyze clinical and other types of data in an automated way. We will specify this approach to analyze how the spread of COVID-19 has advanced in every county of the USA and look for similarities in a specific timeframe. Focusing on the geometric properties of datasets, we unveil a set of unrelated features that could have caused similarities in the pandemic spread.

Researchers worldwide have constructed multiple data models related to the spread of COVID-19; in the current experiment, however, the analysis focuses on the early stage of the pandemic. The objective is to identify why in some regions of the USA the pandemic spread much faster after the first case had been confirmed than in the others where the transmission was much slower. Healthcare, social, economic, demographic, geographic, and other factors, which could influence similarities in pandemic spread, are addressed.

The analysis was performed on a county-by-county basis, yielding a topological data model in form of a graph in which each of 3,142 nodes corresponds to one county, and two nodes are connected if they share similarities. The graph was built using TDA. The dataset incorporated a variety of outcomes corresponding to the number of confirmed cases and deaths in each county over a specified time interval.

The key focus of the experiment is to investigate the spread of the pandemic since its outset. Thus, the first confirmed and recorded case of COVID-19 in each county was taken as the starting point of the observation interval. Given the speed with which the pandemic has advanced, it was critical not only to select an appropriate starting point but also to limit the analysis by carefully selecting the endpoint of the time interval to make it relevant without overloading the model.

After the graph was built, real-world data were used to integrate into the model any predictors which might be responsible for similarities in the early stage spread of the pandemic. Over 250 predictors from different publicly available sources were used during the course of the experiment. Further, we performed a statistical analysis of discovered patterns to explain similarities in the spread of the pandemic based on the predictors integrated into the model. At any time, additional predictors can be added into the model to expand the search of unrelated features that might be responsible for similarities in the pandemic spread.

1. TOPOLOGY-BASED WORKFLOW FOR DATA ANALYSIS

Topological data analysis (TDA) is a novel approach of building a visual representation of a complex dataset. This analysis allows the extraction of comprehensive graphs from a dataset to provide a compressed graphical representation of a multidimensional set of interrelated outcomes. When applied to clinical data, this graph consists of nodes corresponding to patients participating in a clinical study and edges connecting those who share similarities. In this section, we present a general introduction to TDA in clinical trials. This general approach will be specified in Section 2 and applied to the COVID data of the pandemic spread in the USA.

1.1. Topology and data mining

Topology is a field of mathematics that deals with the properties of objects that remain invariant under continuous deformation. Imagine a surface that is made of a very thin and elastic material. The surface can be bent, stretched, or crumpled in any way; however, it cannot be torn and its parts cannot be glued together. As the surface is deformed, it changes in many ways, but some properties remain the same. The idea underpinning topology is that some geometric properties depend not on the exact shape of an object but, rather, on how its parts are combined.

As a simple example, consider geometric figures on the plane representing the numerical digits 0, 1, 2, … 9. For a topologist, various representations of the digit 0 are equivalent since they can all be continuously transformed into each other without cutting or gluing (see Figure 1 a-d). It is possible to change the size, thickness, or slope of the digit 0 through continuous deformation; however, one property remains invariant: the object separates the plane into two regions, namely the interior and the exterior. At the same time, 0 is not topologically equivalent to 1 or 8: 1 does not encircle a region and 8 contains two holes (see Fig. 1e). The topological classification of the digits 0, 1, 2, … 9 results in the following five classes:

{0}, {1, 2, 3, 5, 7}, {4}, {6, 9}, {8}.

The digits in any of the classes are topologically identical, but no two digits taken from distinct classes are equivalent from the topological point of view.

The number of holes in a geometric object is a basic topological property. Another significant property is connectedness. Intuitively, an object is connected if it consists of a single piece. For example, the curve representing 0 is connected; if any two points are removed from it, it will become disconnected. Pieces of a disconnected object that are themselves connected are referred to as connected components. In the mathematical study of topology, all of these intuitive concepts are examined on a rigorous basis and generalized to higher dimensions.

a) b) c) d) e)

Figure 1. Different representations of the digit 0 (a-d) are topologically equivalent. All share a common topological property: they divide the plane into an interior region and an exterior region. The digit e) is not equivalent to 0 since it encloses two internal regions

Topology deals with abstract mathematical entities, such as curves and surfaces, that consist of an infinite number of points. In practice, however, all datasets are necessarily finite. Recently, a new field has emerged at the crossroads of topology and data science. TDA aims to extract topological data, that is, qualitative information, from finite sets of data points. It involves exploring datasets (viewed as finite clouds of points in multidimensional space) at multiple scales or resolutions, from fine- to coarse-grained. Given a complex dataset, TDA can be used to extrapolate the underlying topology and build a compressed yet comprehensive topological summary of the dataset. TDA exploits various methods and algorithms stemming from computational topology and geometry, statistics, and data mining. For detailed expositions of the mathematical theories that underpin TDA and certain applications in biology, see [1, 2] and the references therein.

1.2. Understanding clinical data using topology

Topology was originally developed to distinguish between the qualitative properties of geometric objects. It can be used in conjunction with the usual data-analytic tools for the following tasks:

1. **Characterization and classification.** Topological features succinctly express qualitative characteristics. In particular, the number of connected components of an object is of importance for classification.
2. **Integration and simplification.** Topology is focused on global properties. From the topological perspective, a straight line and a circle are locally indistinguishable; however, they are not equivalent if they are considered as a whole. Topology offers a toolbox to integrate local information about an object into a global summary. Thus, topology can provide the researcher with a natural "big-picture" view of complex, multidimensional data.
3. **Features extraction.** Topological properties are stable. The number of components or holes is likely to persist under small perturbations or measurement errors. This is essential in data mining applications because real data is always noisy.

In the context of clinical research, the dataset under study is typically a table of outcomes in a particular clinical trial or study. The table rows correspond to individual participants in the clinical trial, and the columns contain information on specific outcome measures of interest, such as lab tests, vitals, questionnaires, etc. Given a table of clinical outcomes, the following parameters are required to generate a graph using TDA:

1. A *distance function* as a similarity/dissimilarity measure between patients (i.e., similarity between the rows of the table of outcomes). Patients with similar outcomes, e.g., with sufficiently small distance, are connected with an edge. In application, this is done based on the projection:
2. A *projection function* that is chosen to capture topological features of interest for the dataset by stratifying patients into certain subpopulations (bins).
3. A *projection specification*, which includes:
    - the number of stratifying bins,
    - quantitative overlap of the bins,
    - a threshold for the value of the distance function below which the nodes are connected with an edge within a stratifying bin (i.e., a threshold below which the patients "have similar outcomes").

Graph nodes representing similar patients (in terms of a predefined sequence of clinical outcomes) are connected with edges if they have similar outcomes within each projection bin (measured by distance function).

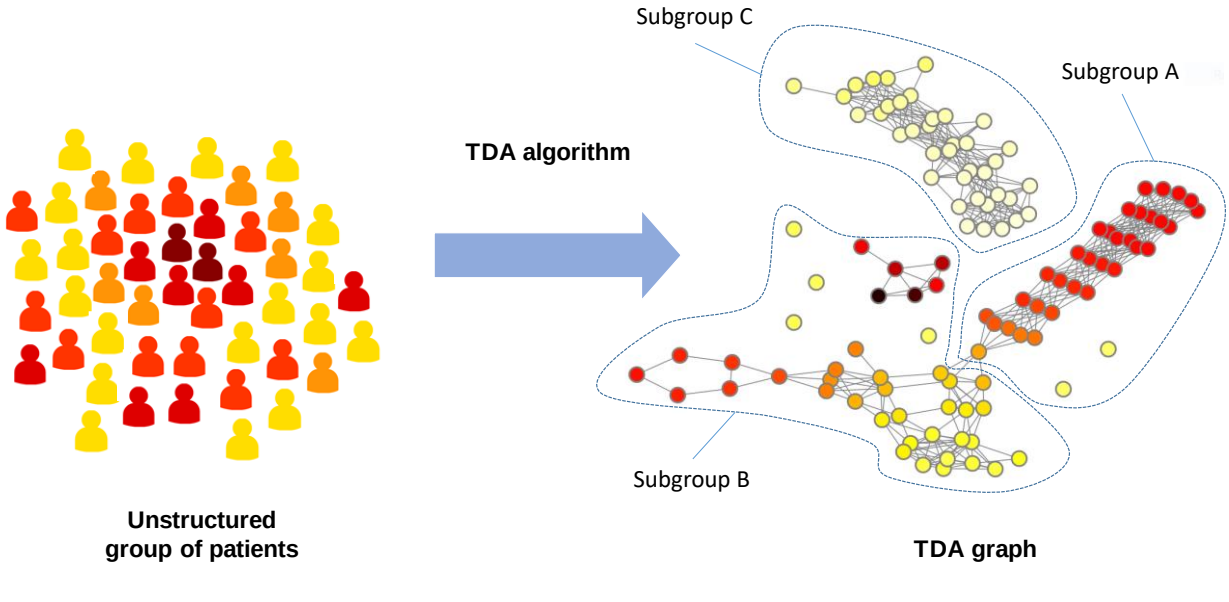


Figure 2. Discovery of multivariate patterns in clinical trial outcomes. The graph represents groups of patients structured according to the similarity of their outcomes

The core idea of clinical data mining using TDA relies on the visual discovery of subgroups of related patients in a graph (see Figure 2) that presents the relevant information about the dataset in a compact and efficient manner. For the clinical dataset, the following criteria have to be met to perform the analysis:

- **Each node represents a patient**: a graph extracted from a clinical dataset is actually a graphical representation of the dataset in which each node represents an individual (trial subject).

- **Similar nodes are connected**: two nodes representing similar patients (in terms of a predefined set of clinical outcomes) are connected with an edge.
- **Coloring focused on specific outcomes**: the color of the nodes helps to highlight emerging patterns in the data and identify subgroups of patients related to the distribution of a variable of interest.
- **Discovery of subgroups**: communities of nodes on a graph reflect a segmentation of patients that may indicate robust patterns within the data.

TDA was successfully applied in the context of clinical studies (see, e.g., [3, 4]).

To be considered for further analysis, a graph extracted from the dataset using TDA algorithms should meet certain requirements. Namely, it should:

- accurately represent the original dataset;
- eliminate the features of the dataset that are not relevant to the purpose of the study;
- reduce the complexity of the features that are shown on the data map; and
- be insensitive and robust to small noise, such as errors of measurement, or missing data.

For illustrative purposes, let us consider a simple two-dimensional dataset with the data points arranged in a "zero-like" shape.

In order to show the robustness of the topological approach, some data points from the dataset were intentionally omitted at random, and additional graphs were built for the modified datasets whereby 50% and 90% of the original data points were missing (see Figure 3).

The graphs show certain geometrical stability even in the case of 90% missingness. The shape of the graphs built on the remaining data points is structurally similar to the shape of the graph corresponding to the complete dataset. Therefore, in this example, graphs representing a relatively small portion of the data still have a similar shape to the graph representing a complete dataset.

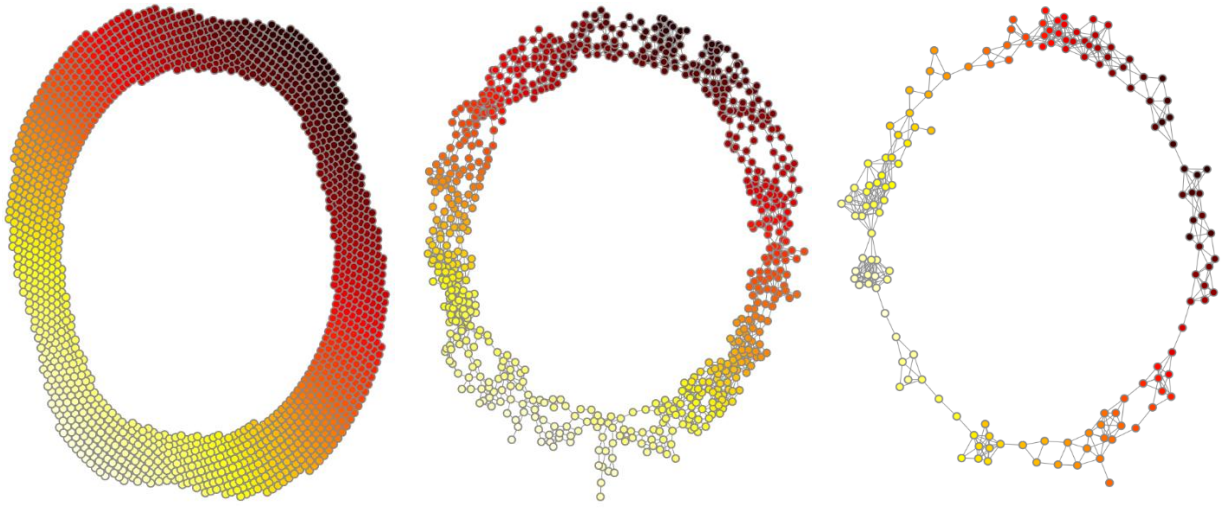


Figure 3. Graphs representing a dataset with varying proportions of randomly missing data. Graphs produced by the TDA algorithm for a complete dataset (left panel) and datasets where 50% and 90% of the data points are missing at random (middle and right panels, respectively). This example illustrates that even with 90% of the data missing, the cyclic shape of the dataset is preserved in the corresponding graph.

1.3. A workflow for graph-based data analysis

Topological Data Analysis is used to create a flexible and versatile workflow to perform graph-based data analysis. This workflow can be adapted to a variety of scenarios and types of data in order to identify hidden patterns. The key steps are summarized and highlighted in Figure 4. We will see their implementation in our experiment in Section 2. All of the steps in the workflow except Steps 1, 5, 8, are performed automatically using machine-leaning algorithms.

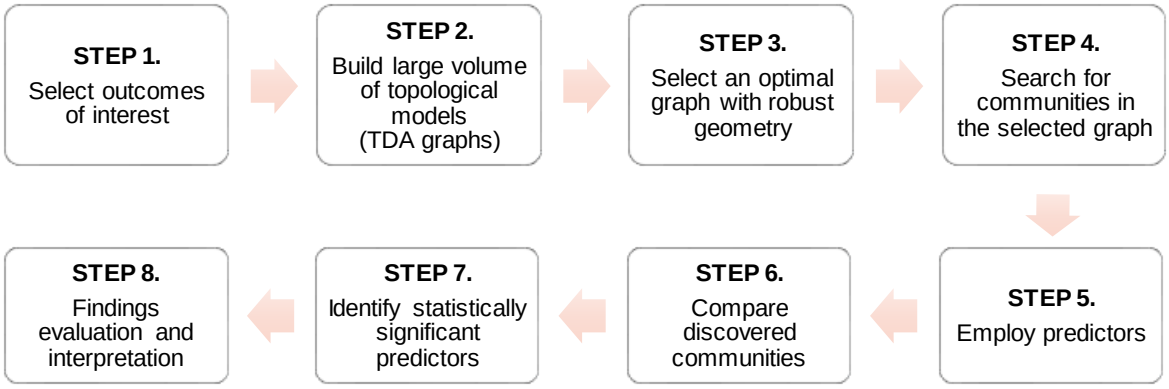


Figure 4. Workflow of graph-based data analysis

Let us expand on the steps in our workflow:

STEP 1. From a given dataset, the outcomes of interest are selected. At this step, some pre-processing of data might be required to deal with irregularity, e.g., to account for missing data, to aggregate noisy data, etc.

STEP 2. Using the selected outcomes of interest, a large volume of TDA-graphs is built by varying parameters of the TDA algorithm, e.g., the parameters in the distance function, in the projection function, etc. (see Section 1.2).

STEP 3. The most robust and representative graph is selected based on an array of criteria, e.g., adapter modularity score, Kolmogorov complexity, etc. (see Section 1.4). In many applications, the most representative graph is selected by the majority vote or the cumulative ranking among the optimality scores.

STEP 4. A selection of community detection algorithms is applied to the most representative graph at this step to reveal hidden patterns within data in form of communities on the graph (see Section 1.5). The discovered communities are highlighted on the graph by coloring and are subject to further analysis.

STEP 5. A selection of predictors of interest is integrated into the model to explain the detected communities, and hence to explain hidden similarities within the dataset of study. At any time, additional predictors can be incorporated into the model at this step to expand the search of unrelated features.

STEP 6. Communities on the graph correspond to subsets of patients. A comparison of communities is performed at this step, e.g., by comparing sizes, overlap, persistence over different community detection methods, etc.

STEP 7. Further pairwise or community-against-the-rest comparison of communities is done at this step using statistical analysis based on predictors. Statistically significant predictors are selected. This step helps to identify the key variables that are driving the community structure and involves a large volume of automatic statistical tests.

STEP 8. At this final step, the statistically significant predictors of the discovered community structure are being further interpreted, e.g., using subject-matter expertise.

1.4. Selection of the optimal graph

When applying the TDA algorithm to a complex dataset based on various parameters, the algorithm generates a large volume of graphs, including those that capture non-relevant noise. The important step is to determine the most representative graph based on available parameters of the topological model (the number and placement of stratifying bins, parameters of the projection, etc.). There is no strict definition, but usually by a representative graph we mean one that

is well structured with well-recognized geometric features: dense clusters, flares, loops, or other patterns that may indicate the robust geometric patterns in the data. The other graphs we refer to as dull graphs.

In this subsection, we briefly discuss our approaches to tackling this problem. It should be noted that usually one should select a number of candidates to be the "best graph" and then look through all of them. The ways to define the most representative graph may include:

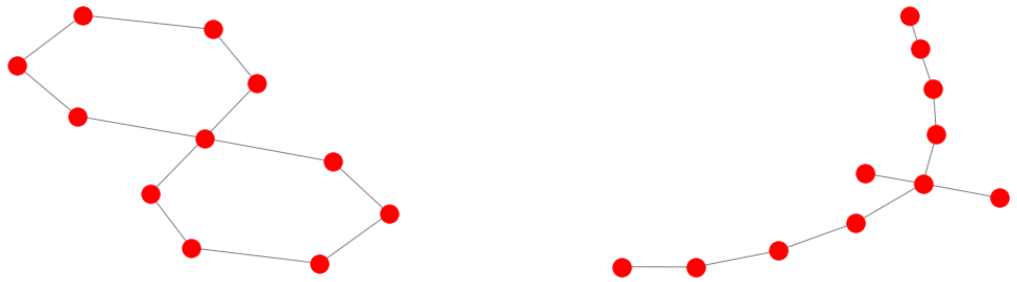
- finding the graph that has the larger optimality score;
- aggregating certain family of graphs;
- finding the most stable graph.

It should be noted the above approaches are not mutually exclusive. They both intersect (e.g., an optimal score could be a measure of graph stability or aggregation of graphs could yield a stable graph) and complement each other.

1.4.1. Defining the most representative graph using an optimality score

A. Null-model based modularity

One of the scores that measures the quality of a partition of the graph into communities is the classical score of Newman's modularity [5]. Its idea is to evaluate how far the graph with a given partition of nodes is from some null-model graph, where the null-model graph is a random graph with the same number of nodes and the same given degrees of vertices. This idea can be generalized to define some other "null-model based" modularities as measures of "well-structuredness" (i.e. "irrandomness") of the graph. The null-models of random graph we use in experiments take into consideration the partition of the data cloud into stratifying bins. Such random graphs are considered dull. We assume the graph with the largest null-model based modularity (i.e., the largest difference between the graph under consideration and a random graph in some null-model) to be the most well-structured and hence the most representative.



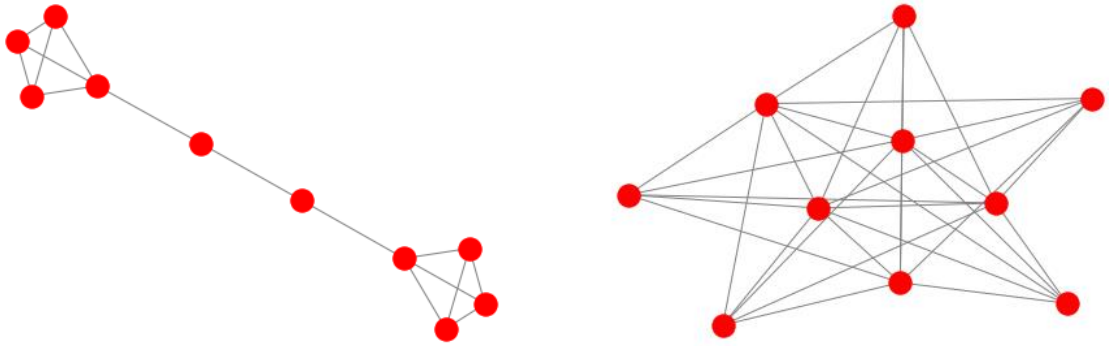
i. a well-structured graph having high linear covering modularity (LCM = 0.38)

ii. a random graph having low linear covering modularity (LCM = 0)

Figure 5. Linear covering modularity

Our null-model based modularities include:

- *linear covering modularity* (LCM), where the null-model is a random graph with the same list of edges connecting neighboring stratifying bins (see Figure 5);
- *generative modularity*, where the null-model is a random graph with the same distribution by the stratifying bins;
- *multidimensional modularity*, which is Newman modularity normalized by some nodes' similarity measure, based on the number of occurrences in the same stratifying bins;
- *random walk-based modularity* (RWM), and its variations (e.g. *mixing rate* and *time*), where the null-model is a graph with uniformly distributed random walks (see Figure 6).



a) a well-structured graph having high Random walk-based modularity (RWM = 40.48)

b) a random graph having low Random walk-based modularity (RWM = 2.31)

Figure 6. Random walk-based modularity

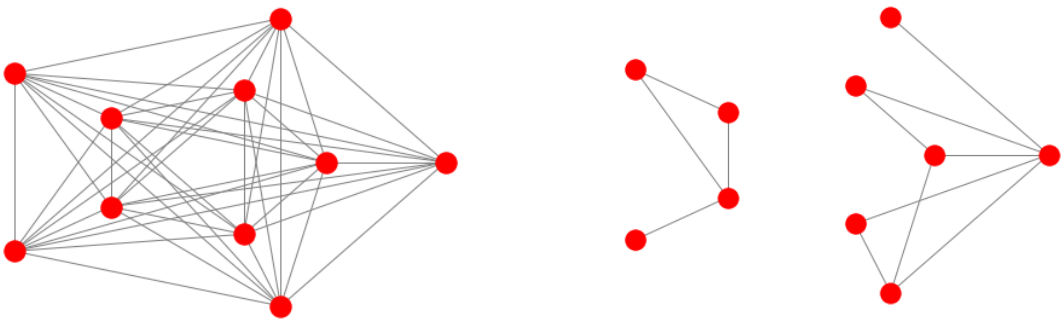
B. Kolmogorov complexity

Kolmogorov complexity originated in theoretical computer science and is a measure of information contained in a string or an arbitrary array of letters and digits. It measures how well an object (array, graph, text, etc.) can be compressed.

For a given string (array), the Kolmogorov complexity is determined as the smallest length of the compressed description of the string (array). For example, the string "11111111111111111111" can have the following compressed description: "1 × 20 times". However, in the string "10011011110101011110" the digits seem to be spread randomly and it is difficult to give a shorter description of the string than a direct list of the digits within it. Therefore, the Kolmogorov complexity of the second string is higher than that of the first string.

The measure of information contained in graphs can be similarly estimated by using Kolmogorov complexity. Specifically, when the nodes of a graph are numbered, it can be described by a graph adjacency matrix, in which "1" is placed at the intersection of row i and column j if the nodes numbered i and j are connected, and "0" is placed at the intersection of row i and column j if the nodes numbered i and j are not connected.

Graphs with a regular structure of edges have low Kolmogorov complexity, while graphs with a random structure of edges have high Kolmogorov complexity. Figure 7 a) shows a complete regular graph with 10 nodes that has a Kolmogorov complexity of 103.79, while b) shows a random graph with 10 nodes that has a Kolmogorov complexity of 437.61:



a) a regular graph having low Kolmogorov complexity (K.C. = 103.79)

b) a random graph having high Kolmogorov complexity (K.C. = 437.61)

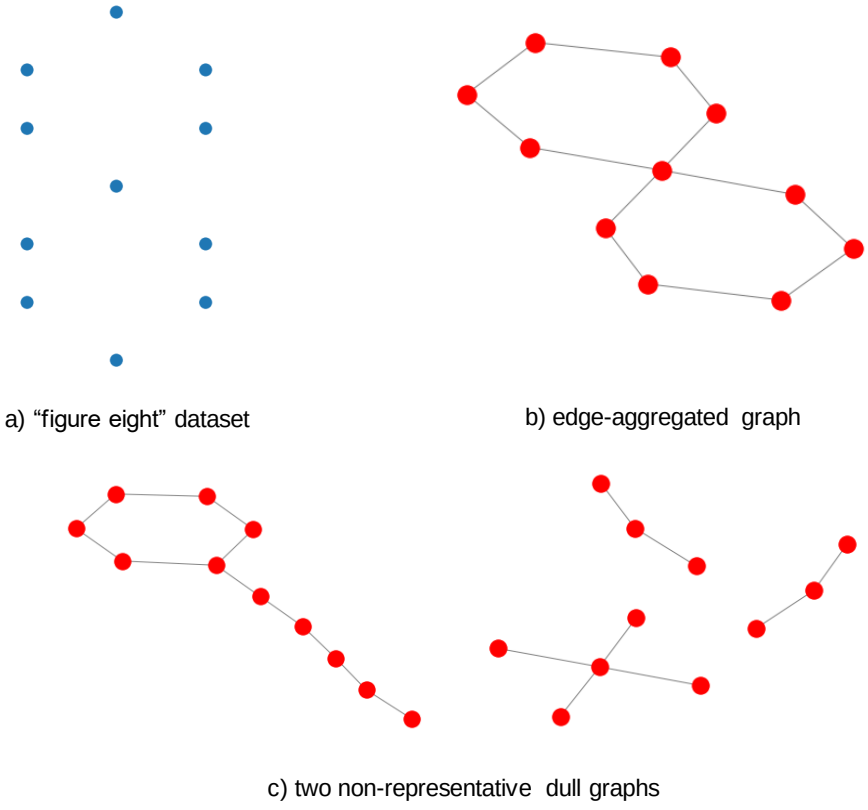
Figure 7. Kolmogorov complexity

Although it is known that the Kolmogorov complexity cannot be calculated in a general case, there are various methods for estimating it; a bit loosely, the term "Kolmogorov complexity" is still used when referring to these estimates. An example is the block decomposition method, which includes the decomposition of a string (array) into blocks having a limited length, estimation of the Kolmogorov complexity of each block, and summing up the estimates according to the information theory rules [6]. Experiments have shown that in graphs with low Kolmogorov complexity, nodes within the same stratifying bin tend to combine into groups (clusters) with a high density of edges. Conversely, graphs with high Kolmogorov complexity have a more uniform distribution of edges; that is, the selected parameters of the projection in such graphs distribute the nodes over the stratifying bins more evenly. Thus, in graphs having high Kolmogorov complexity, the nodes within the stratifying bins tend to group into clusters of approximately equal size.

In the current experiment, when the Kolmogorov complexity of the graph was estimated, consideration was given to partitioning the nodes of the graph into stratifying bins and the number of nodes in each bin, as well as to partitioning the graph into separately connected subgraphs and the number of connected subgraphs. The Kolmogorov complexity was measured within each stratifying bin separately. It was assumed that high Kolmogorov complexity implies a random distribution of edges in subgraphs of the graph, which in turn makes the influence of the node-clustering algorithm less significant than the influence of the projection, metric, and cover parameters.

1.4.2. Defining the most representative graph using graph aggregation

If some family of graphs is constructed for a given dataset and a number of predefined parameters, the aggregated graph can be built from this family. The nodes of aggregated graph correspond to data points from the dataset. The edge between vertices u and v is constructed based on the frequency of appearance of edges between u and v in the given family of graphs. The resulting graph is called an edge-aggregated graph (see Figure 8).



a) "figure eight" dataset

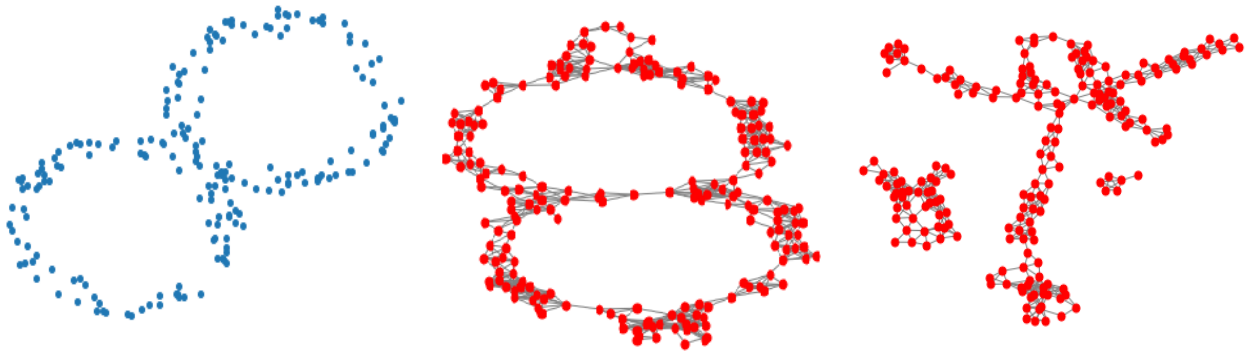
b) edge-aggregated graph

c) two non-representative dull graphs

Figure 8. Aggregating graphs for the "figure eight"

1.4.3. Defining the most representative graph using graph stability

The topological model constructed using TDA should reveal robust properties of the dataset, and hence can be searched based on stability properties of graphs. Given a topological model G, one way to understand its stability is to construct a "derivative" graph using the matrix of graph distances between all pairs of vertices of G, with and without taking the weights of the edges into account. This results in a pair of derivative graphs DG and DG_{weight}. Recall that the graph distance between a pair of vertices u and v in a graph is the combinatorial length (i.e., the number of edges) of the shortest path along edges that connects u and v. The graph distance can also take into account weights of edges. The smaller the difference between DG and DG_{weight}, the more stable the original graph G (see Figure 9).



a) the original point cloud

b) the topological model with small distance between its derivative graphs ($|DG - DG_{weight}| = 24.04$)

c) the topological model with large distance between the derivative graphs ($|DG - DG_{weight}| = 63.84$)

Figure 9. Selection of the optimal topological model based on graph stability

1.5. Community detection on graphs

In many problems, it is convenient to represent data as a point cloud in a multidimensional space. The points in the point cloud often tend to form dense subgroups. In these subgroups the points are closer to each other than to the points from the rest of the point cloud. We call these dense subgroups *patterns* as they help to identify and exploit relationships of interest in the dataset [7]. Searching for robust geometric patterns becomes one of the most important tasks in data analysis. In this section, we consider community detection methods as the way to search for such patterns.

The modern approach to data science frequently employs graphs to enhance understanding of complex systems. A variety of problems can be represented and studied using graphs as described in previous subsections. The key feature of a graph is a community structure, which relates to the way the nodes are organized in communities. Specifically, many edges connect nodes within the same community, while comparably few edges connect nodes between different communities [8, 9]. These communities can be considered to represent independent structures within the graph, and the detection of those independent communities is one of the key goals in the analysis of large graphs.

In graphs that represent real-world systems or data gathered in a study, the distribution of edges over subgroups of nodes is usually non-uniform. This reflects the possible presence of a hidden structure and patterns in the graph and, hence, in the data based on which the graph was created. Specifically, some groups of nodes may have high concentrations of intra-edges, while the concentrations of inter-edges between these groups of nodes may be low. The groups of densely connected nodes are referred to as communities. Figure 10 illustrates an example of a community structure within the graph that contains three groups of nodes (vertices) with dense internal connections within each group and comparably fewer connections between groups.

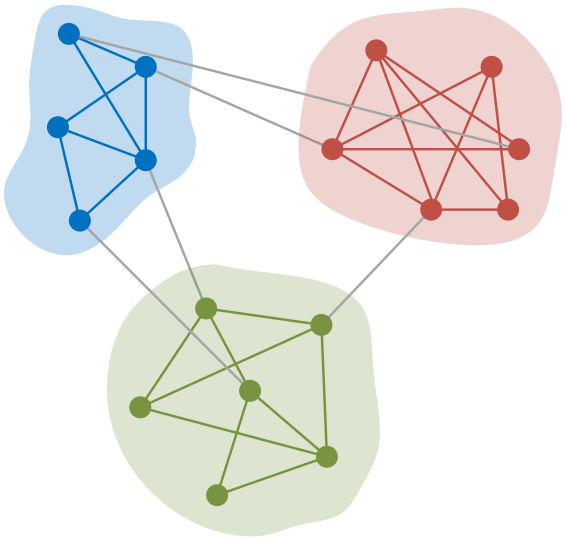


Figure 10. The schematic representation of the simple graph that has a community structure. The graph contains three communities of densely connected nodes with a much lower density of connections (gray edges) between the communities.

A number of algorithms have been developed for community detection (see [8, 9] for a survey), in particular:

- hierarchical methods, which build some dendrogram of communities by merging or splitting them, e.g., the **Girvan-Newman algorithm** described in subsection 1.5.1. Such dendrograms can be cut at some level to communities; the number of communities depends on the cut;
- methods based on the specific nature of communities, e.g., methods for maximizing so-called *modularity*, assuming that communities should have structure different in some sense from that of a random graph;
- propagation methods assuming that some structure (like a clique in the **clique percolation method**, described in subsection 1.5.2, or the most frequent neighbors label as in the label propagation method) "propagates" through the graph;
- random walk methods which are based on the assumption that the random walker spends more time inside a community and change communities with low probability.

In this paper, we give emphasis to the Girvan-Newman algorithm and the clique percolation method (with modifications made to these methods by the authors, see [10]) as the most efficient and most popular methods when applied for analyzing data (see also [7] for a comparison between community detection on TDA graphs and clustering analysis).

1.5.1. The Girvan-Newman algorithm

The Girvan-Newman algorithm [11] attempts to identify the edges that are located "between" some pairs of nodes in the graph. In the algorithm, the distance between all pairs of nodes, i.e., the shortest edge-based path, is calculated. Such paths define the edge betweenness characteristic of the edges. The edge betweenness characteristic of an edge is the number of shortest paths between pairs of nodes that run along the edge.

The method of community detection using the Girvan-Newman algorithm is based on calculating the edge betweenness characteristic for all edges in the graph. The method includes steps of removing the edge having the highest edge betweenness characteristic and recalculating the edge betweenness characteristic for all edges affected by the removal. The steps are repeated until no edges remain. The edges that have the highest edge betweenness characteristic are the most "loaded" and, hence, are considered to lie the most "between" communities. The removal of the edges having the highest values of the edge betweenness characteristic from the graph results in the nodes falling into communities. The removal of the edges that have the further highest values of the edge betweenness characteristic separates further communities within the graph. See the main steps in Figure 11.

The Girvan-Newman algorithm has been widely applied to a variety of graphs, e.g., graphs of human and animal social networks, metabolic graphs, gene graphs, graphs representing collaborations between scientists and musicians, and so forth. However, this algorithm is computationally intensive and takes $O(m^2n)$ times for a graph with m edges and n nodes. In view of the large amount of time required to perform the calculations, the use of the algorithm is limited to graphs that contain less than a few thousand nodes. Furthermore, the algorithm does not show how many edges need to be removed to provide the most optimal community detection.

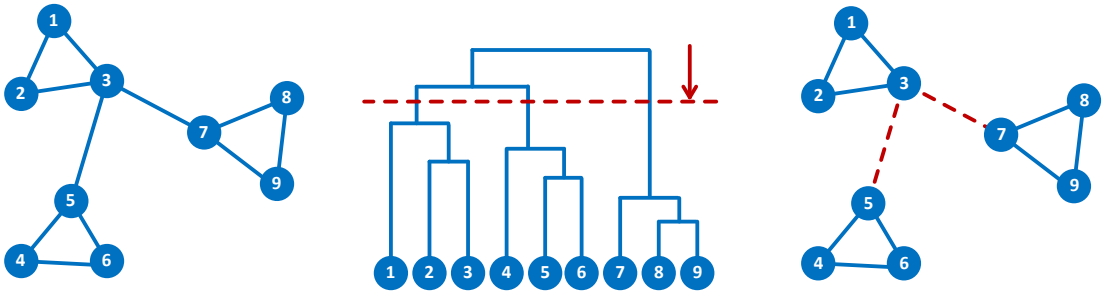


Figure 11. A hierarchical decomposition of the graph.
As one moves down the dendrogram, the detailed partitioning into communities starts to appear.

1.5.2. The clique percolation method

The clique percolation method [12] operates based on the assumption that internal edges within a community form *k*-cliques (i.e., subgraphs with *k* nodes in which every pair of nodes is connected by an edge) and edges that lie between the communities are not likely to form cliques.

The use of this method is based on the assumption that if a clique can "move" in the graph, the clique will get trapped inside the community and will not manage to pass in between two communities due to a lack of connecting paths. In this method, one clique can be "moved" to another if they share all but one node, and a community is defined as a maximal connected subgraph of the original graph so that each node in this graph belongs to some *k*-clique that lies entirely in the subgraph. The classical clique percolation method receives a value of *k* as an input and produces the list of all possible communities (as described above for the given value of *k*) as an output. The peculiarities of the method include the ability of some nodes to belong to several communities as several *k*-cliques may pass through these nodes and the ability of some nodes to occur out of communities as no *k*-clique contains them.

The example of the 3-clique percolation method can be seen at Figure 12: the 3-clique cannot pass through node 3, but the blue community and the green community overlap in node 3.

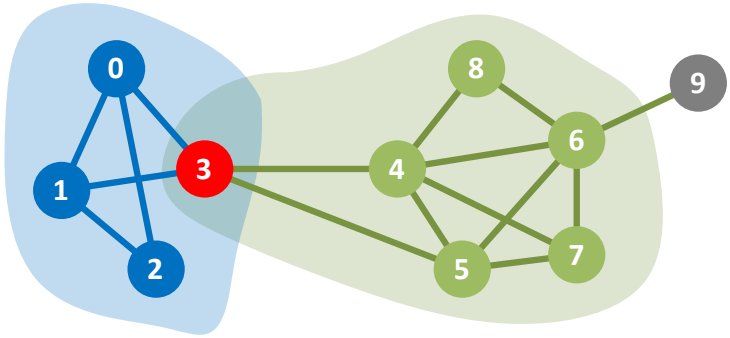


Figure 12. Example of overlapping community detection by the 3-clique percolation on a simple graph

Although theoretically the clique percolation method is computationally intensive because the detection of maximal cliques requires processing time that runs exponentially to the size of the graph, it was shown by the practical applications of this algorithm to real-world systems that this method works reasonably fast due to a limited number of cliques in real-world-based graphs.

2. THE COVID-19 EXPERIMENT

The goal of the experiment is to find similarities in the spread of COVID-19 since the beginning of the pandemic in February 2020. The start of the pandemic is defined by the date when the first confirmed cases were reported in the United States. The analysis is performed on a county-by-county basis, which will yield a topological model in the form of a graph in which every node corresponds to one county (total 3,142 nodes), and two nodes are connected if they share similarities. Focusing on the geometric properties of datasets, the experiment aimed to unveil a set of unrelated features that could have caused similarities in the spread of the pandemic. A representative graph will automatically highlight possible similarities in pandemic spread patterns and discover meaningful subgroups within the data as communities on the graph. Further, by integrating various predictors from different areas, such as demographic, social, geographic, etc., the revealed similarities will be described.

2.1. Experiment workflow

In this subsection, we show how to specify the versatile TDA workflow presented in Figure 4 for exploration of the advance of COVID-19 in the United States at the initial stages of the pandemic. Table 1 illustrates the order of steps in the experimental set-up.

Table 1. Phases in the experimental set-up

1) Define outcomes	Integrate the number of confirmed cases and deaths in every county of the United States over a specified timeframe since the beginning of the pandemic
2) Build the topological model (graph)	Based on defined outcomes and using TDA, build a large volume of topological models (TDA graphs). Each node in these graphs corresponds to a single county and encodes its number of confirmed cases, number of deaths, and a piece of the epidemic curve within the chosen timeframe
3) Select the graph	Select an optimal graph with robust geometry revealing the geometric properties of the dataset and hidden interdependencies. The choice is based upon one or more methods described in Section 1.4.
4) Search for communities	Perform an automatic search for communities by Girvan-Newman and percolation algorithms
5) Employ predictors	Integrate real-world data (from a variety of publicly-available sources) which may influence similarities in the spread of the pandemic. (The number of predictors can be expanded at any time after the model is built)
6) Compare discovered communities	Pairwise comparison of identified communities by outcomes and predictors, hypotheses
7) Identify statistically significant predictors	Perform a statistical analysis of discovered patterns to explain similarities in the spread of the pandemic based on the predictors integrated into the model
8) Findings evaluation and interpretation	Results interpretation, suggesting explanations and drawing conclusions

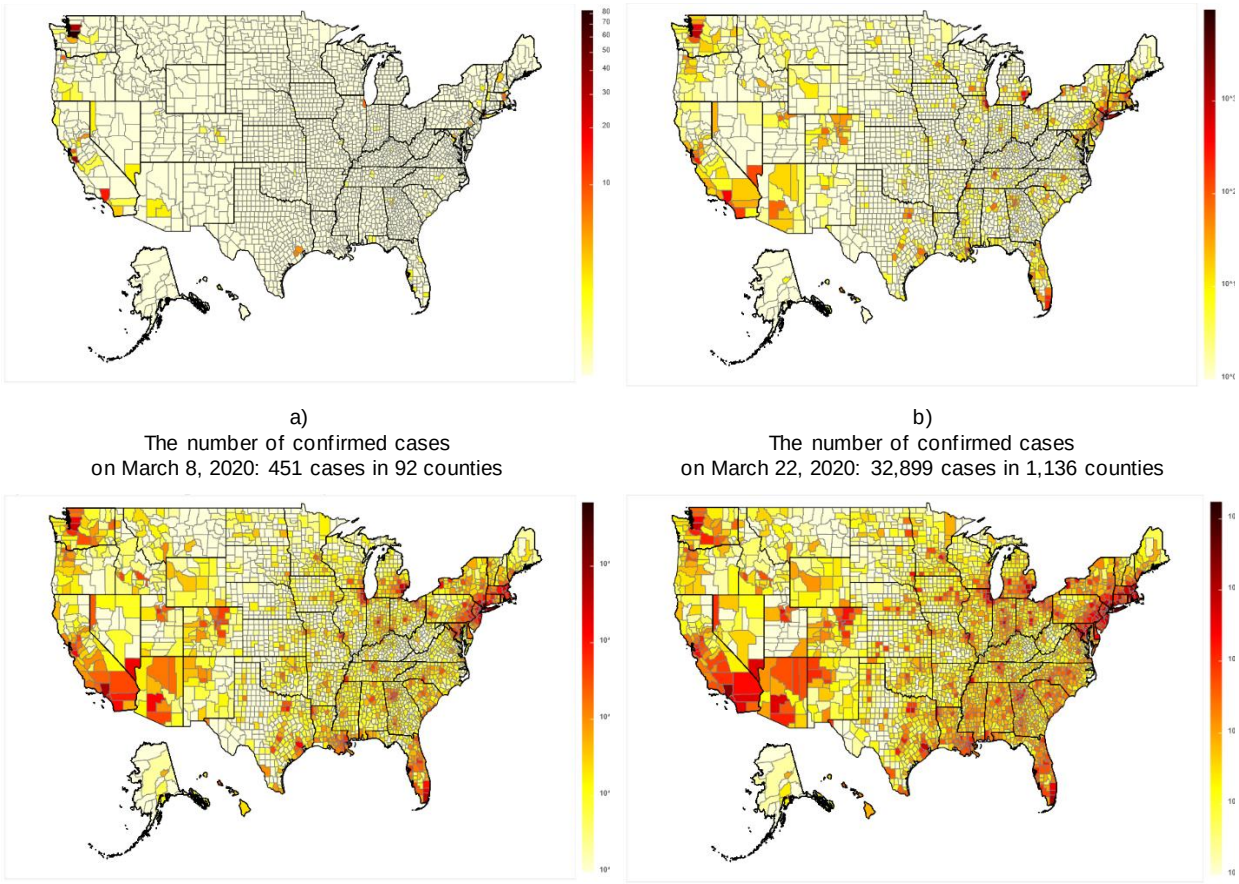
2.2. Sources of data used to define outcomes

Since the outbreak of the COVID-19 pandemic, multiple organizations and authorities across the globe have continuously collected statistical data related to the occurrence and spread of the disease and, as a result, accumulated a large volume of complex real-world data. In this paper, we use data obtained from an open-source data repository

available at Github and named *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University* [13]. Collected data include aggregated data sources such as the World Health Organization, European Centre for Disease Prevention and Control, US Centers for Disease Control and Prevention, etc. In this dataset, data related to US statistics are represented at state or county/city level, including data from public health authorities all over the US states and counties, while non-US data sources are aggregated at country/region or province level. The current research focuses on data which corresponds to the location and number of confirmed COVID-19 cases and deaths in all affected counties of the US.

The data used to set up the outcomes for the model were collected at the level of counties (administrative or political subdivisions of a state in the United States) and county equivalents (other functionally equivalent subdivisions under US jurisdiction). In total, the resulting dataset includes data related to 3,142 counties and county equivalents. The selected dataset includes the number of confirmed COVID-19 cases and the number of deaths reported by each county.

Data collection started on January 22, 2020, when the first confirmed case of COVID-19 was reported in King County, Washington State, and has since been continuously undertaken. Figure 13 illustrates the rapidity of the spread of the pandemic across the United States during a short period. There were only 451 confirmed cases on the 47th day (March 8) in 92 counties. Just 14 days later, on March 22, there were 32,899 cases, with 1,136 counties affected. Snapshots c) and d) illustrate how fast the pandemic advanced thereafter, reaching 330,384 and 1,151,933 cases on April 5 and May 3, respectively.



a)
The number of confirmed cases
on March 8, 2020: 451 cases in 92 counties

b)
The number of confirmed cases
on March 22, 2020: 32,899 cases in 1,136 counties

c)
The number of confirmed cases
on April 5, 2020: 330,384 cases in 2,406 counties

d)
The number of confirmed cases
on May 3, 2020: 1,151,933 cases in 2,831 counties

Figure 13. The spread of the pandemic across counties

2.3. Selection of observation time interval

The key focus of the experiment was to target the spread of the pandemic since its outset. Thus, the first confirmed and recorded case of COVID-19 in each county was taken as the starting point of the observation interval. Many counties recorded no new cases for several days after the first case was reported, so the day the second confirmed case was recorded was taken as the start day to reduce the gap in the number of days for which no new data were collected.

In order to implement ML algorithms for time series, it was necessary to select a time interval of the same fixed length for every county, as this would allow the construction of a graph using TDA based on an n-dimensional vector representing each county over a specific time interval. Given the speed with which the pandemic has advanced, it was critical not only to select an appropriate starting point but also to limit the analysis by carefully selecting the endpoint of the time interval to make it relevant without overloading the model.

Governmental and public health authorities across the United States implemented a stay-at-home policy to prevent the spread of COVID-19. During the stay-at-home period, typical measures included shuttering non-essential business operations and requiring people to stay home unless performing essential activities or attending medical facilities. The stay-at-home restriction was adopted in most states. The experiment aimed to analyze how COVID-19 spread in every county at the beginning of the outbreak and during the stay-at-home period when contacts among people were limited in most counties and conditions for spreading the disease in different counties were as similar as possible.

For our model, the first day after the stay-at-home restrictions were lifted for each county was set as a maximum limit of our observation time interval. In other words, for each county, the observation time interval was determined to lie within **Start_day** – the second confirmed case in the county, and **End_day** – the day of release from the stay-at-home order based on statewide orders [14]. The jump in the number of cases was significant, going, for example, from 451 to 330,384 between March 8 and April 5, with an additional 821,549 cases between April 5 and May 3 (see Figure 13); therefore, our time interval was set at 39 days so as not to overload the model. Figure 14 below illustrates the selection of the start day and end day of the observation interval, whereby the start day was selected as the day of the second confirmed case and the end day was selected as the release day of the stay-at-home period.

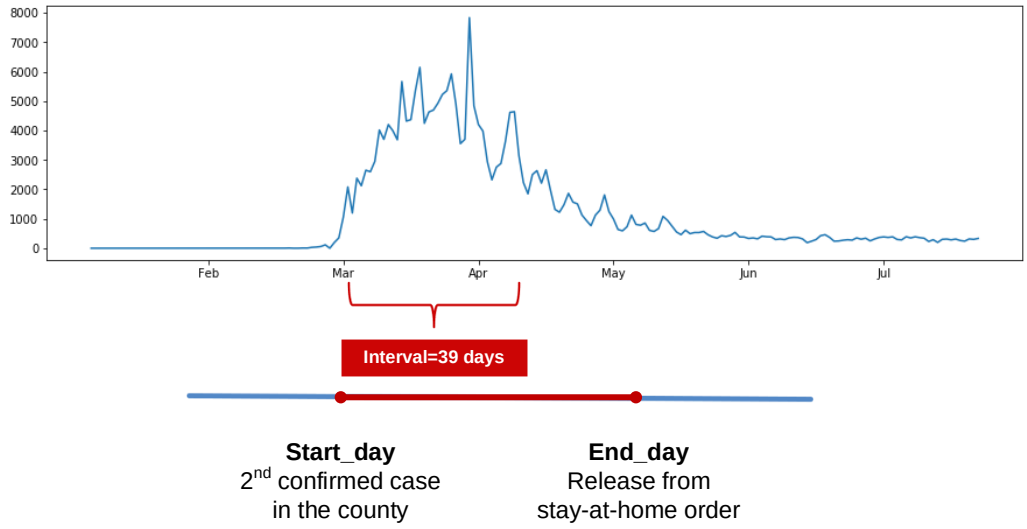


Figure 14. Selection of time series observation interval

In the scenario when more than 39 days elapse between start_day and end_day in a specific county, the observation interval was limited to the 39th day after the second reported case. In the opposite scenario when there were not enough data between start_day and end_day to cover a 39-day interval, the empty days were filled with zeros while the observation interval was expanded by an additional 5 days. It was assumed that social activities in the first five days after the release day from stay-at-home order did not affect the number of confirmed cases because of the disease incubation period and the delay in the availability of COVID-19 testing results. Figure 15 illustrates the selection of the start_day and end_day of the observation interval. The first diagram shows the observation interval within the 39-day limit, while the second demonstrates the scenario in which not enough data points are available to cover the 39-day time series.

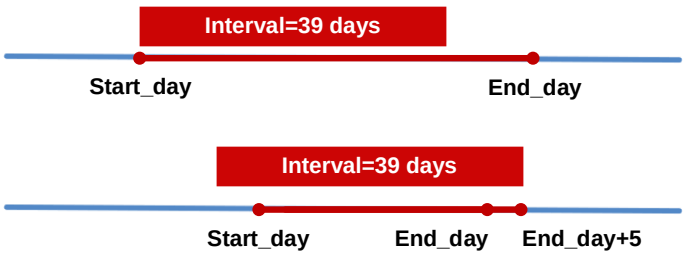


Figure 15. Selection of the start_day and end_day of the observation interval

This 39-day observation interval was used to structure a table of outcomes for the model. The n-dimensional vector built for every county combined 13 data points of confirmed cases and 13 data points of deaths, with data aggregated every three days. This vector, with the additional transformation described in Section 2.5, was further processed by the computational platform to build a graph using TDA.

2.4. Analysis of epidemic curves

Epidemic curves are an important component of public health, especially during a pandemic outbreak, allowing epidemiologists to represent the onset of disease cases over time visually. An epidemic curve is a histogram or plot illustrating the onset and progression of an outbreak of infectious disease in a particular population over a specific time [15]. The time interval is displayed on the X-axis and case numbers are shown on the Y-axis. This visual representation provides useful information on the size, pattern of spread, time trend, and exposure period of the outbreak.

For the purpose of the current experiment, epidemic curves were built for every county based on the number of confirmed cases. The resulting epidemic curves are different in terms of a shape (e.g., the peak of confirmed cases differs across counties, falling at the beginning, middle, or end of the curve) and of total number of confirmed cases of the disease reported during the observation period. To incorporate all counties into the model during the pre-processing stage, the data were first normalized using the standard score (also known as the z-score). The normalized epidemic curves were used to compare different counties in terms of the varying number of confirmed cases within the observation interval.

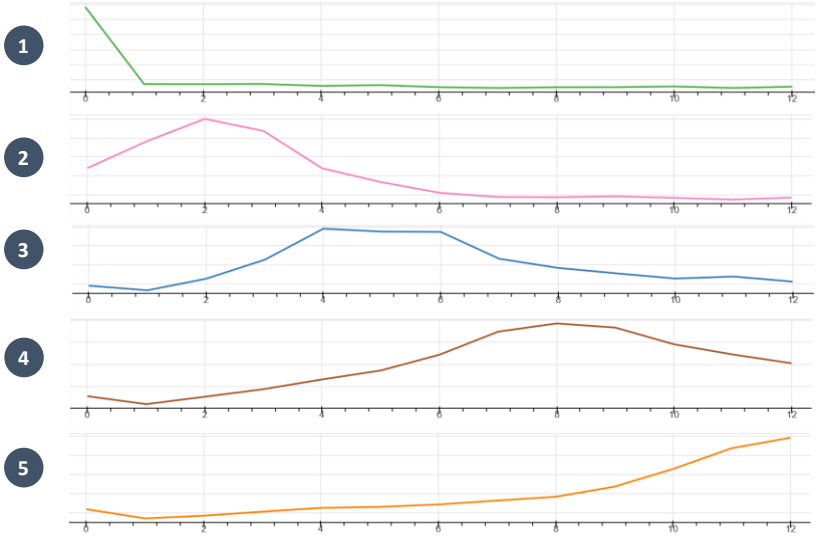


Figure 16. Five groups of epidemic curves built according to the number of confirmed cases

After normalization, the epidemic curves were grouped according to the similarity of shape of the epidemic curves. The k-means clustering method was selected as a tool for grouping the epidemic curves into clusters. The k-means method starts with the selection of the number of groups (i.e., clusters) into which the epidemic curves should be grouped. In the current experiment, the number of groups was set to five types of epidemic curves. When performing the clustering, the k-means method enables calculation of the center of a cluster, which is then used to determine the type of shape of the epidemic curves within the cluster. Next, all counties were grouped according to similarity of epidemic curves and assigned to one of five types to illustrate similarity in terms of the pattern of disease spread (see Figure 16).

For illustrative purposes, we assigned a specific color to the epidemic curve of each county, using five determined types. By combining administrative and geographical [16] maps of the United States (see Figure 17), it was possible to visually identify the similarity of pandemic spread based on geographical features. For example, counties located in mountainous areas, which probably indicates a lower population density, have a similar pattern: a peak at the beginning and a flattening of the curve thereafter (type 2 epidemic curve; colored pink). This visual approach, combined with the additional analysis of a variety of predictors (social, demographic, etc.), provides a much broader picture of the reasons behind the spread of the pandemic. The same approach was used to analyze the number of deaths integrated into the model for each county.

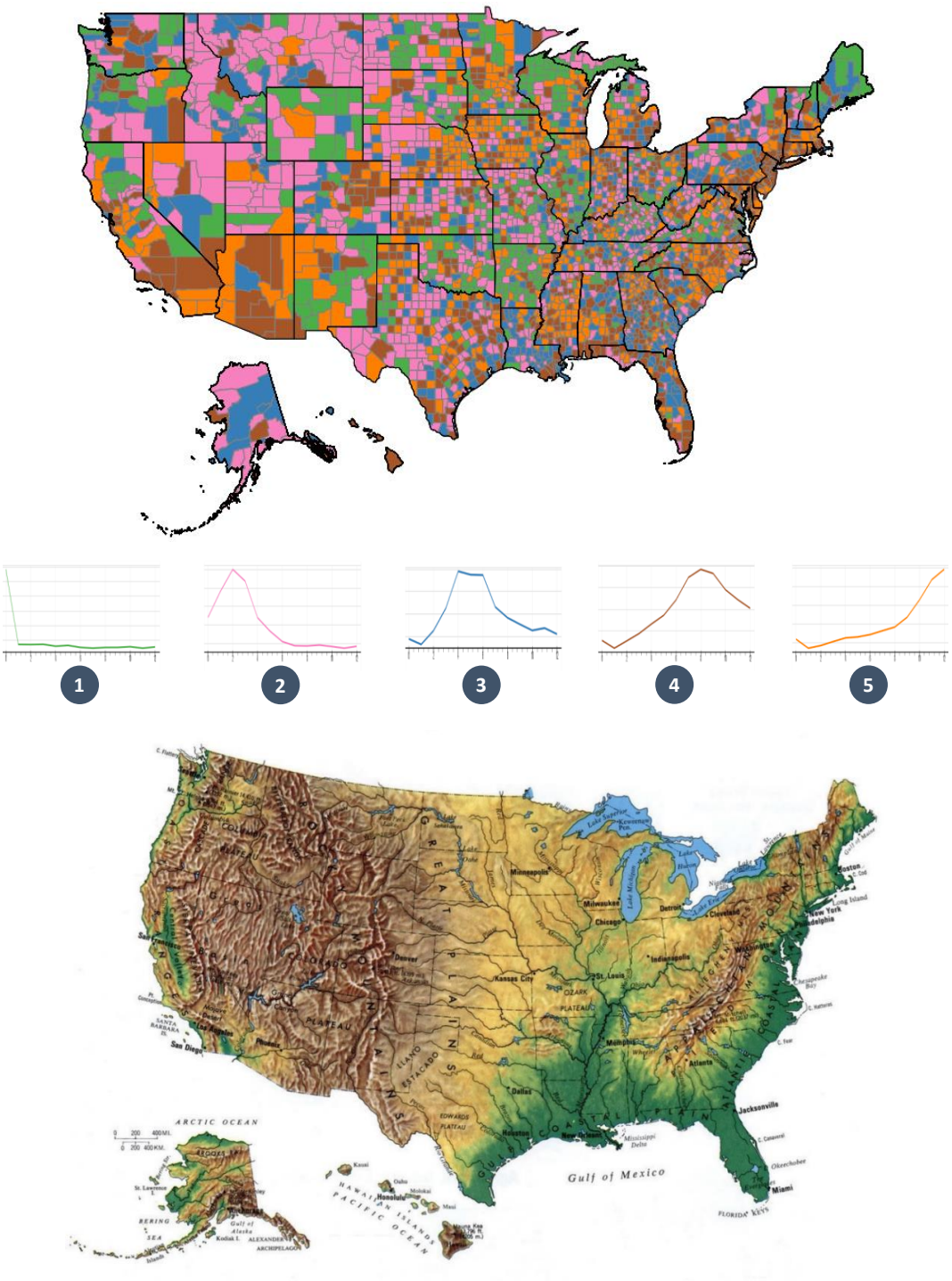


Figure 17. Five types of epidemic curves on administrative and geographical maps

2.5. Outcomes selected for the model

In the present experiment, TDA was used to build a graph that visually represents the geometric properties of the dataset. In other words, TDA allowed a comprehensive graph to be extracted from the original dataset to provide a compressed graphical representation of a multidimensional set of interrelated outcomes. The graph consists of 3,142 nodes corresponding to each county and edges connecting the counties that are similar in terms of outcomes.

As described above, the analysis was performed with respect to a 39-day observation interval. Thus, the outcomes selected in this study include 13 (39 / 3) normalized numbers of confirmed cases, the logarithm of the total number of confirmed cases, 13 (39 / 3) normalized numbers of cases of death, and the logarithms of the total number of deaths calculated for each county. Therefore, the total number of outcomes for each county is 13 + 1 + 13 + 1 and equals 28 (see Figure 18).

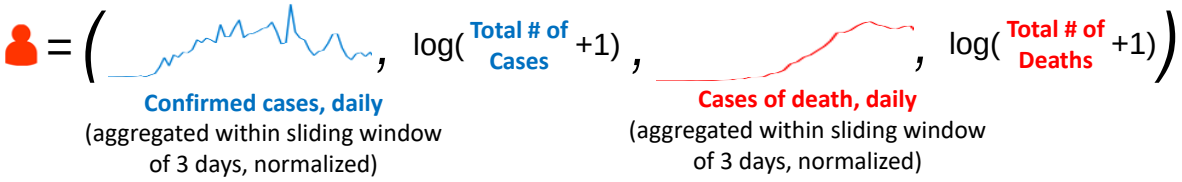


Figure 18. Outcomes selected for each county

Therefore, the relevant information about the spread of COVID-19 in each county is encoded by a 28-dimensional vector as illustrated in Figure 19:

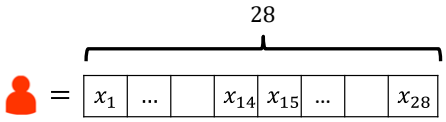
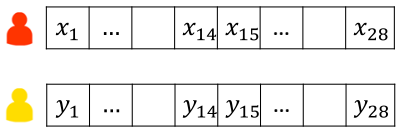


Figure 19. 28-dimensional vector constructed for each county

Encoding this information into this vector allowed investigation of the dataset based on four characteristics of the spread of COVID-19: the shape of epidemic curves built on confirmed cases, the total number of confirmed cases, the shape of the epidemic curve built on the number of deaths, and the total number of deaths.

To determine similarities between two counties while building a graph, it is necessary to calculate the distance between two 28-dimensional vectors. For that purpose, a distance function based on modified Euclidean distance was used to measure the similarity in the spread of the pandemic between the counties (see Figure 20).



$$d(\,,\,) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_{14} - y_{14})^2} + \sqrt{(x_{15} - y_{15})^2 + \ldots + (x_{28} - y_{28})^2}$$

Figure 20. The distance function measuring similarities between two counties

Thereafter, using the two-dimensional projection based on density and centrality of nodes with respect to selected outcomes, the resulting TDA graph was built.

2.6. Defining the most representative graph

By changing parameters (the number of bins and percentage of their overlap) of the topological model, we built a family of more than 200 graphs. The most representative graph (see Figure 21) was selected from this family using the approaches described in Section 1.4. This graph has received high scores by methods such as random walk-based modularity, Kolmogorov complexity, edge-aggregated graph selection.

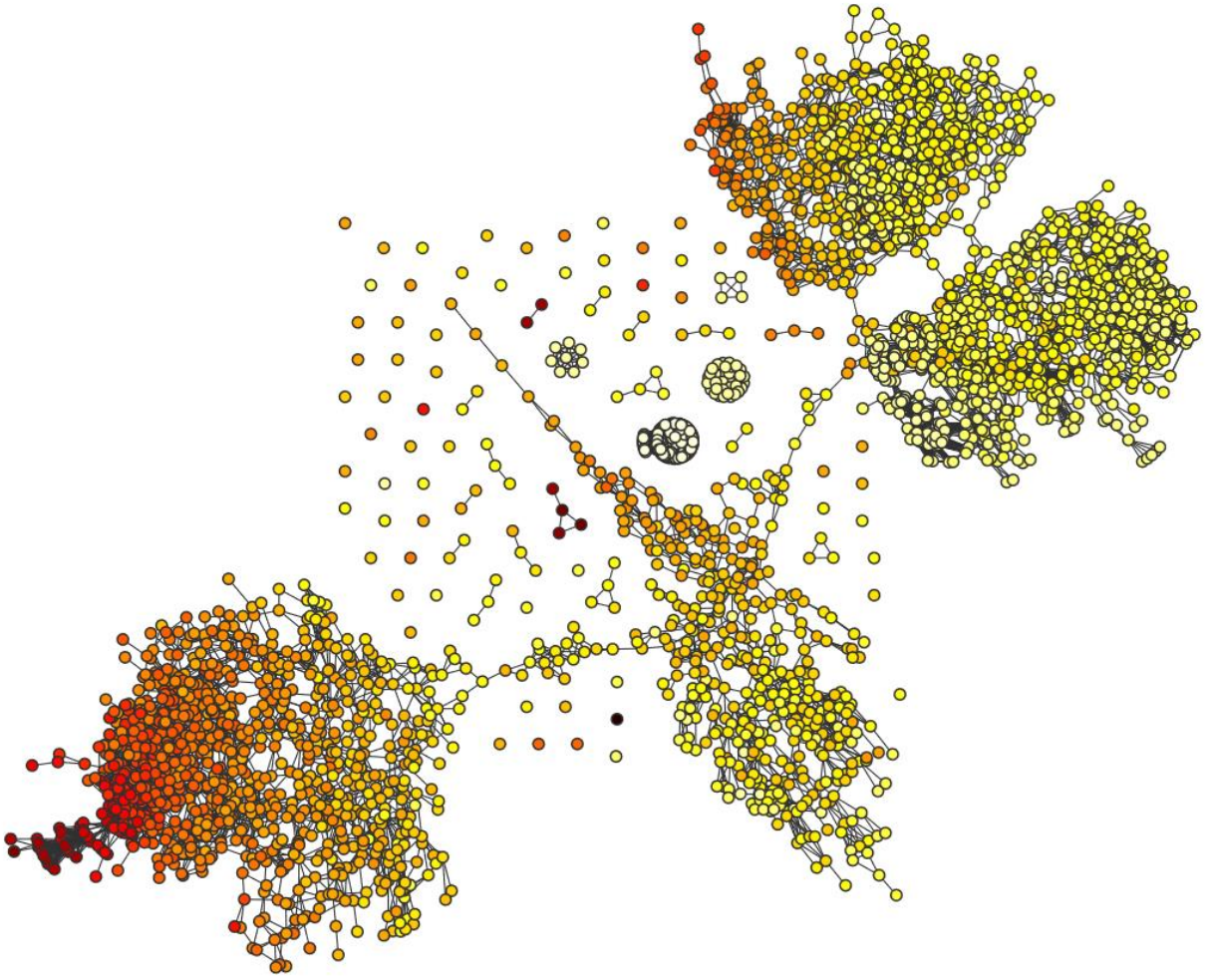


Figure 21. The topological model of COVID-19 spread over the United States

The most representative graph capturing key features of our dataset with selected outcomes will be further analyzed using ML algorithms, visual exploration, and statistical analysis of predictors. To summarize, here are a few essential points that describe this graph:

- Each node on the graph corresponds to one US county (total 3,142 nodes).
- Two nodes are connected with an edge if they are similar in terms of predefined outcomes.
- Outcomes integrate the number of confirmed cases and deaths within a 39-day observation interval from the start of the pandemic.
- The similarity in pandemic spread is measured by a modified Euclidean distance function.

2.7. Predictors used for statistical analysis

After a topological model is extracted, the researcher visually explores the graph to discover interesting subgroups within the data. These subgroups can be further studied by utilizing standard statistical methods to determine the predictors that may be responsible for the similarity of the pandemic spread observed within the identified subgroup of counties.

A very common situation in statistics occurs when the distribution of an outcome (or response variable) is related to one or several predictors (or explanatory variables). A standard approach used by researchers to study the relationship between a predictor and an outcome is the application of a suitable statistical model. The model selection depends on the data types of the predictor and outcome (quantitative, binary, categorical, etc.) and often involves additional assumptions concerning the distribution of the outcome. To describe similarities in pandemic spread, we integrated over 250 predictors into our model, available from a variety of public sources (see Figure 22). At any time, additional predictors can be easily integrated into the analysis without any changes in the topological model.

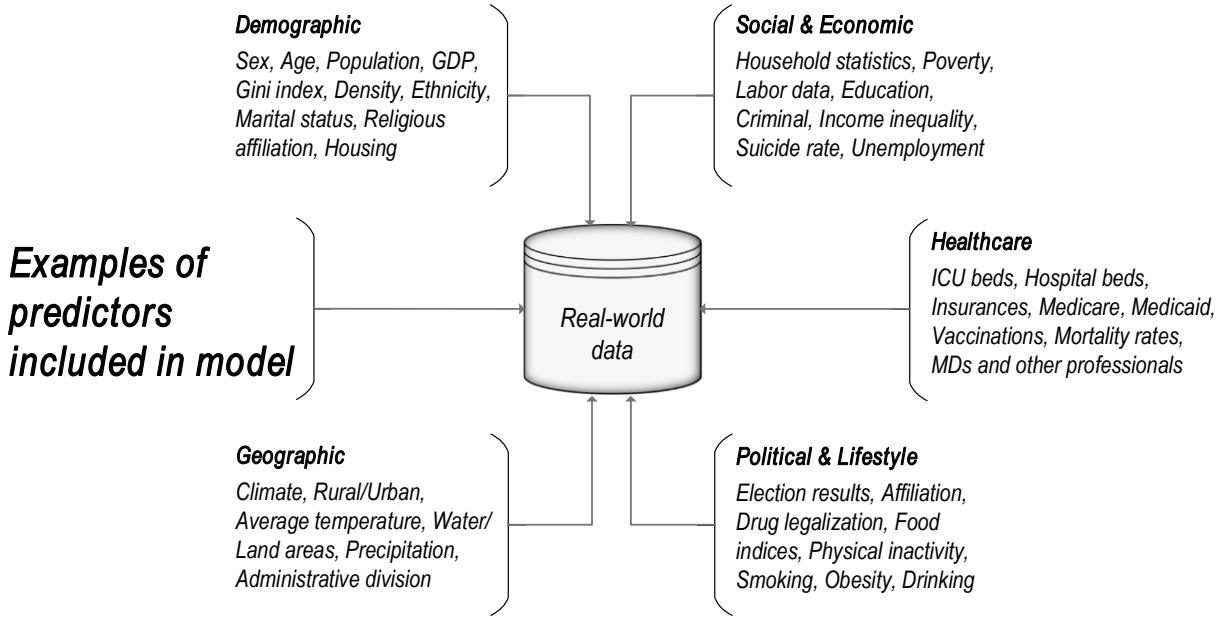


Figure 22. Group of predictors incorporated into the model.

Interactive visualization provides researchers with an opportunity to manually perform a visual inspection of a graph to identify regions of interest. For example, the nodes that form fork-like structures or loops might be of interest for further research. In addition, isolated components or highly concentrated groups of nodes that form communities may indicate meaningful relationships in the outcomes dataset. While performing a visual inspection, the researcher can also recolor the graph in accordance with the value of an outcome or predictor selected from the corresponding datasets. The use of color codes may highlight how a subgroup of nodes represented by a given region of the graph might be different from the rest of the nodes.

The researcher can select any region of the data map that exhibits interesting geometric properties to perform further statistical analysis. After running statistical tests, a table of predictors with the corresponding p-values can be calculated to determine whether the distribution of the predictors for the selected subgroup of nodes is different from that of the rest of the nodes. If the desired significance level of any predictor is found to be statistically significant, the researcher can construct a histogram that represents normalized frequency distributions of the predictor for both the nodes in the selected region of the graph and the rest of the nodes. The same can be done when comparing two different selected regions of nodes with each other.

For the purpose of the statistical analyses undertaken for the present experiment, continuous, mixed, binary, and categorical (non-binary) univariate predictors were differentiated according to a variable type. Continuous predictors were examined using the standard two-sample Kolmogorov-Smirnov test. This method verifies whether two data samples are obtained from the same distribution. The test assumes that the underlying distributions are continuous (no ties in the variables' values are allowed). To examine the statistical association between two samples within the categorical data, Fisher's exact test and the χ^2 test were used for the binary and non-binary categorical variables, respectively.

Figure 23 describes the variable type, data source, and other features of the predictors that were integrated into the model.

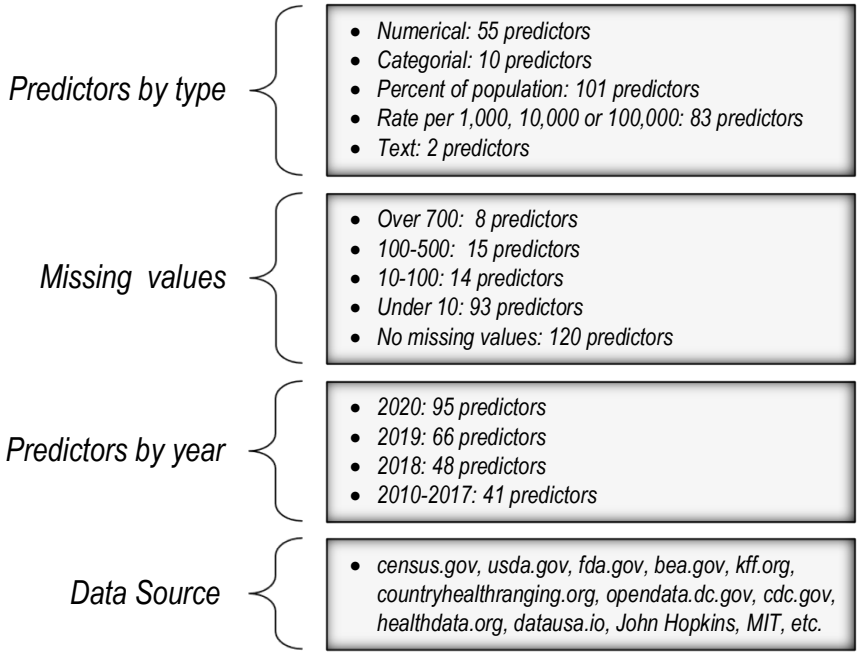


Figure 23. The characteristics of predictors integrated into the model

Subsequently, a number of communities were identified in the model using the methods of automatic community detection described in Section 1.4. Comparing these communities with each other, as well as comparing each community with the rest of the graph nodes, more than **5300** statistical tests were performed on predictors, of which about **3700** were statistically significant with p-value ≤ 0.05.

In the following section, we will describe several such comparisons that we found particularly interesting.

3. FINDINGS EVALUATION AND INTERPRETATION

3.1. Girvan-Newman communities

By constructing a graph representing the original dataset with the number of confirmed cases and deaths, we were able to undertake a visual exploration to discover sub-populations within the data. For example, isolated components of the graph or highly interlinked groups of nodes may indicate meaningful relationships in the dataset. As datasets can span a large number of nodes and edges, visual inspection, and further discovery of sub-populations within the graph can be challenging or even misleading. Thus, we rely on some known machine-learning algorithms applied to the automatic detection of sub-populations using community search on the graph (see Section 1.5).

Looking at the geometric properties of the graph extracted from the dataset that embeds the number of confirmed cases and deaths within a 39-day interval (see Figure 14), we can clearly distinguish four main regions. Using the Girvan-Newman method, four communities were found on the graph that have more similarities (number of edges) with each other than with the rest of the counties. Figure 24 illustrates the communities to which we refer.

The total number of counties within these four communities is 2,736 (out of 3,142), with the largest, the purple community located on the lower left, corresponding to 1,029 counties. Let us first consider the map of the USA and the distribution of these four communities. We colored the counties on the US administrative map according to the colors of the nodes corresponding to the counties in the detected communities (see Figure 25). An analysis of the map clearly indicates that counties colored red and green lie in regions with a lower population density. In contrast, purple counties are located in densely populated areas.

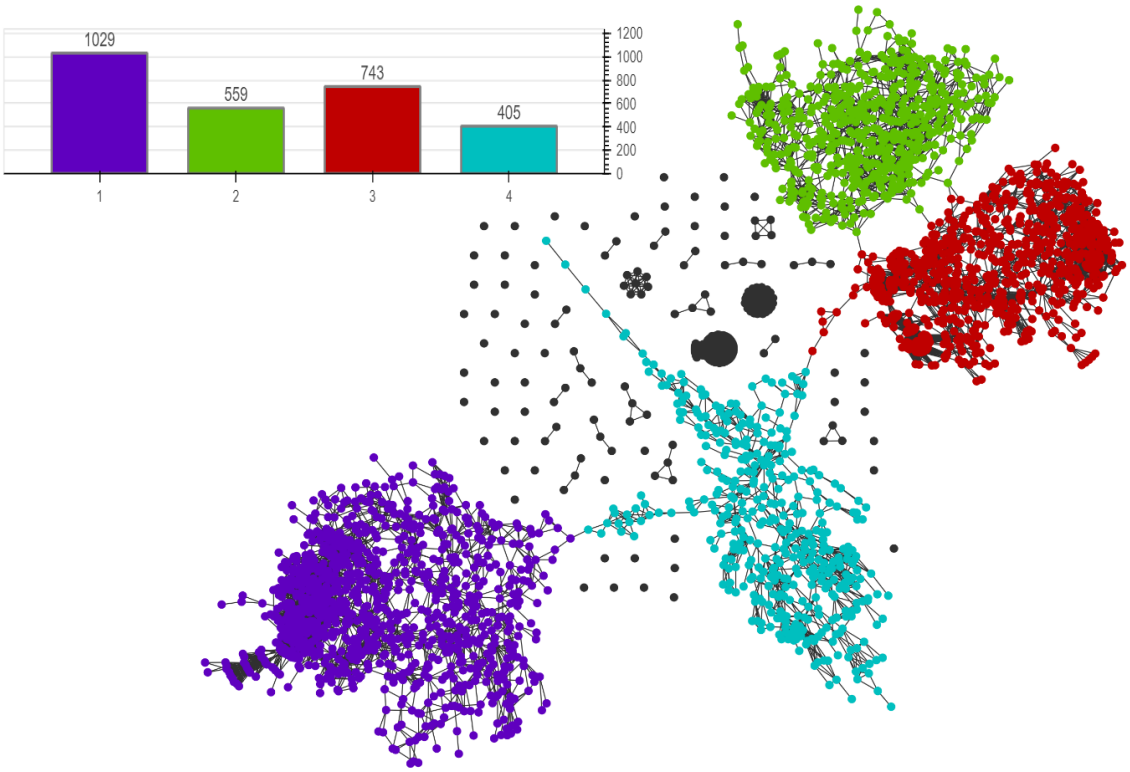


Figure 24. The four Girvan-Newman communities detected on the graph

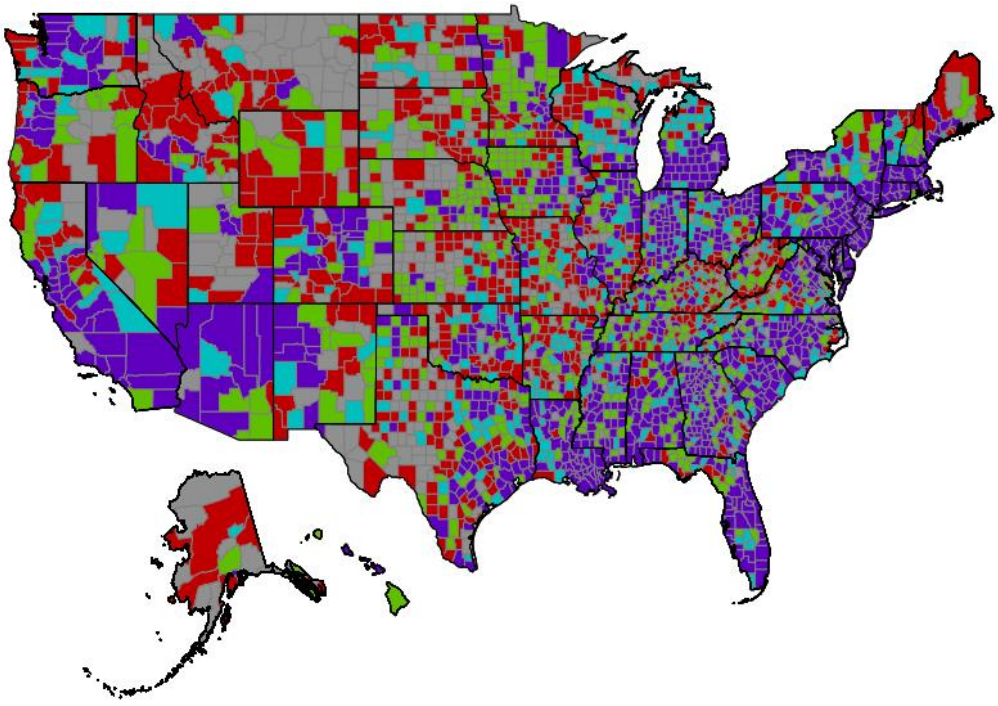


Figure 25. The US administrative map with the four Girvan-Newman communities detected on the graph

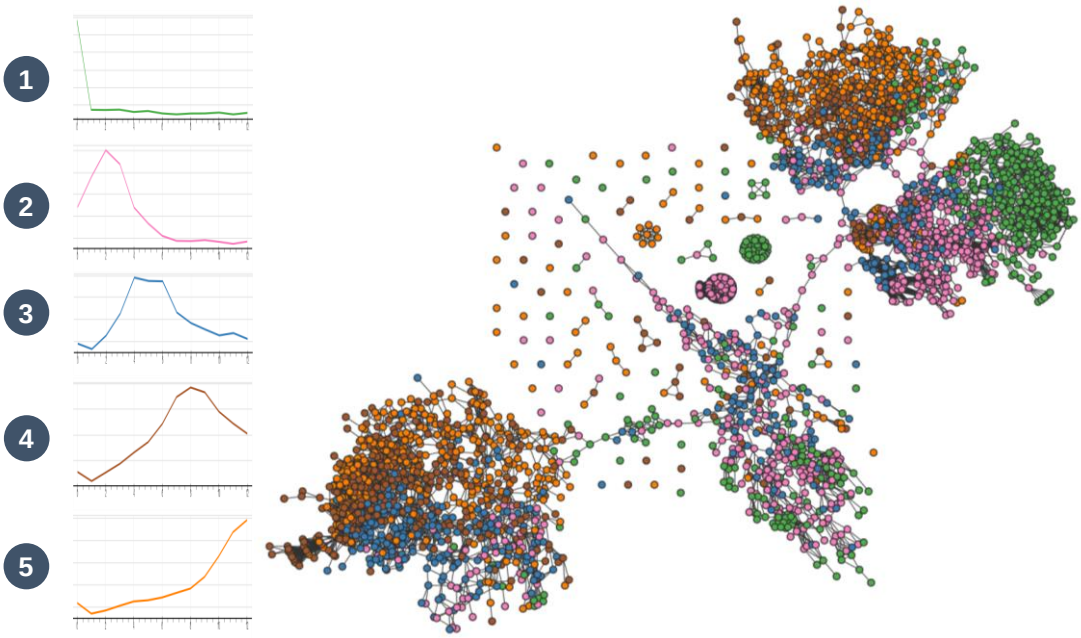


Figure 26. The graph colored according to the **epidemic curves** within the observation interval

Let us next analyze how the four major communities differ in terms of pandemic spread patterns. To determine the significance of the type of onset and progression of an outbreak of the disease, the graph was re-colored according to the epidemic curves described earlier (see Figure 16). As we can see in Figure 26, the lower-left and upper-right communities are similar in terms of epi-curves. In these two communities, the nodes are mostly colored blue (3), brown (4), and orange (5), which corresponds to the delayed outbreak of the disease. The lower-right and central communities are mostly colored green (1) and pink (2), which corresponds to the disease outbreak at the beginning of the observation interval, with a further significant reduction of confirmed disease cases.

To further determine specific features of these four communities in terms of the predefined outcomes, the graph was colored according to the number of confirmed cases within the 39-day observation interval (see Figure 27) and number of deaths (see Figure 28). The dark color indicates a higher number of cases, while the yellowish color indicates the opposite.

Looking at these two colorings, we can clearly see that a lower-left (circled in purple) and upper-right community (circled in green) have similar patterns in regard to the number of confirmed cases as compared with the other two communities (cyan in the center and red on the lower right). On the other hand, referring to Figure 28, these two communities, purple and green, have different patterns in the number of deaths: the green community, on the upper right, has far fewer deaths. In other words, when the pandemic started, a similar pattern of spread was seen in both purple and green communities. At the same time, the number of deaths indicates that counties located in the green community handled the virus more effectively, which substantially reduced the number of fatalities. Even though there are almost twice as many counties in the purple than in the green community (1,029 versus 559), this is still important for further analysis.

These two communities, purple and green, will become of interest for further analysis to determine **why the number of deaths was significantly different** while at the same time the spread of the disease showed a clear similarity, based on the shapes of epi-curves and the number of confirmed cases. For this purpose, we perform a statistical analysis of discovered patterns to explain similarities in the spread of the pandemic based on the predictors integrated into the model.

After running statistical tests in these two communities, a table of predictors consisting of 267 rows with corresponding p-values, of which 211 were statistically significant with p-values ≤ 0.05, was calculated to determine whether the distribution of the predictors for the selected purple subgroup of nodes was different from that of the green subgroup. If the desired significance level of any predictor was found to be statistically significant, we were able to construct a histogram representing normalized frequency distributions of the predictor for both the nodes in the selected purple community and those from the green community.

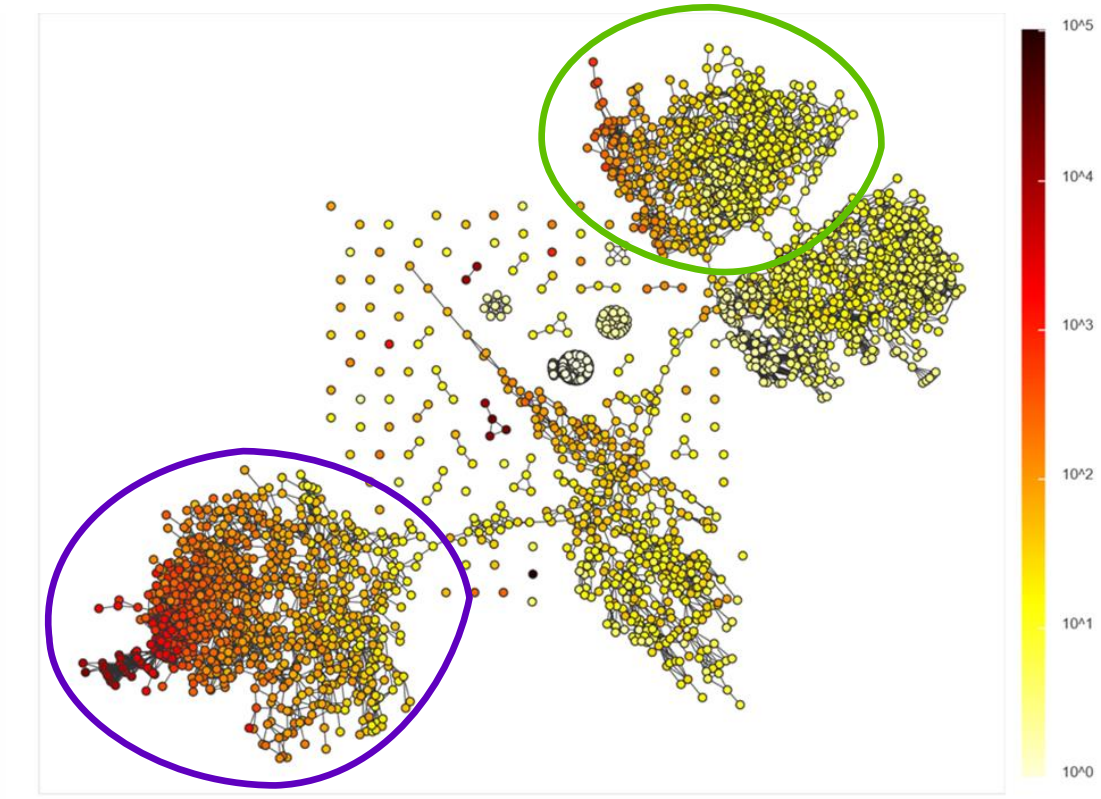


Figure 27. The graph colored according to the **number of confirmed cases** within the observation interval

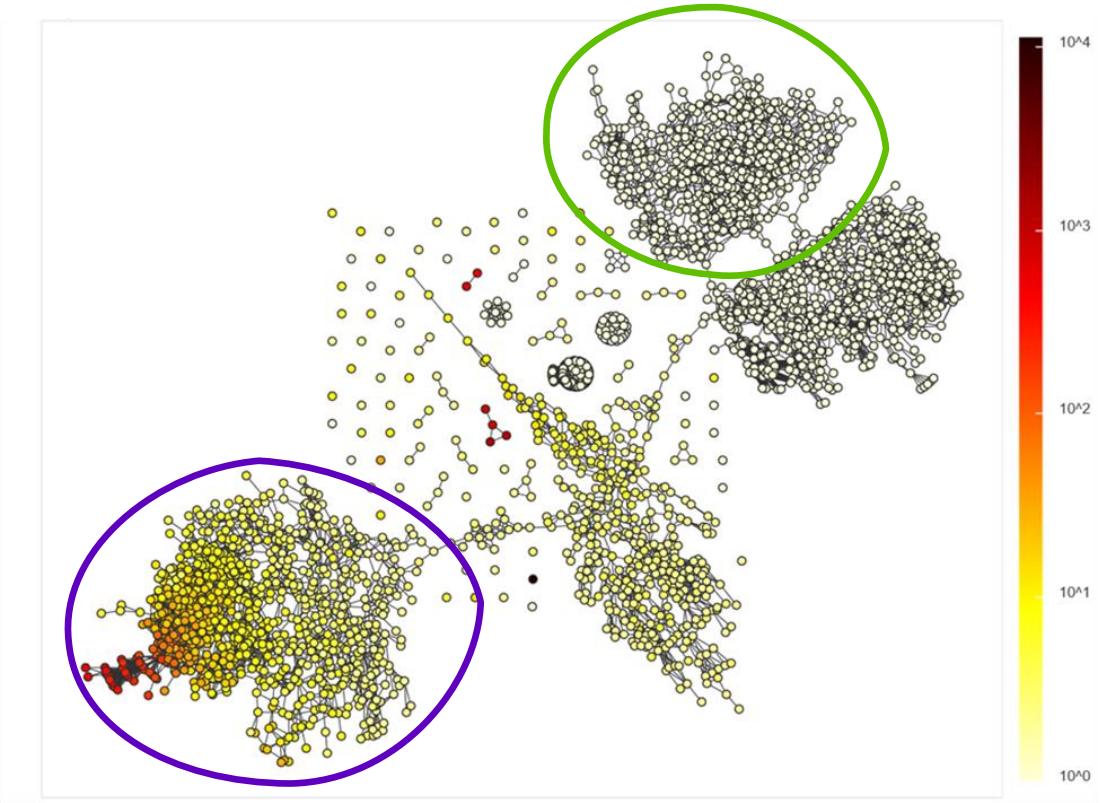
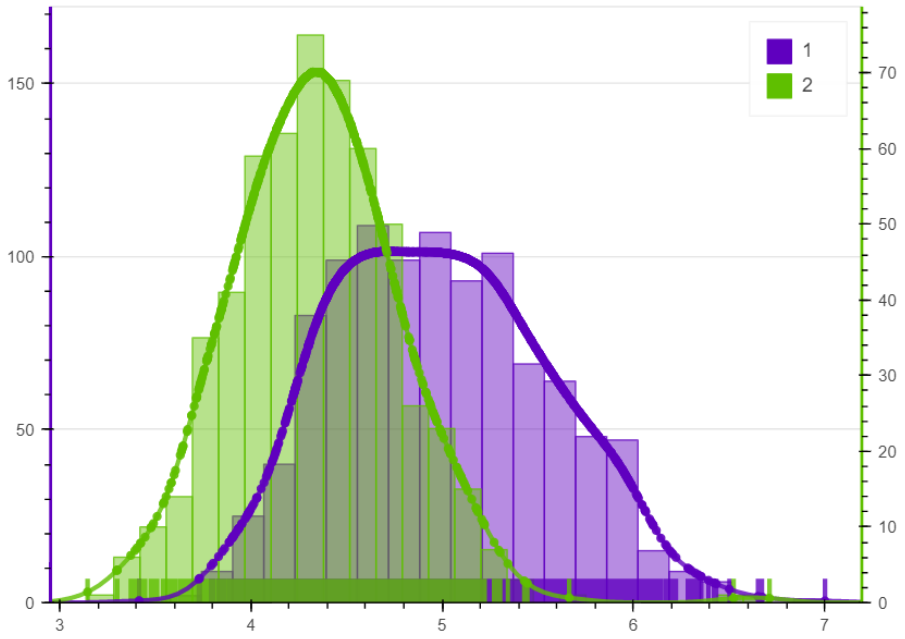


Figure 28. The graph colored according to the **number of deaths** within the observation interval

For the purpose of the statistical analyses undertaken for the present experiment, continuous, mixed, binary, and categorical (non-binary) univariate predictors were differentiated according to variable type. Continuous predictors were examined using the standard two-sample Mann–Whitney–Wilcoxon test. To examine the statistical association between two samples within the categorical data, Fisher's exact test and the χ^2 test were used for the binary and non-binary categorical variables, respectively.

In the next step, we analyzed various predictors with statistically significant p-values ≤ 0.05 to describe the differences between the purple and green communities. Thus, Figure 29 shows a histogram that compares the population in both. The purple community appeared to have more densely populated counties, including cities and big urban centers. In contrast, the green community turned out to have moderately populated counties, mostly located in rural areas (see Figure 25 for more detail).



Count	Mean	Median	Min	Max	Std
1029	4.992	4.969	3.416	7.002	0.567
559	4.343	4.323	3.146	6.712	0.439

Figure 29. The histogram showing the **logarithm of population** in the green and purple communities

The next statistically significant predictor is the urban influence code, also called an "urban/rural score." In the histogram illustrated in Figure 30, the value "1" on the scale corresponds to big urban centers while "12" corresponds to villages. The purple community mostly encompassed big urban centers, while the green community turned out to include counties in non-urban areas.

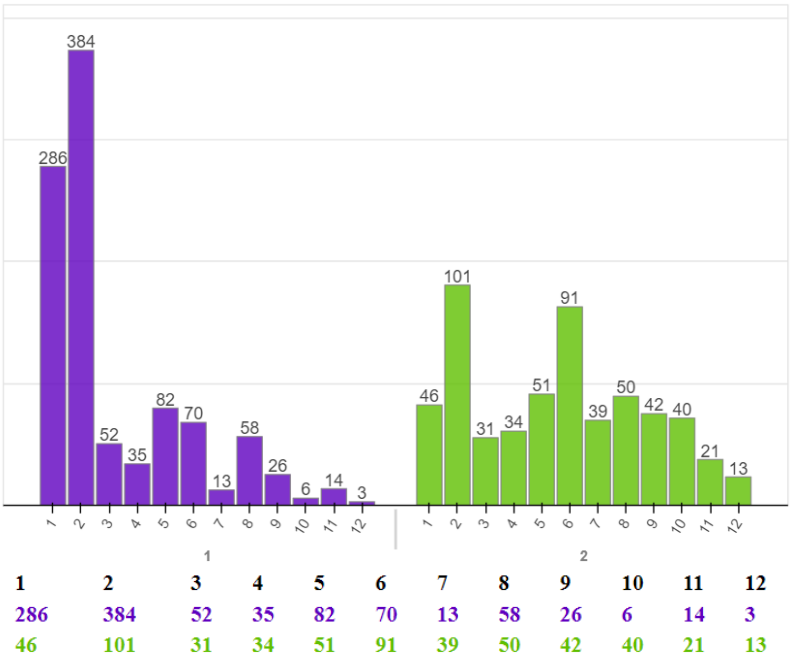


Figure 30. The histogram showing the **urban influence code** in the green and purple communities

Another predictor found to be significant is the public transportation system, which might be responsible for the difference in the number of deaths between the green and purple communities. We compared these communities based on the public transportation score (see Figure 31). The public transportation score indicates how well a location is served by public transit. The purple community appeared to have a much higher public transportation score than the green one.

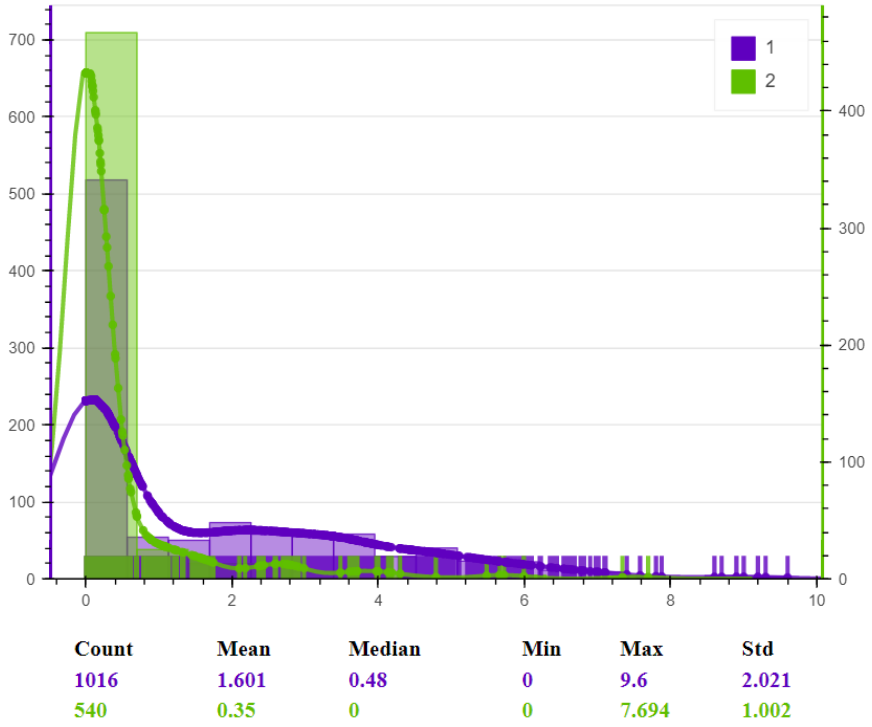
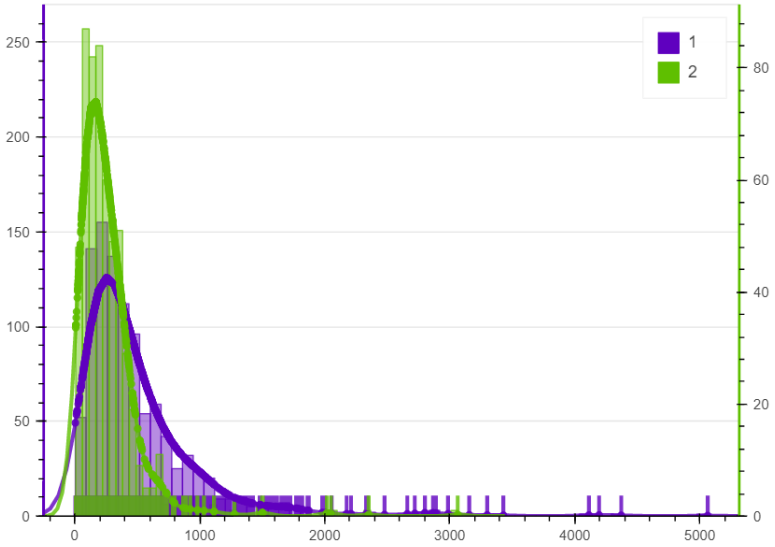


Figure 31. The histogram showing the **public transportation score** in the green and purple communities

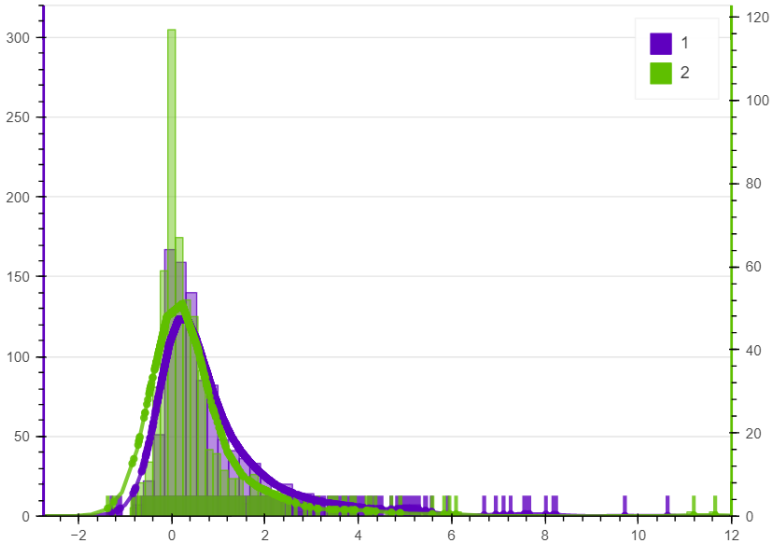
Analyzing more predictors in the transportation segment, Figure 32 shows a histogram comparing the purple and green communities based on highway length per 1,000 square kilometers. This predictor was found to be statistically significant with the p-value ≤ 0.05. The purple community appeared to have a much greater highway length than the green one.



Count	Mean	Median	Min	Max	Std
1029	516.08	370.132	5.497	5063.183	520.057
557	265.703	207.978	8.802	3064.092	267.615

Figure 32. The histogram showing the **highway density** in the green and purple communities

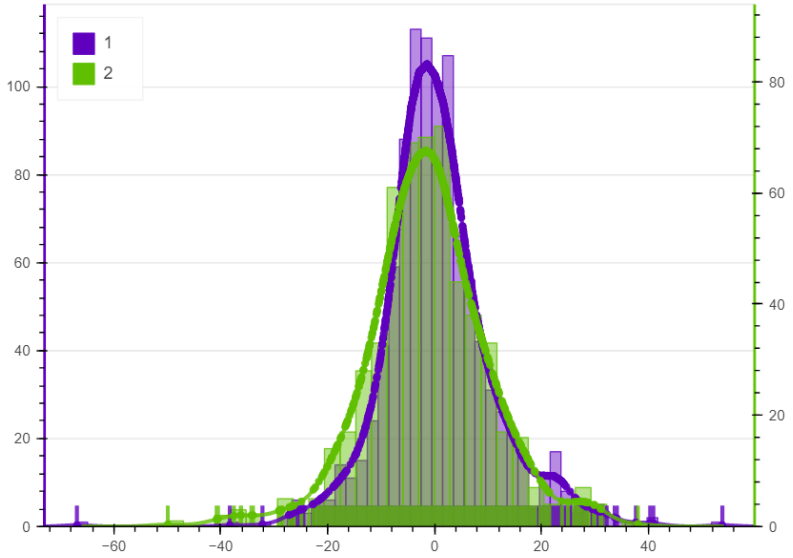
The international migration rate within the purple and green communities was also compared. This rate illustrates the difference between the number of people coming to the county from outside of the USA and the number of people leaving the county for international destinations throughout the year. Figure 33 shows this histogram. The purple community appeared to have approximately twice the international migration rate of the green community.



Count	Mean	Median	Min	Max	Std
1029	0.937	0.467	-1.292	13.998	1.45
559	0.584	0.192	-1.374	26.042	1.617

Figure 33. The histogram showing the **international migration rate** in the green and purple communities

Likewise, we analyzed the domestic migration rate, which, in contrast with the international migration rate, refers to the difference in the number of people coming into and leaving the county from within the USA (not internationally). Figure 34 shows a histogram comparing the purple and green communities according to domestic migration rate. The purple community had a positive domestic migration rate, which means a surplus in arrivals into the counties of this community. The green community had a negative domestic migration rate, which means more people depart this community's counties than arrive.



Count	Mean	Median	Min	Max	Std
1029	0.669	-0.35	-66.935	53.881	10.271
559	-1.287	-1.373	-49.92	38.099	11.413

Figure 34. The histogram showing the **domestic migration rate** in the green and purple communities

The next statistically significant predictor for the purple and green communities was found to be a climate region. In the purple community, the number of counties located in northeast, east north central, central northwest, west, southwest, south, southeast climate regions were higher than in the green community. In contrast, in the green community, the number of counties belonging to the west north central climate regions and Alaska was higher than in the purple community. Figure 35 shows the USA's climate regions, and Figure 36 shows histograms comparing the green and purple communities according to the climate region.

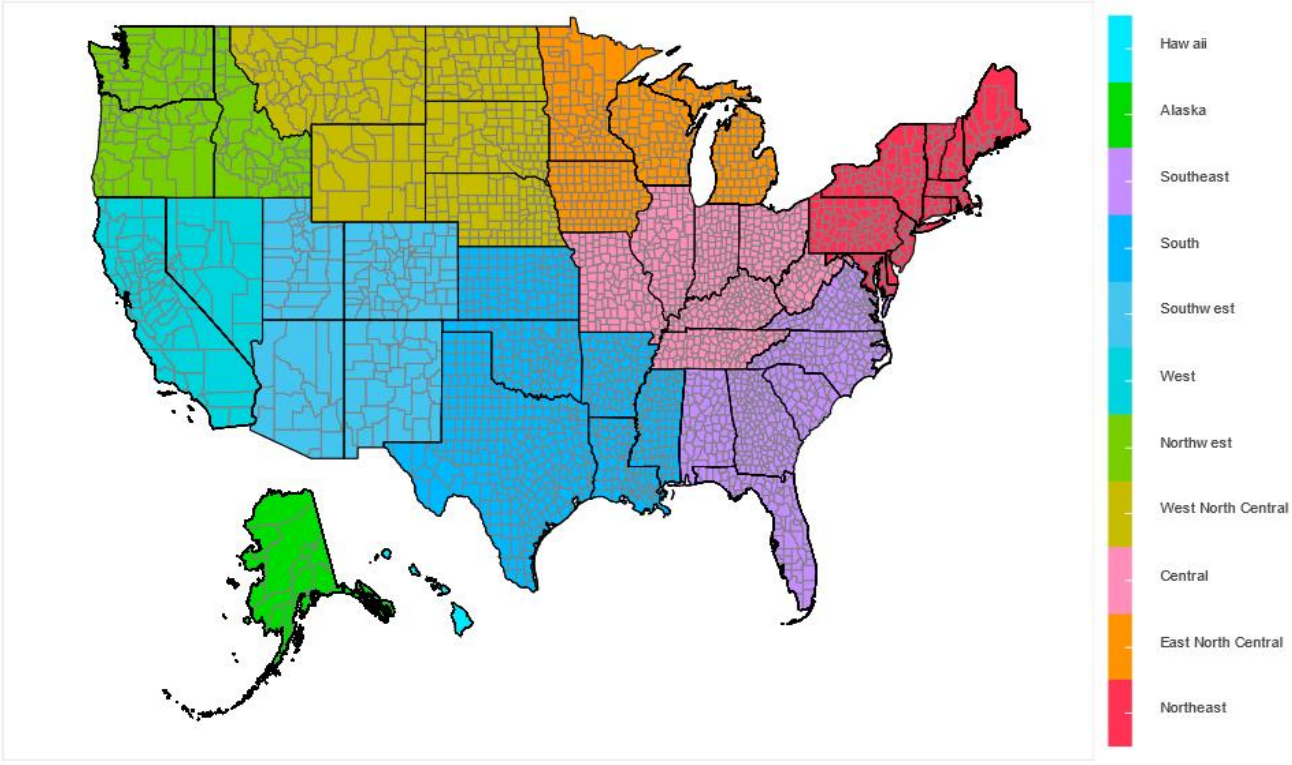
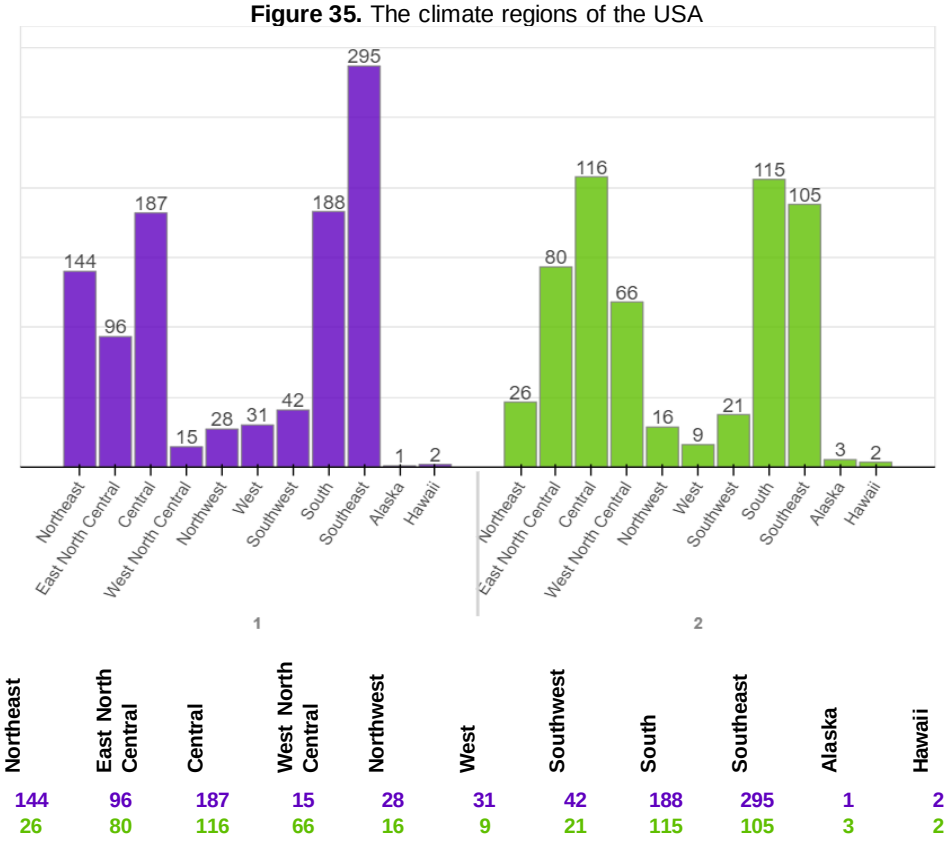


Figure 35. The climate regions of the USA



Northeast	East North Central	Central	West North Central	Northwest	West	Southwest	South	Southeast	Alaska	Hawaii
144	96	187	15	28	31	42	188	295	1	2
26	80	116	66	16	9	21	115	105	3	2

Figure 36. The histograms comparing the green and purple communities according to the climate region of the counties

Further, several predictors related to the race of the population are statistically significant. Figure 37 (a-d) shows the histograms comparing the green and the purple communities according to the percentage of different races and ethnicities. The green community appeared to have a higher percentage of White populations than the purple community did. The purple community had a higher percentage of African American, Hispanic, and Asian populations than the green community.

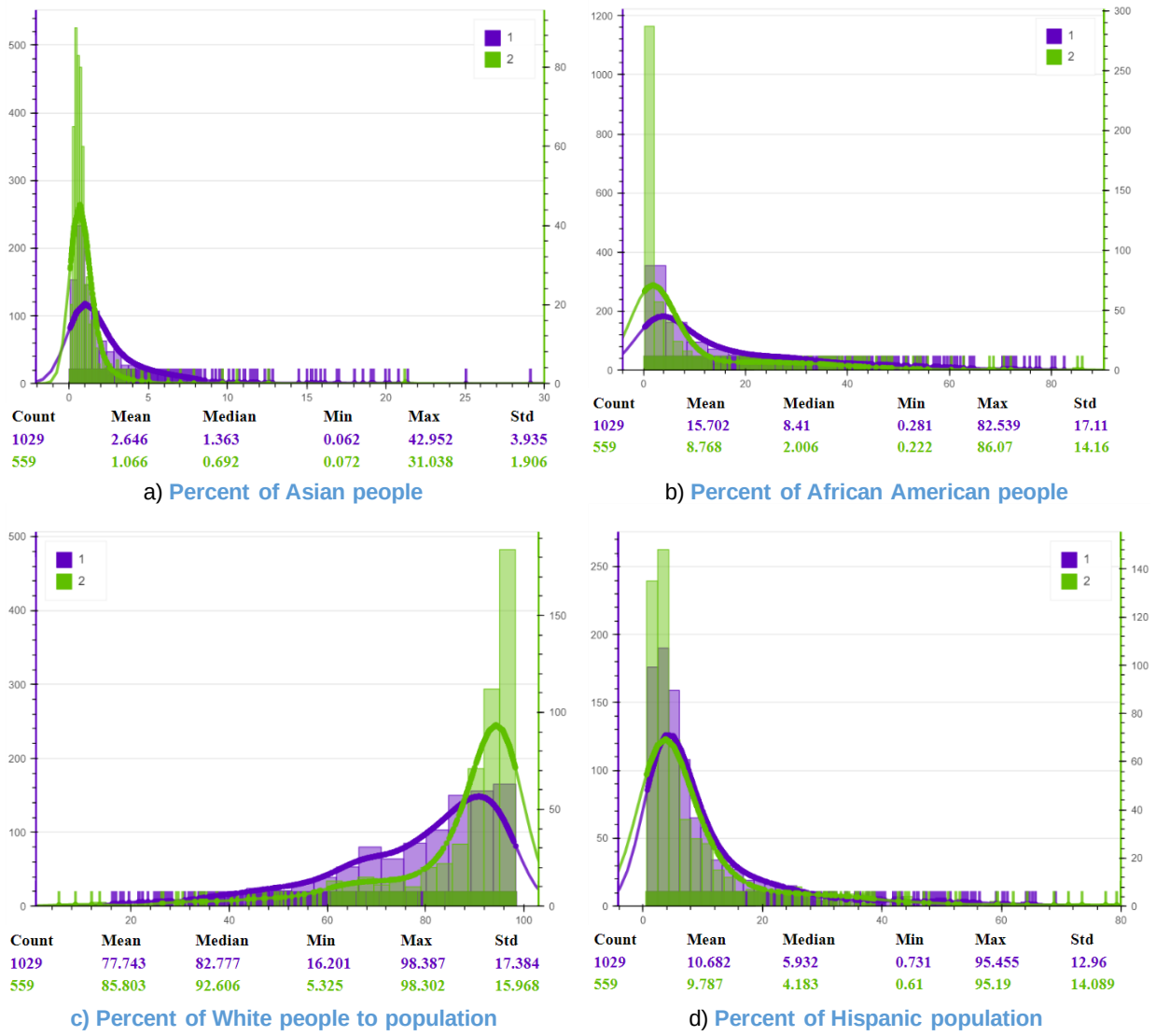
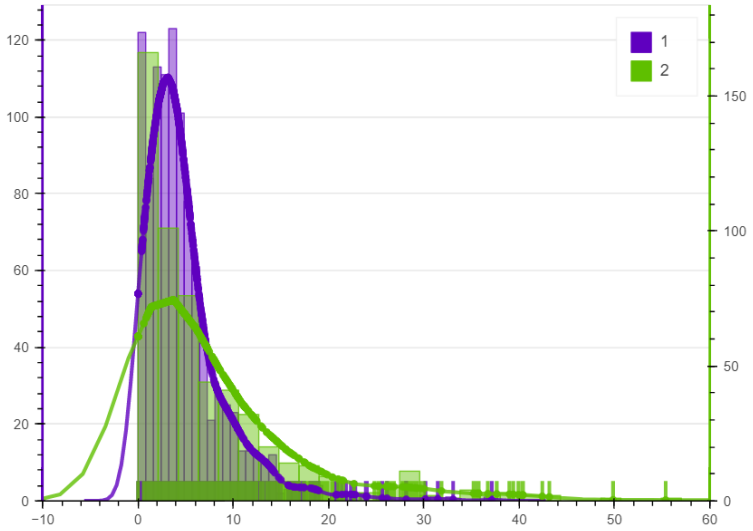


Figure 37. The histogram comparing the green and purple communities based on the percentage of different races to population within the communities

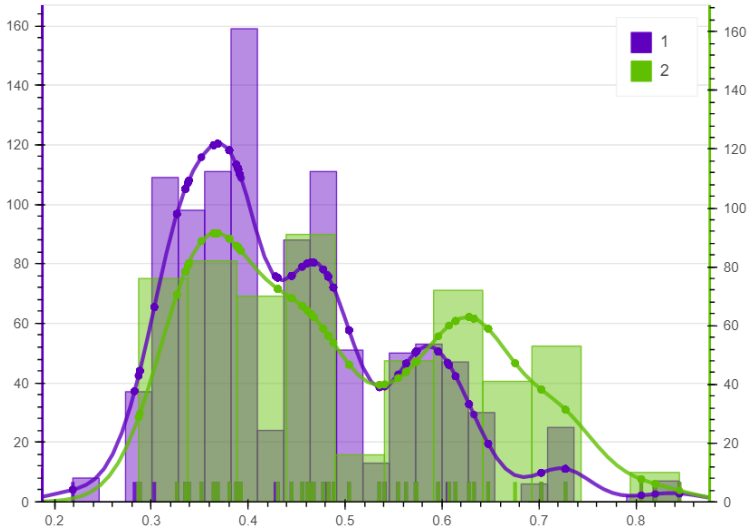
Another group of predictors that might be responsible for the difference in the number of deaths across the purple and green communities relates to the healthcare system. In the next step, the graph was recolored according to the number of hospitals per 100,000 people. Figure 38 shows a histogram comparing the purple and green communities. The green community appeared to have twice the number of hospitals per 100,000 people than the purple one. This could be one reason that counties in the green community handled the pandemic outbreak more effectively even though they experienced a similar pattern in the number of cases as counties from the purple community.

Analyzing the healthcare system further, the purple and green communities were also compared according to the number of general practice physicians per 1,000 people. Figure 39 shows the histogram comparing the purple and green communities. The green community appeared to have a slightly higher number of general practice physicians per 1,000 people than the purple one.



Count	Mean	Median	Min	Max	Std
1029	4.857	3.765	0	37.116	4.537
559	7.9	4.569	0	200.602	12.563

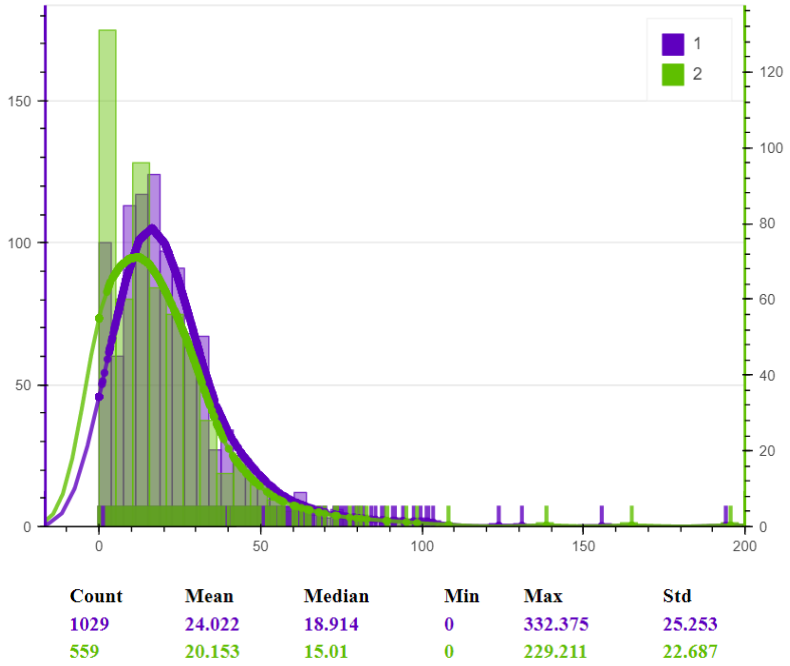
Figure 38. The histogram showing the **number of hospitals per 100,000 people** in the green and the purple communities



Count	Mean	Median	Min	Max	Std
1029	0.443	0.392	0.219	0.845	0.115
559	0.498	0.468	0.287	0.845	0.136

Figure 39. The histogram showing the **number of general practice physicians per 1,000 people** in the green and the purple communities

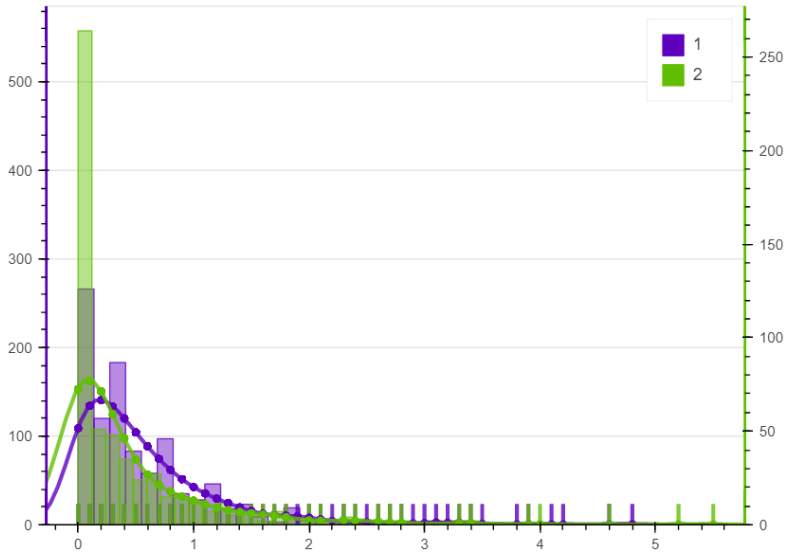
The number of staffed beds per 10,000 people was also analyzed in relation to our communities. Figure 40 shows the histogram comparing the purple and green communities. The purple community appeared to have a higher number of staffed beds per 10,000 people than the green one. Whereas there are more beds in the purple community, there are more hospitals in the green community. More small hospitals and more general practice physicians proved to be preferable for lower death rates.



Count	Mean	Median	Min	Max	Std
1029	24.022	18.914	0	332.375	25.253
559	20.153	15.01	0	229.211	22.687

Figure 40. The histogram showing the number of staffed beds per 10,000 people in the green and the purple communities

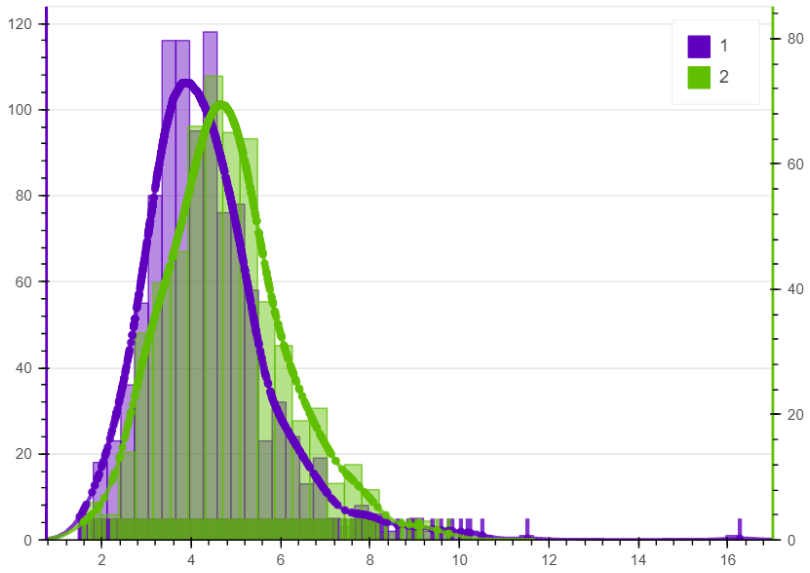
The population 65 years and over with no health insurance coverage to total population within the purple and green communities was also compared. Figure 41 shows this histogram. The percentage of the population without insurance in the purple community is greater than in the green one.



Count	Mean	Median	Min	Max	Std
1029	0.572	0.4	0	4.8	0.662
559	0.441	0.2	0	5.5	0.72

Figure 41. The histogram showing the population 65 years and over with no health insurance coverage to total population in the green and the purple communities

Likewise, we analyzed the population 65 years and over with *Medicare coverage only* to total population. Figure 42 shows the histogram comparing the purple and the green communities according to this predictor. Unlike the previous predictor, the percentage of the population with Medicare insurance is higher in the green community.



Count	Mean	Median	Min	Max	Std
1029	4.36	4.17	1.5	16.27	1.439
559	4.799	4.69	1.59	9.76	1.364

Figure 42. The histogram showing **the population 65 years and over with Medicare coverage only to total population** in the green and purple communities

Multiple other predictors integrated into the model were found to be statistically significant with p-value ≤ 0.05 and could be responsible for the discrepancy between the green and purple communities in terms of the number of deaths. At any time, new predictors can be added to the model to find additional features that might be responsible either for similarities or dissimilarities in the patterns within selected communities.

3.2. Percolation-detected communities

The clique percolation method was applied to our graph to detect communities that might not be obviously identifiable when only visual exploration techniques are used. Figure 43 illustrates the result of the automatic percolation community search algorithm. We focus our **analysis on the comparison of blue and yellow communities, since they showed similar patterns** in the number of confirmed cases and number of deaths (see colors of nodes in Figure 27 and Figure 28). Also, concentrating our analysis on differences between these two communities may enable us to discover reasons why the TDA algorithm has split green and red communities, which were among the four Girvan-Newman communities described earlier (see Figure 24).

To analyze these two communities, statistical tests were performed and a predictor table consisting of 267 rows with corresponding p-values was calculated. In this predictor table, 111 p-values were found to be statistically significant p-values ≤ 0.05.

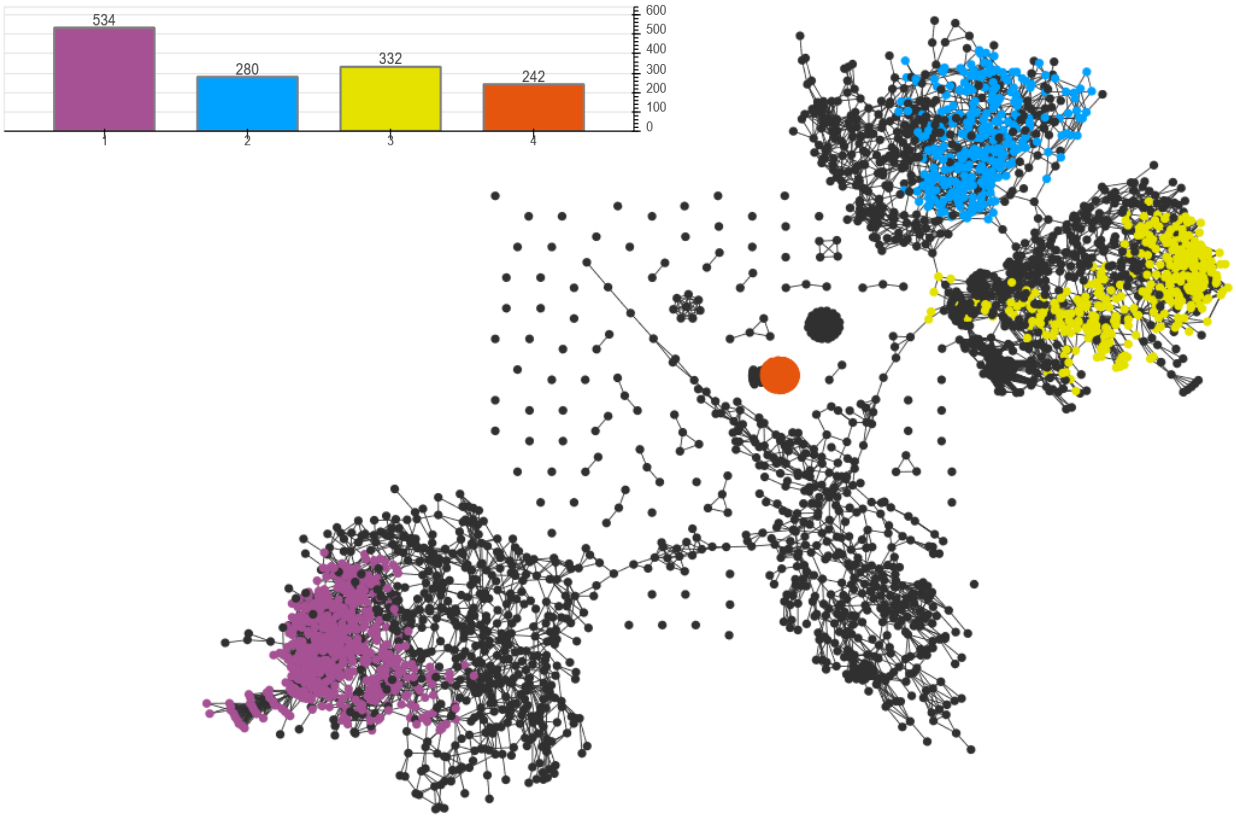


Figure 43. Communities detected on the graph using the clique percolation method

We further compared these two communities using the statistical methods described above to analyze various predictors integrated into the model. Figure 44 shows a histogram comparing the blue and yellow communities based on the population density predictor, which was statistically significant with the p-value ≤ 0.05. The blue community appeared to have a higher population density than the yellow one.

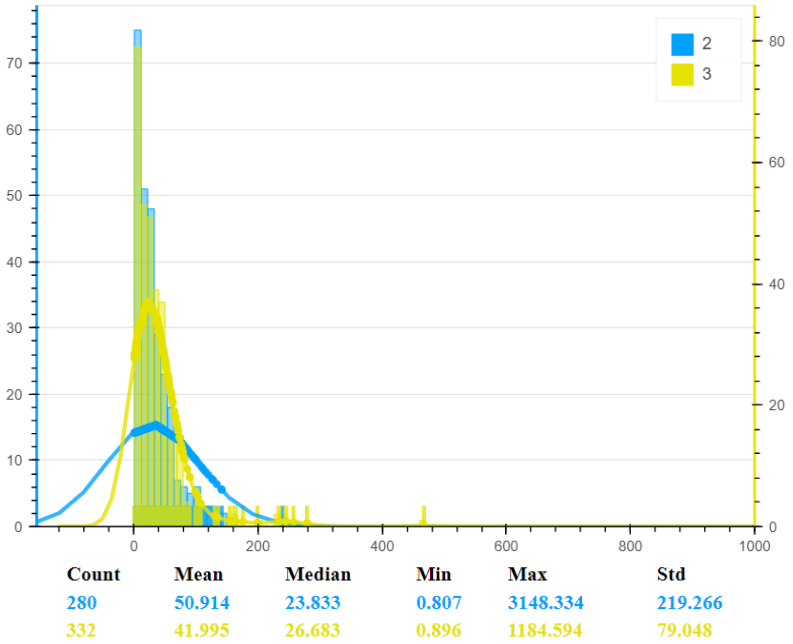
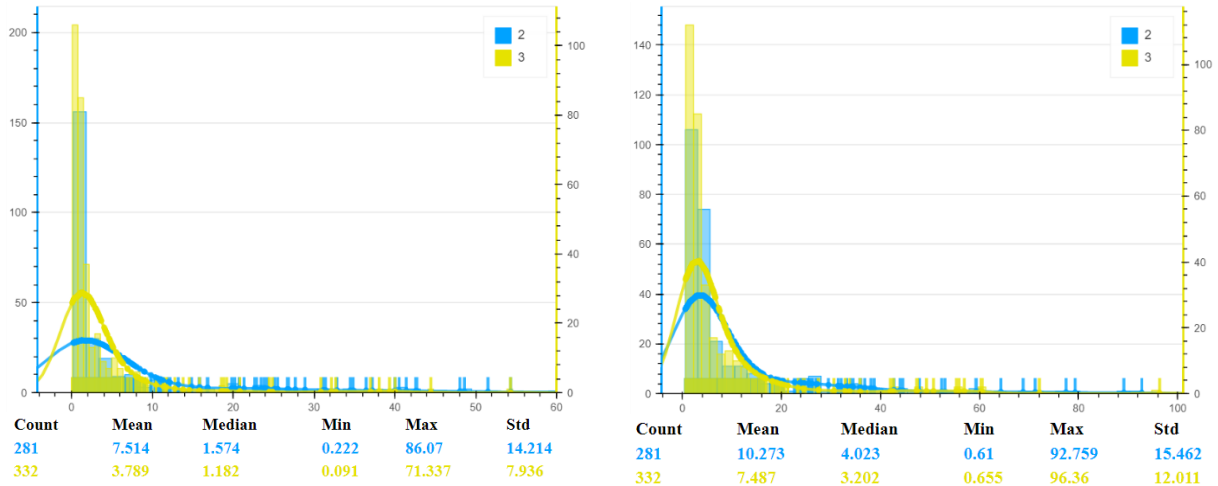


Figure 44. The histogram comparing the blue and the yellow communities based on population density

Also, several predictors related to the race or ethnicity of the population are statistically significant. Figure 45 shows histograms comparing the blue and the yellow communities according to the percentage of different races and ethnicities. The blue community appeared to have a higher percentage of African American and Hispanic populations than the yellow one.



a) Percent of African American people b) Percent of Hispanic population

Figure 45. The histogram comparing the blue and yellow communities based on the percentage of different races to population within the communities

The next group of predictors that might be responsible for the difference of the blue and the yellow communities relates to the healthcare system.

The histograms shown in Figure 46 and Figure 47 compare these communities according to the number of hospitals per 100,000 people and the number of professionally active specialist at emergency medicine per 1,000 people. The histogram analysis shows that the number of hospitals is greater in the blue community, while the number of specialists at emergency medicine is greater in the yellow community. In addition, the histogram in Figure 48, showing the percent of population with no health insurance coverage, allows us to conclude that there are more people without insurance coverage in the blue community.

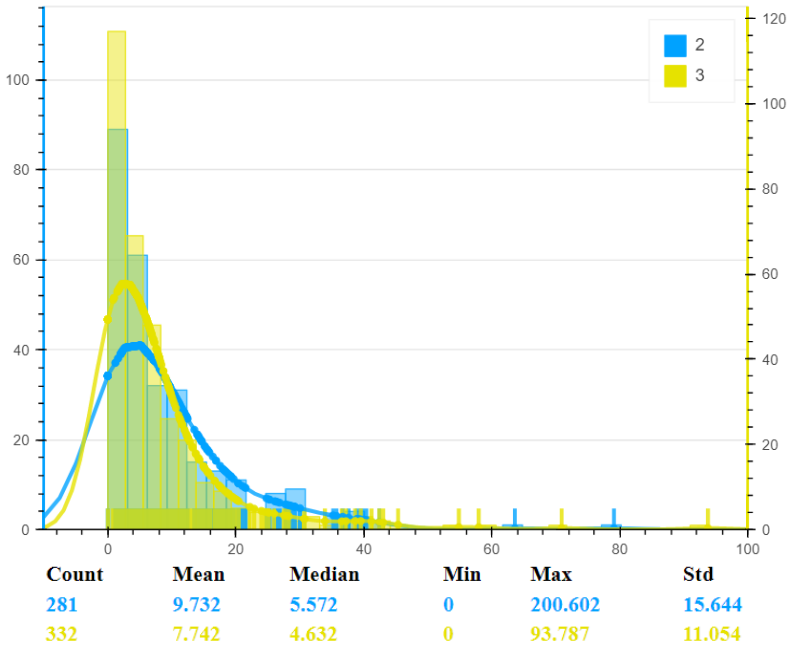
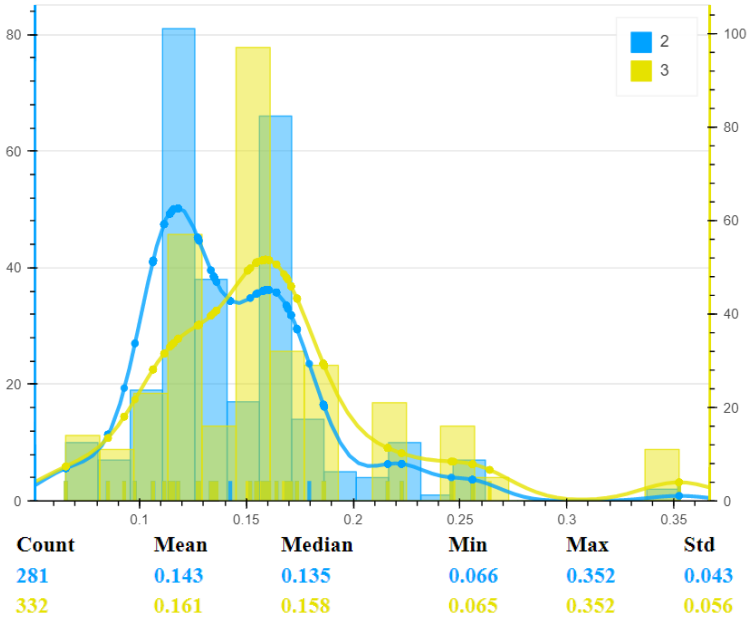
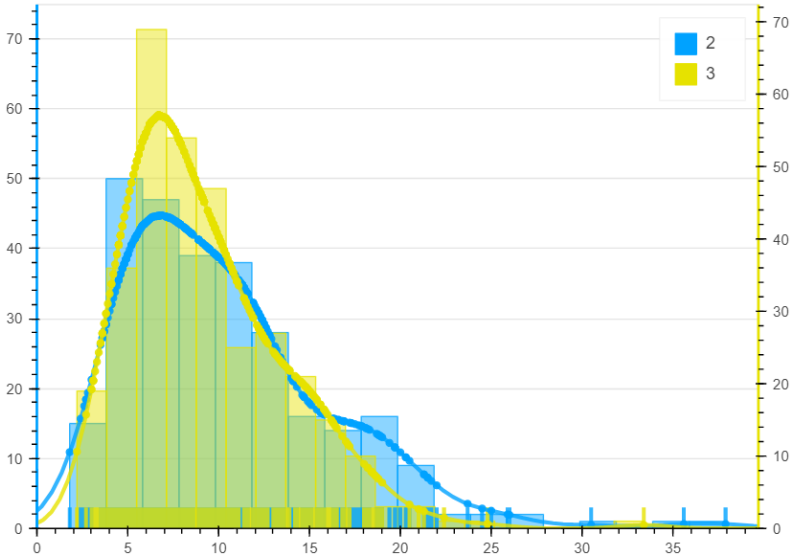


Figure 46. The histogram showing the number of hospitals per 100,000 people in the blue and yellow communities



Count	Mean	Median	Min	Max	Std
281	0.143	0.135	0.066	0.352	0.043
332	0.161	0.158	0.065	0.352	0.056

Figure 47. The histogram showing **the number of professionally active specialist at emergency medicine per 1,000 people** in the blue and the yellow communities



Count	Mean	Median	Min	Max	Std
281	10.475	9.4	1.8	37.9	5.735
332	9.305	8.35	2.2	33.4	4.38

Figure 48. The histogram showing the **percent of population with no health insurance coverage** in the blue and the yellow communities

The next statistically significant predictor for the percolation-detected communities appeared to be the number of days which passed between the date on which the stay-at-home restriction was adopted and the start day of the observation interval, i.e., the date of detection of the second confirmed case in the county.

Figure 49 shows a histogram comparing the blue and yellow communities according to the number of days which passed between the date on which the stay-at-home restriction was adopted and the start date of detecting confirmed cases in each county. In the yellow community (with the smaller number of confirmed cases of the disease), the delta (number of days) between the stay-at-home restriction date and the start day of the observation interval appeared to be negative in most counties. This means that the stay-at-home restriction was adopted before the second confirmed case of the disease was detected in these counties (see Figure 49). It was determined that in the blue community counties, the stay-at-home restriction was adopted after the start day.

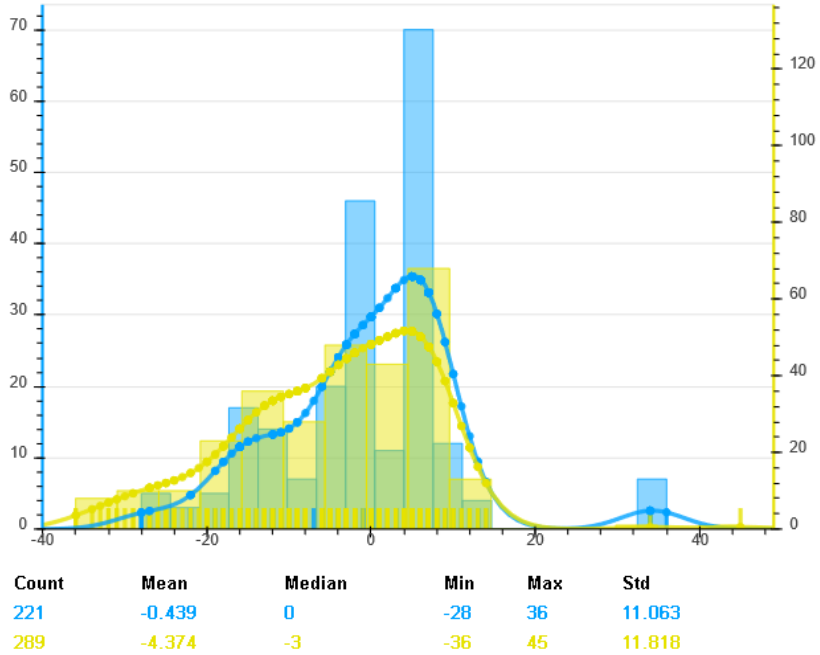


Figure 49. The histogram comparing the yellow and blue communities according to **the number of days which passed between the date on which the stay-at-home restriction** was adopted and the start date of detection of confirmed cases in each county

In the next step, the graph was re-colored based on the delta between the date on which restrictions were introduced for bars and restaurants and the start day of the observation interval (see a histogram comparing the yellow and blue communities in Figure 50). In the yellow community, the bars and restaurants were closed earlier with respect to the start day than they were in the blue community.

The data mining experiments indicated various predictors that could influence the spread of COVID-19 in the USA. These include geographic conditions, population density, urban influence, public transportation, highway length, migration, number of hospitals, number of general practice physicians, number of days passed between introducing the stay-at-home restriction and the start date of detection of confirmed cases of the disease, and many others. At any time, additional predictors can be easily integrated into the model to expand the search of parameters that might be responsible for similarities in pandemic spread discovered by using TDA, machine learning algorithms, and visual exploration techniques.

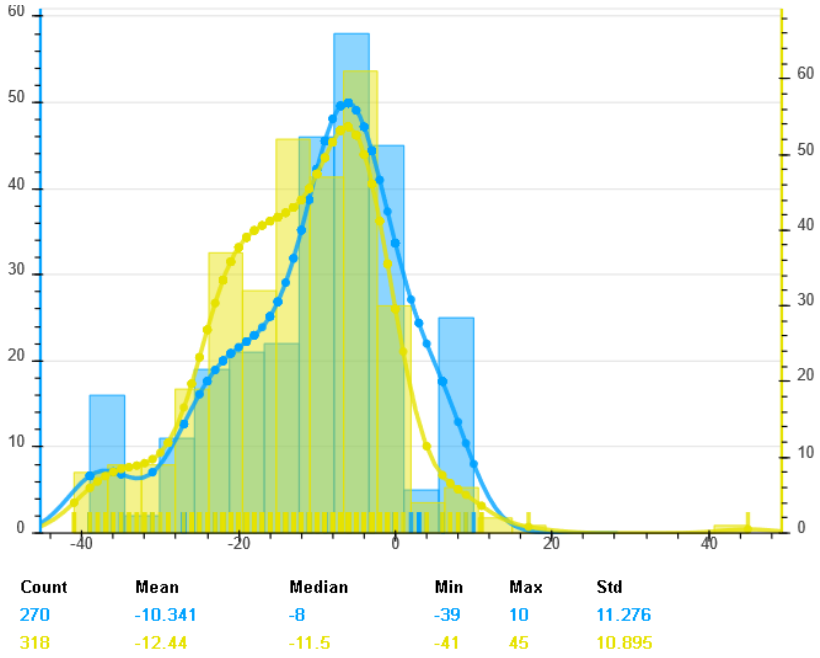


Figure 50. The histogram comparing the yellow and blue communities according to **the number of days which passed between the date on which restrictions were introduced for bars and restaurants and the start date** of detection of confirmed cases in each county

CONCLUSION

In this paper, the authors present novel machine-learning techniques and workflow that rely on graphs as the fundamental tool to structuring and analyzing complex real-world data. The authors illustrate this workflow by analyzing how the spread of COVID-19 has advanced over the USA, looking for similarities in a specific timeframe. The analysis was performed on a county-by-county basis, by extracting from the dataset a topological data model represented as a graph in which each of 3,142 nodes corresponds to one county, and two nodes are connected if they share similarities. After the graph was built, real-world data was used to integrate into the model predictors which might be responsible for similarities in the early-stage spread of the pandemic. Over 250 predictors from different publicly available sources were used during the course of the experiment. The data mining experiments indicated various predictors that could influence the spread of COVID-19 in the USA. These include geographic conditions, population density, urban influence, public transportation, highway length, migration, number of hospitals, number of general practice physicians, number of days passed between introducing the stay-at-home restriction and the start date of detection of confirmed cases of the disease, and many others. At any time, additional predictors can be easily integrated into the model to expand the search of parameters that might be responsible for similarities in pandemic spread discovered by using TDA, machine learning algorithms, and visual exploration techniques.

REFERENCES

[1] G. Carlsson, "Topology and data," *Bulletin of the American Mathematical Society,* vol. 46, no. 2, pp. 255-308, 2009.

[2] A. Zomorodian, Topology for Computing, Cambridge: Cambridge University Press, 2005.

[3] K. Drach and I. Kotenko, "Using the geometric properties of datasets to analyze the accuracy of COVID-19 forecasting models," in *Pharmaceutical Users Software Exchange 2022 (Phuse US Connect, May 1-4, 2022, Atlanta, USA)*, 2022.

[4] S. Glushakov, I. Kotenko and A. Rekalo, "Handling missing data in clinical trials using Topological Data Analysis," in *Pharmaceutical Users Software Exchange 2018 (Phuse EU Connect, November 4-7, 2018, Frankfurt, Germany)*, 2018.

[5] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences,* vol. 103, no. 23, p. 8577–8696, 2006.

[6] H. Zenil, S. Hernández-Orozco, N. Kiani, F. Soler-Toscano, A. Rueda-Toicen and J. Tegnér, "A Decomposition Method for Global Evaluation of Shannon Entropy and Local Estimations of Algorithmic Complexity," *Entropy,* vol. 20, no. 8, p. 605, 2018.

[7] K. Drach and I. Kotenko, "Cluster analysis vs. graph-based machine learning: the battle of the strongest in pattern recognition," in *Pharmaceutical Users Software Exchange 2023 (PHUSE US Connect, March 5-8, Orlando, USA)*, Orlando, 2023.

[8] S. Fortunato, "Community detection in graphs," *Physics Reports,* vol. 486, no. 3-5, pp. 75-174, 2010.

[9] S. Fortunato and D. Hric, "Community detection in networks: a user guide," *Physics Reports,* vol. 659, pp. 1-44, 2016.

[10] S. Glushakov and K. Drach, "Sub-population Detection Using Graph-based Machine Learning," in *Pharmaceutical Users Software Exchange 2019 (Phuse EU Connect, November 10-13, 2019, Amsterdam, The Netherlands)*, 2019.

[11] M. Girvan and M. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences,* vol. 99, no. 12, pp. 7821-7826, 2002.

[12] G. Palla, I. Derenyi, I. Farkas and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature,* vol. 435, pp. 814-818, 2005.

[13] "COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University," [Online]. Available: https://github.com/CSSEGISandData/COVID-19.

[14] "A List of Statewide "Stay-at-home" Orders," [Online]. Available: https://www.littler.com/publication-press/publication/stay-top-stay-home-list-statewide.

[15] M. Torok, "Focus on Field Epidemiology," *North Carolina Center for Public Health Preparedness,* 2003.

[16] "Topographical map of United States," [Online]. Available: https://us-canad.com/topographic-map-usa.html.

ACKNOWLEDGMENTS

We would like to acknowledge **Oleksandr Leonov** (Kharkiv National University, Ukraine), **Lyudmyla Polyakova** (Kharkiv National University, Ukraine), and **Victoriia Shevtsova** (Intego Group, Ukraine) for handling the computational experiment and being the core members of the research team. Without you, this research and the experiment would not have been possible.

We would also like to acknowledge Illia Skliar, Anna Petrenko and Natalia Averianova, students on the Kharkiv National University's Biostatistical Programming training program, who helped in gathering data and preparing the database of predictors used during the course of the experiment.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the group of authors at:

Contact: Sergey Glushakov

Company: Intego Group

Address: 2300 Maitland Center Pkwy, Suite 240, Maitland, FL 32751, USA

Work Phone: +1 407 512-1006

Email: sergey.glushakov@intego-group.com

Web: www.intego-group.com