



INTEGO
GROUP

UNDERSTANDING THE SPREAD OF COVID-19 IN THE UNITED STATES USING TOPOLOGY AND MACHINE LEARNING FOR TIME SERIES



EU Connect 2020

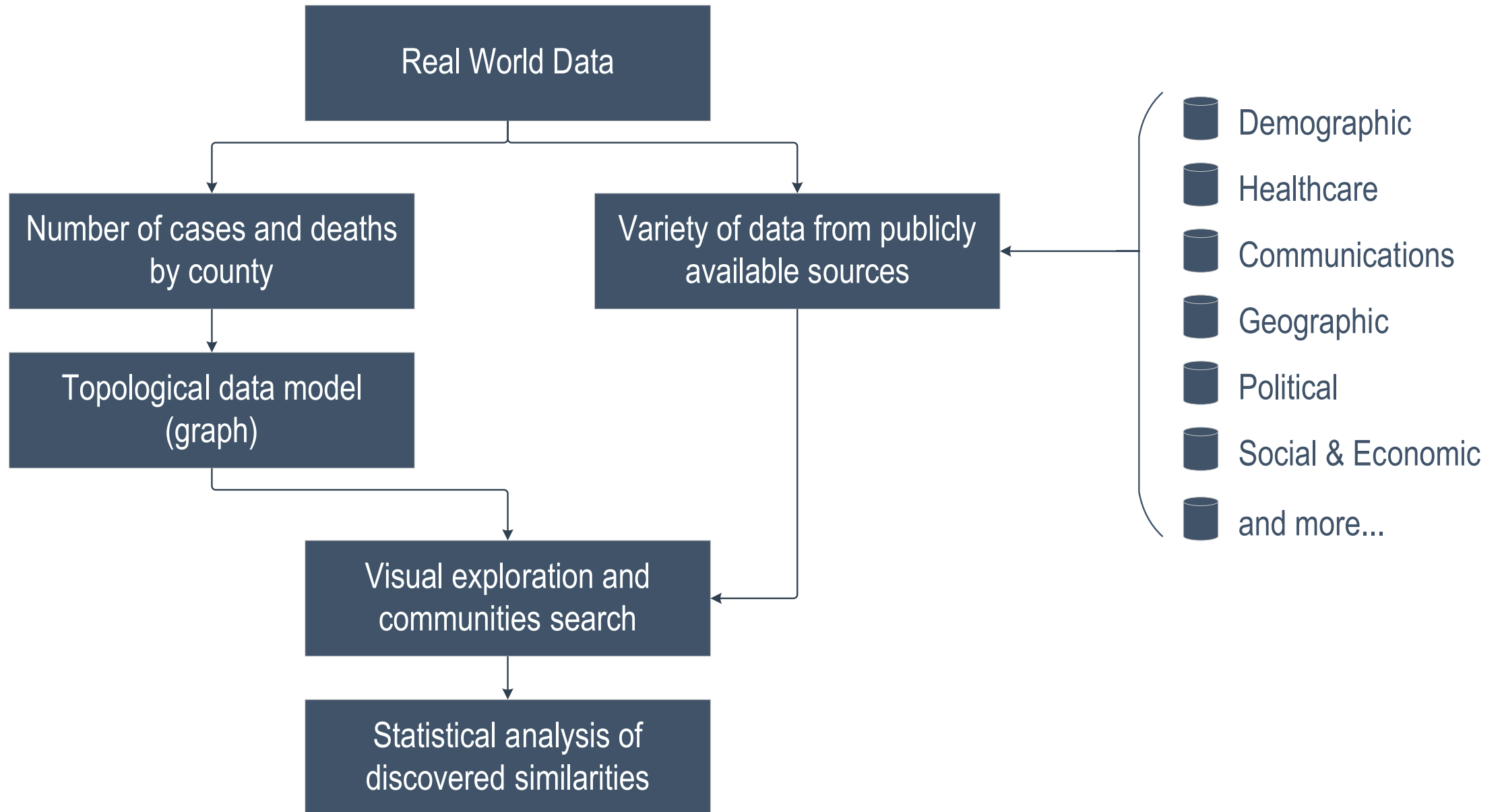


The goal of the experiment:

- Introduce machine-learning techniques that rely on graphs as the fundamental approach to structuring and analyzing complex real-world data
- Analyze how the spread of COVID-19 has advanced in every county of the United States, looking for similarities in a specific time interval
- Focusing on the geometric properties of datasets, unveil a set of unrelated features that could have caused the similarity in the pandemic spread

Experiment step-by-step:

- 1) Build the model to integrate the number of **cases and deaths** in every county of the US
- 2) Build **the graph** to unveil geometric properties of the dataset and hidden interdependencies
- 3) Integrate **real-world data** from a variety of publicly available sources
- 4) Perform **visual exploration** and automatic search for communities
- 5) Statistically analyze discovered patterns to **explain the similarity** of pandemic spread



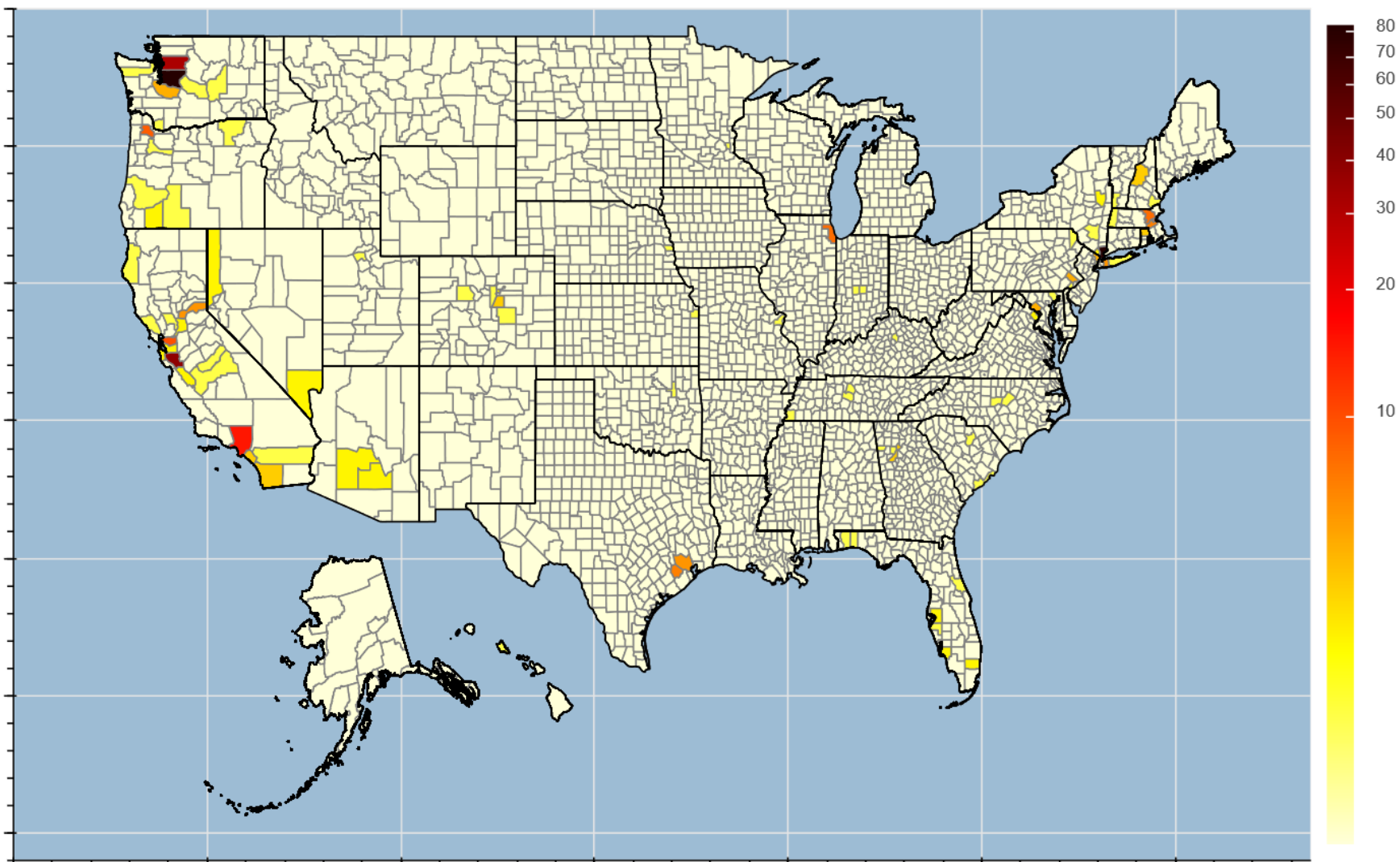


STEP 1.
NUMBER OF CASES AND DEATHS
OVER THE US BY COUNTIES



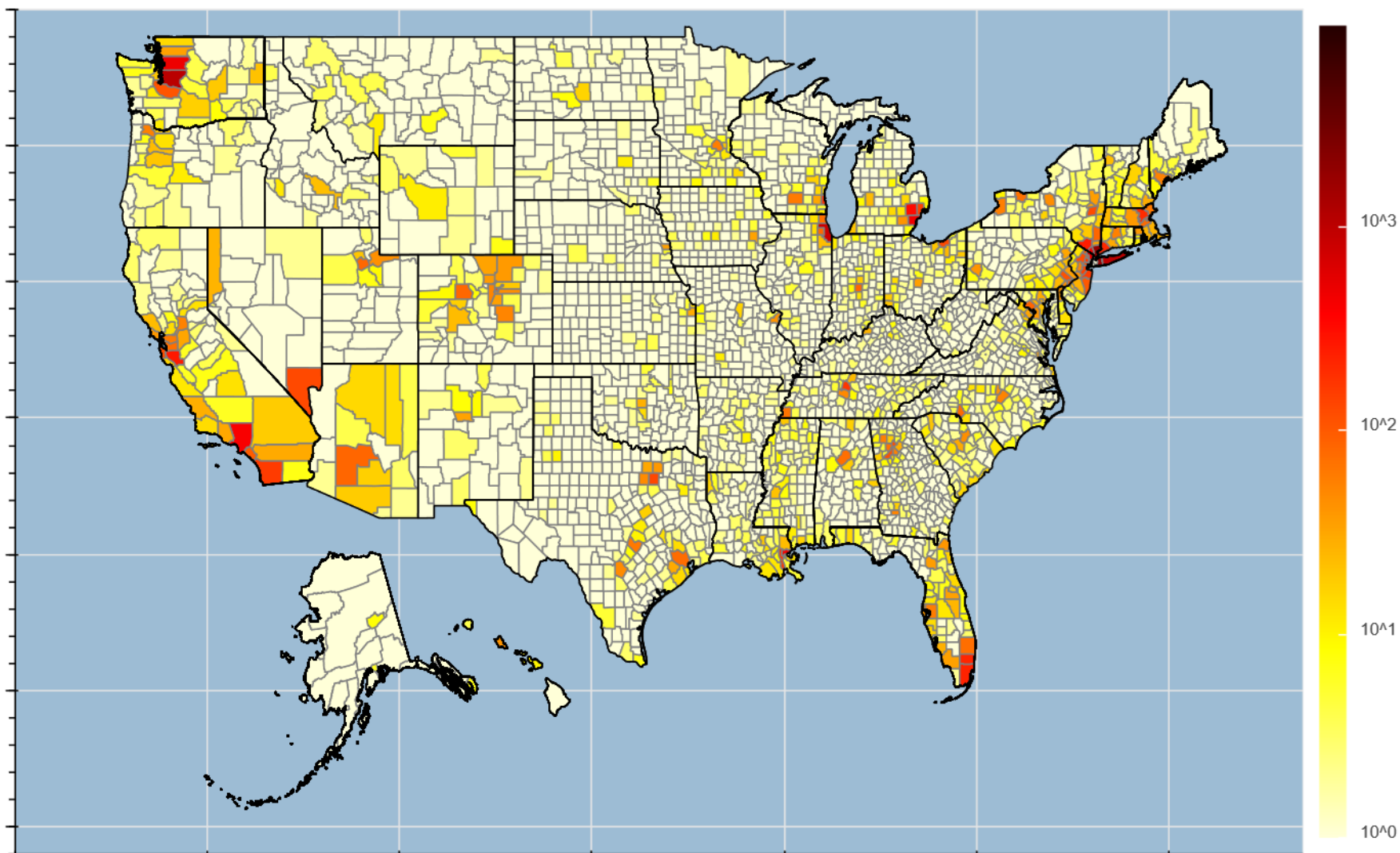
Number of confirmed cases 451

Day 47, March 8



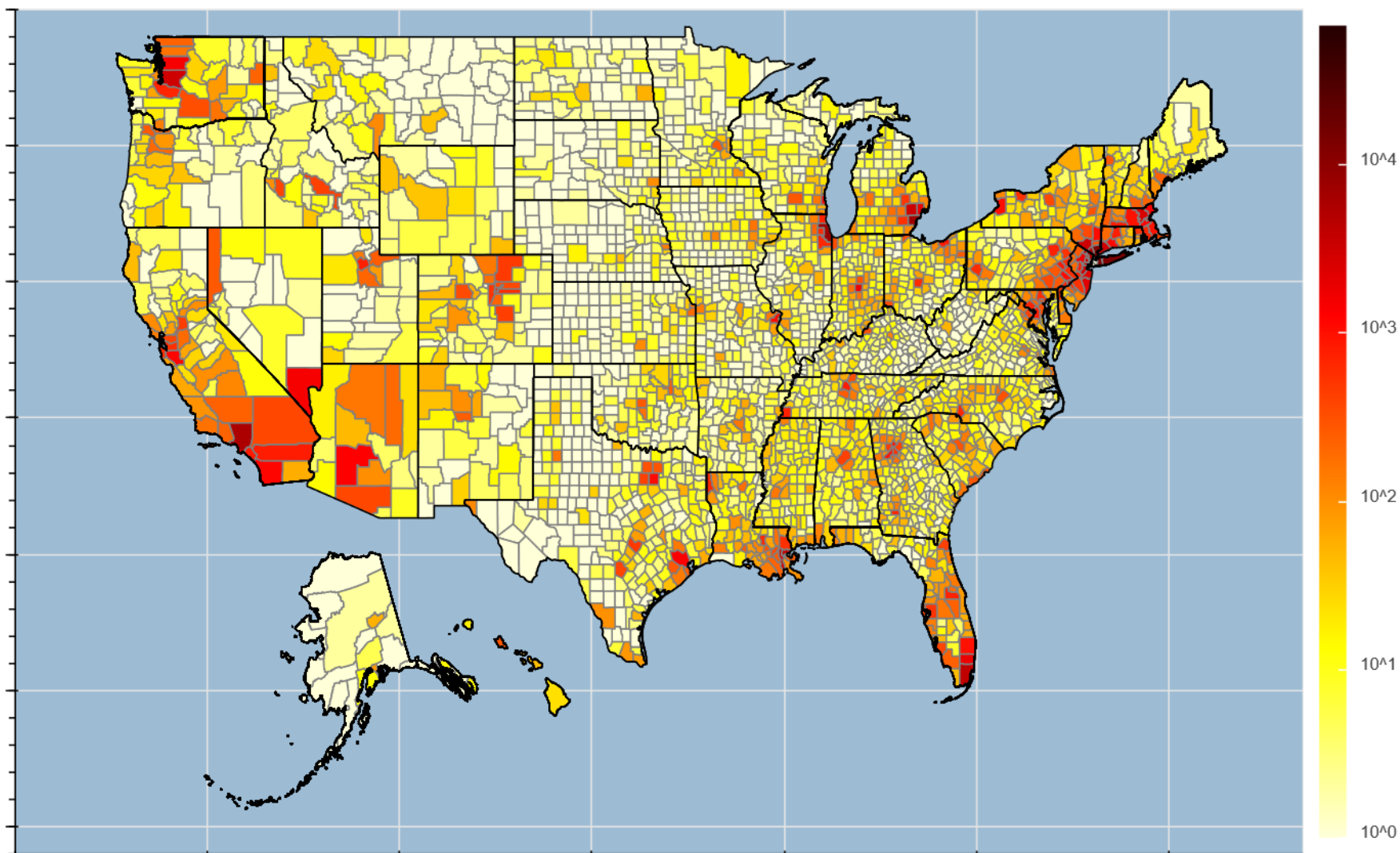
Number of confirmed cases 32,899

Day 61, March 22



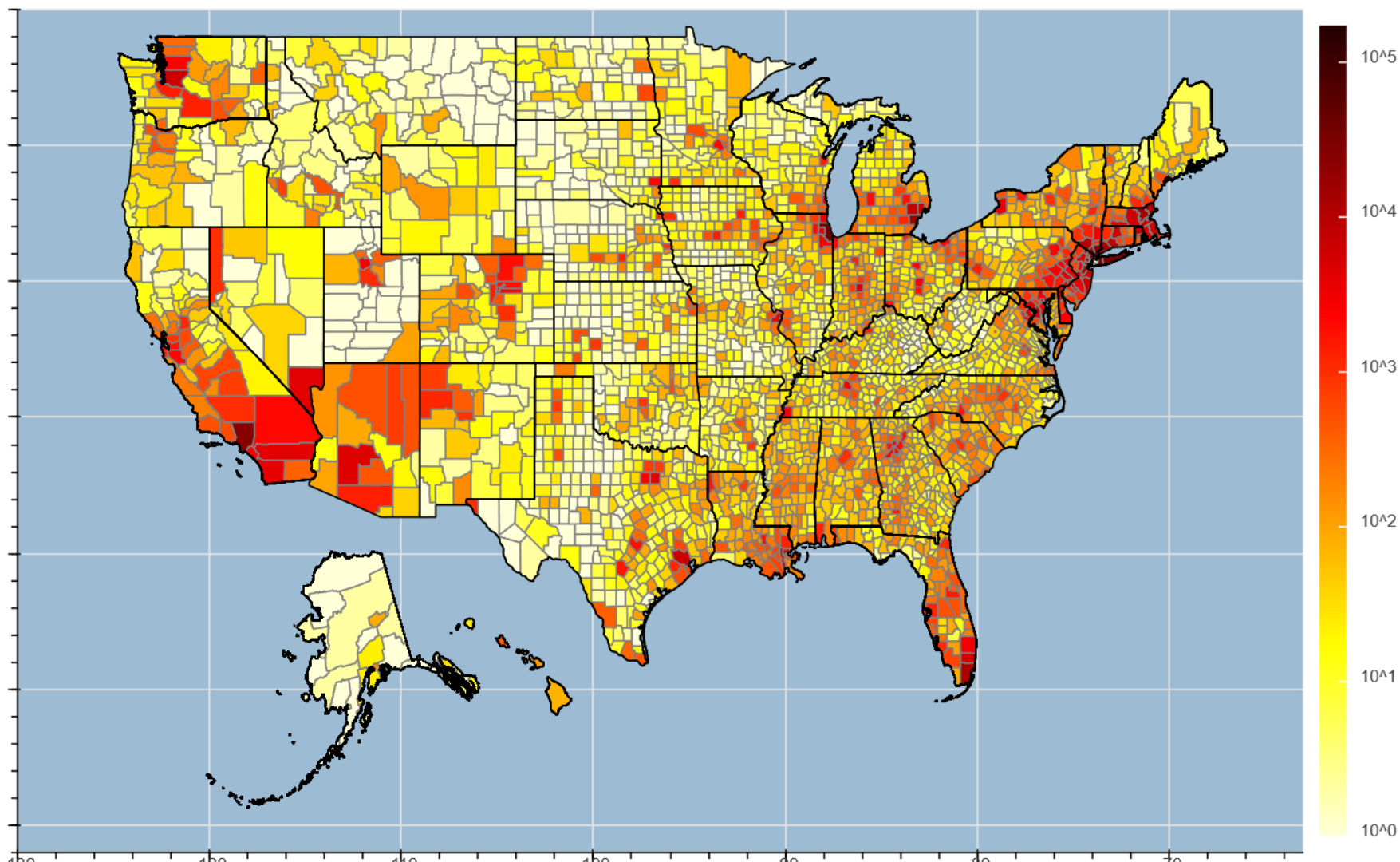
Number of confirmed cases **330,384**

Day **75**, April 5

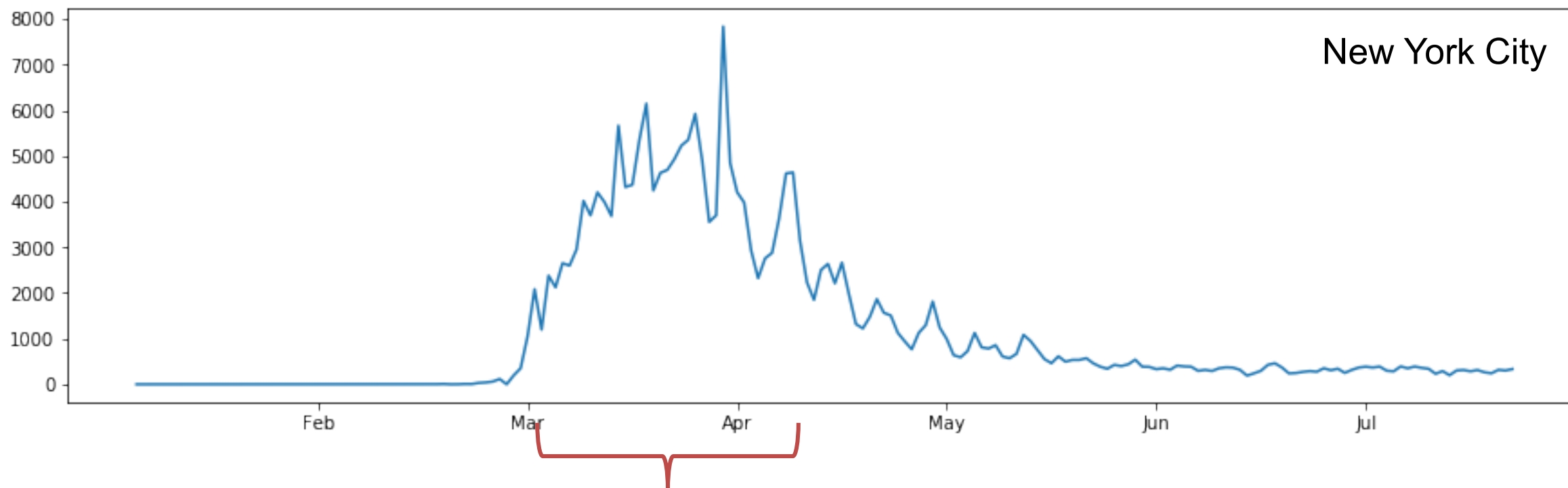


Number of confirmed cases **1,151,933**

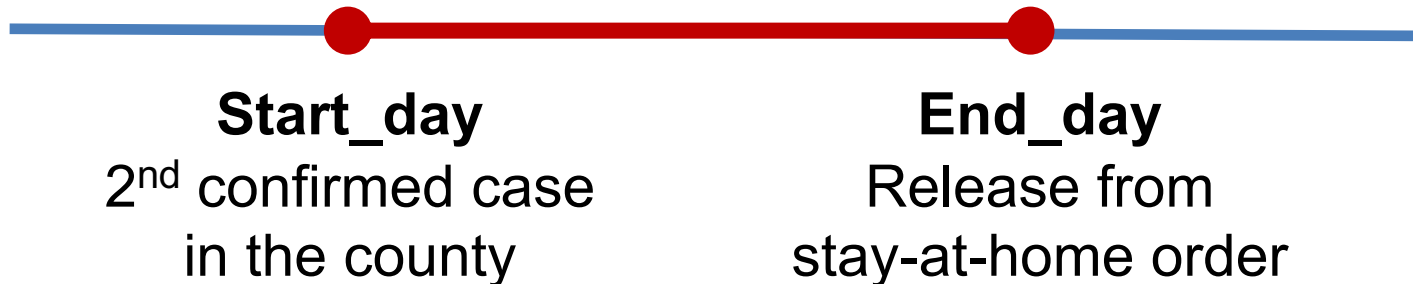
Day **103**, May 3



Time series observation interval



Interval=39 days





STEP 2.

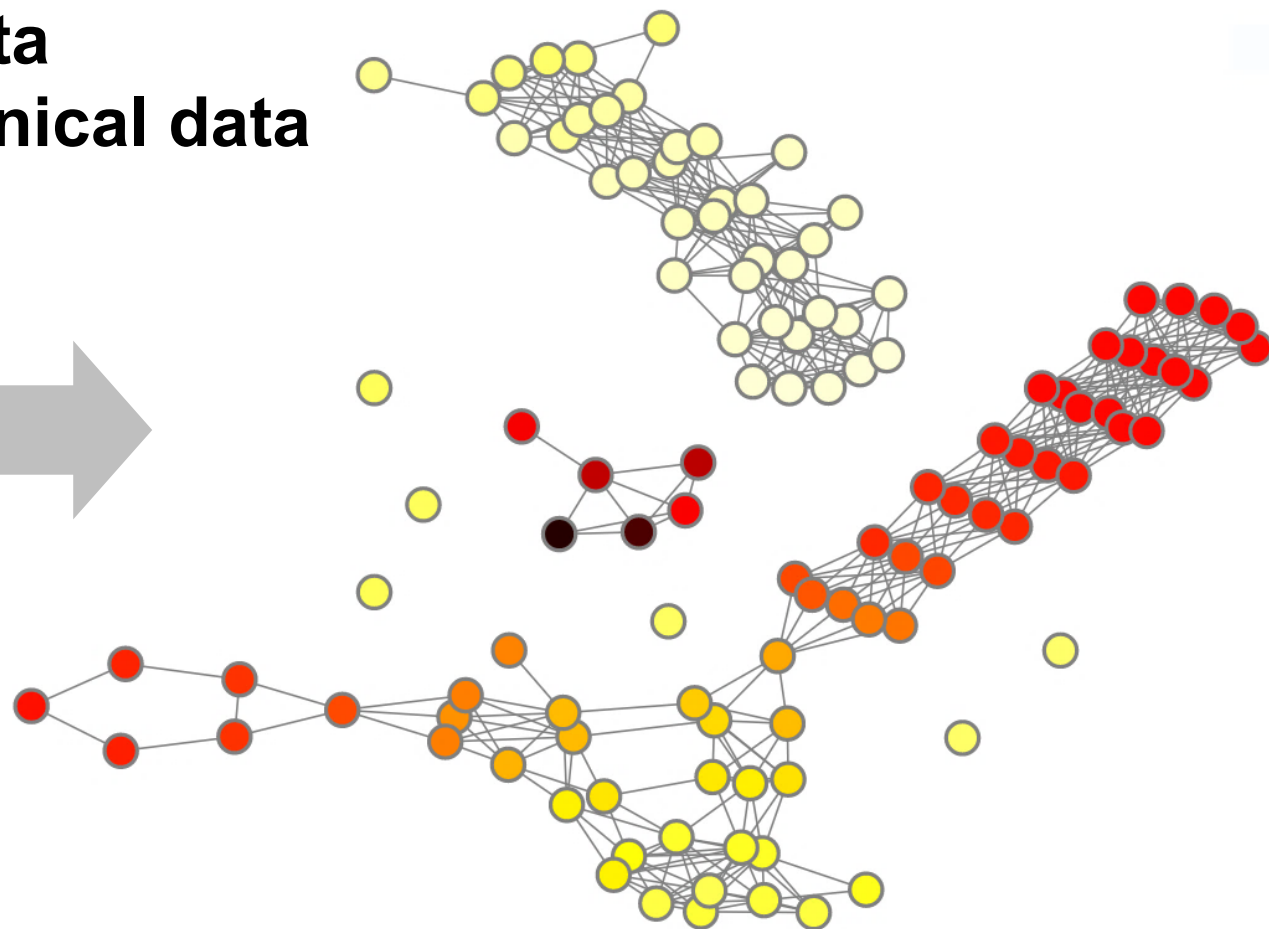
**BUILDING THE GRAPH USING
TOPOLOGICAL DATA ANALYSIS**



Topological Data Analysis for clinical data



Each **node** represents one
subject



Two nodes are connected
if they are **similar**, within **stratifying
bins**, in terms of **outcomes**

In our experiment:

- Each node corresponds to a county of the US
- Outcomes: all time series are within the selected observation interval of 39 days



$$\text{Person} = \left(\begin{array}{c} \text{Confirmed cases, daily} \\ \text{(aggregated within sliding window of 3 days, normalized)} \end{array}, \log(\text{Total \# of Cases} + 1), \begin{array}{c} \text{Death cases, daily} \\ \text{(aggregated within sliding window of 3 days, normalized)} \end{array}, \log(\text{Total \# of Deaths} + 1) \right)$$

Total number of entries per each county is $39/3 + 1 + 39/3 + 1 = 28$

Thus, the relevant information about the spread of COVID19 in a county is encoded by a **28-dimensional row vector**

(first 14 entries encode confirmed cases, second 14 entries encode deaths)

$$\text{Person} = \overbrace{\begin{bmatrix} x_1 & \dots & x_{14} & x_{15} & \dots & x_{28} \end{bmatrix}}^{28}$$

In our experiment:

- Stratifying bins arrange counties into several overlapping groups. Bins are chosen based on **density** and **centrality** of nodes with respect to outcomes.
- Similarity is then measured within each bin separately.



x_1	...		x_{14}	x_{15}	...		x_{28}
-------	-----	--	----------	----------	-----	--	----------



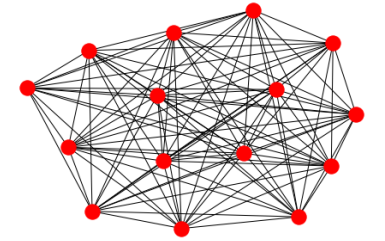
y_1	...		y_{14}	y_{15}	...		y_{28}
-------	-----	--	----------	----------	-----	--	----------

- **Distance function based on Euclidean distance** is used to measure the similarity between counties

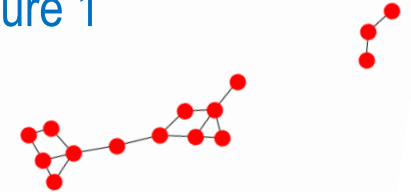
$$d(\text{red person}, \text{yellow person}) = \sqrt{(x_1 - y_1)^2 + \dots + (x_{14} - y_{14})^2} + \sqrt{(x_{15} - y_{15})^2 + \dots + (x_{28} - y_{28})^2}$$

Selection of the most representative topological model: Kolmogorov complexity

- Based on available parameters of our model (number and placement of stratifying bins, parameters in density and centrality estimates, etc.) the most representative graph was chosen using the **Kolmogorov complexity**.
- Kolmogorov complexity, originated in theoretical computer science, measures how well the object (text, array, graph, etc.) can be compressed.
- Graphs with **regular structure** have **low** Kolmogorov complexity (picture 1).
- Graphs with **random structure** have **high** Kolmogorov complexity (picture 2).



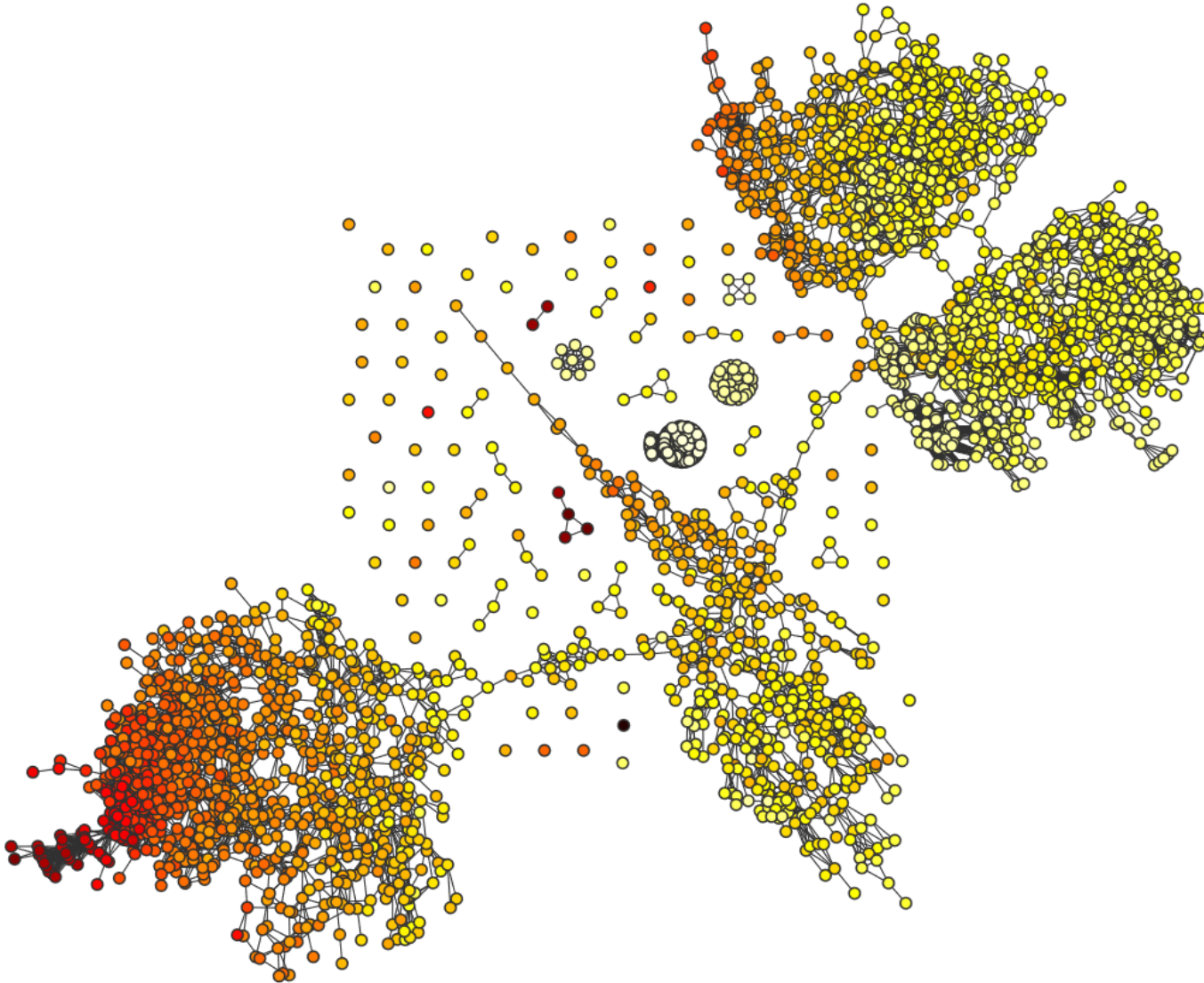
Picture 1



Picture 2

In our experiment, Kolmogorov complexity was measured within each stratifying bin separately, and the graphs with high complexity were selected. For these graphs the nodes within the bins tend to group in approximately equal size clusters.

Topological model of COVID-19 spread over the US



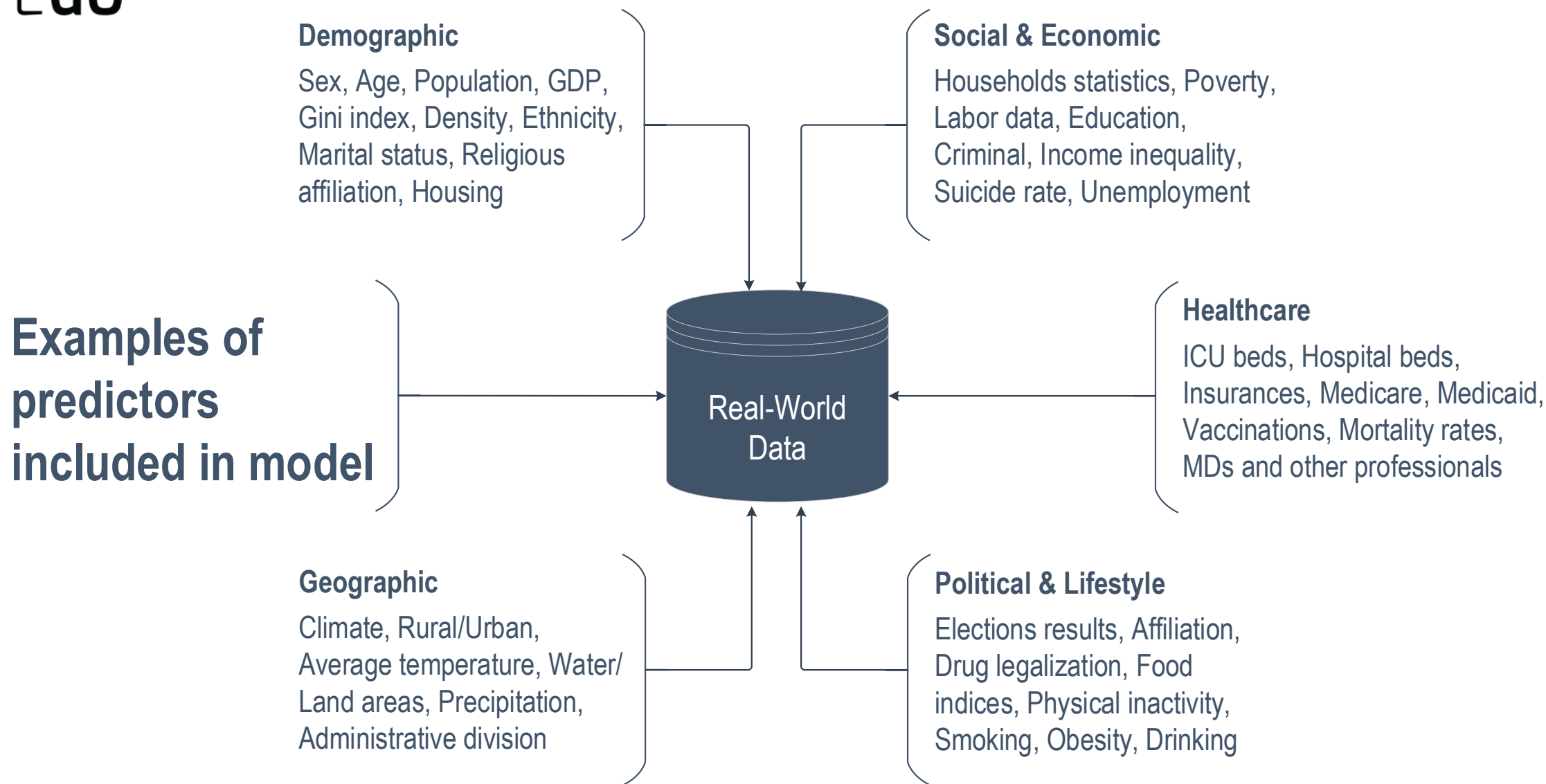
- Each node on the **graph** corresponds to one county of the United States (total **3,142 nodes**)
- Two nodes are connected with the edge if they are **similar** in terms of predefined outcomes
- Outcomes integrate the number of confirmed **cases and deaths** within a **39-day observation interval** from the start of the pandemic
- **Similarity** is measured by a modified Euclidean distance function between time series of outcomes
- The resulting graph is selected based on estimates for its **Kolmogorov complexity**



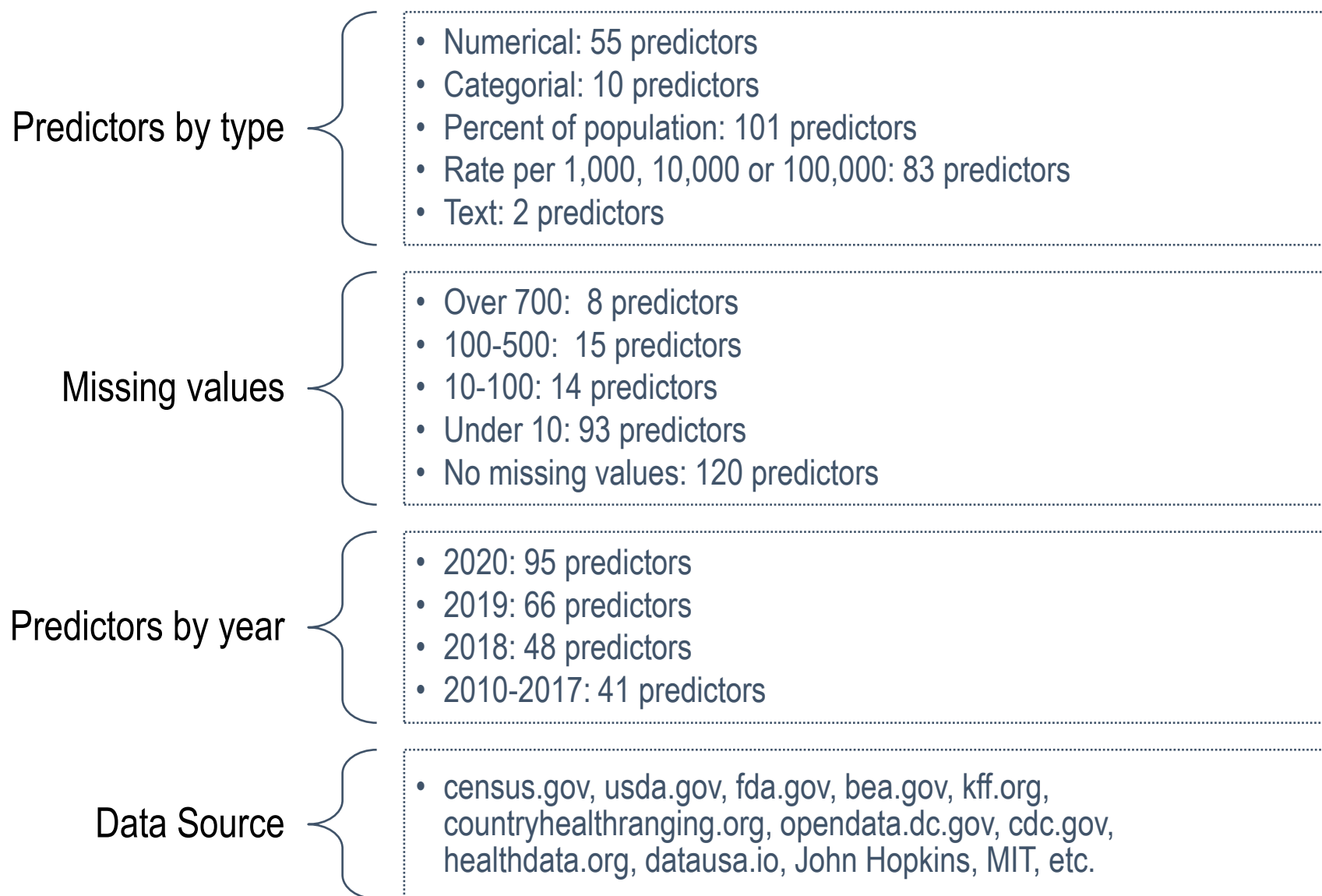
STEP 3.

**REAL-WORLD DATA FROM
PUBLICLY AVAILABLE SOURCES**





250+ predictors were integrated into the model



No limitation on additional predictors after the model is built



STEP 4.

**VISUAL EXPLORATION
AND SEARCH FOR COMMUNITIES**



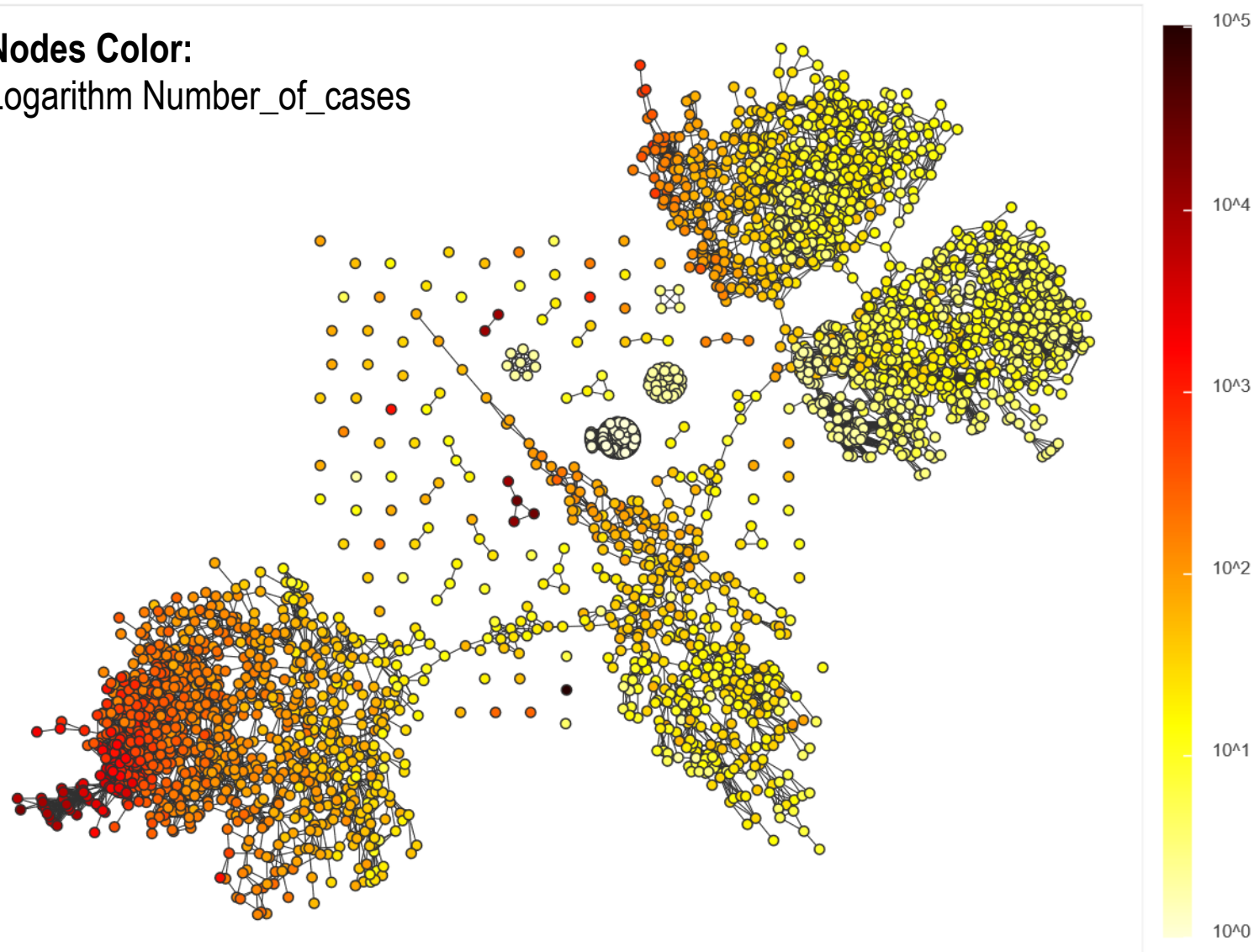
Ways of discovering sup-groups using TDA graphs:

- **visual** discovery of subgroups using **graph features** (connected components, etc.)
- **visual** discovery by changing **color of nodes** (based on outcomes, predictors, composite SQL-queries)
- **automatic** discovery using **Machine Learning algorithms**

Number of confirmed cases

Nodes Color:

Logarithm Number_of_cases

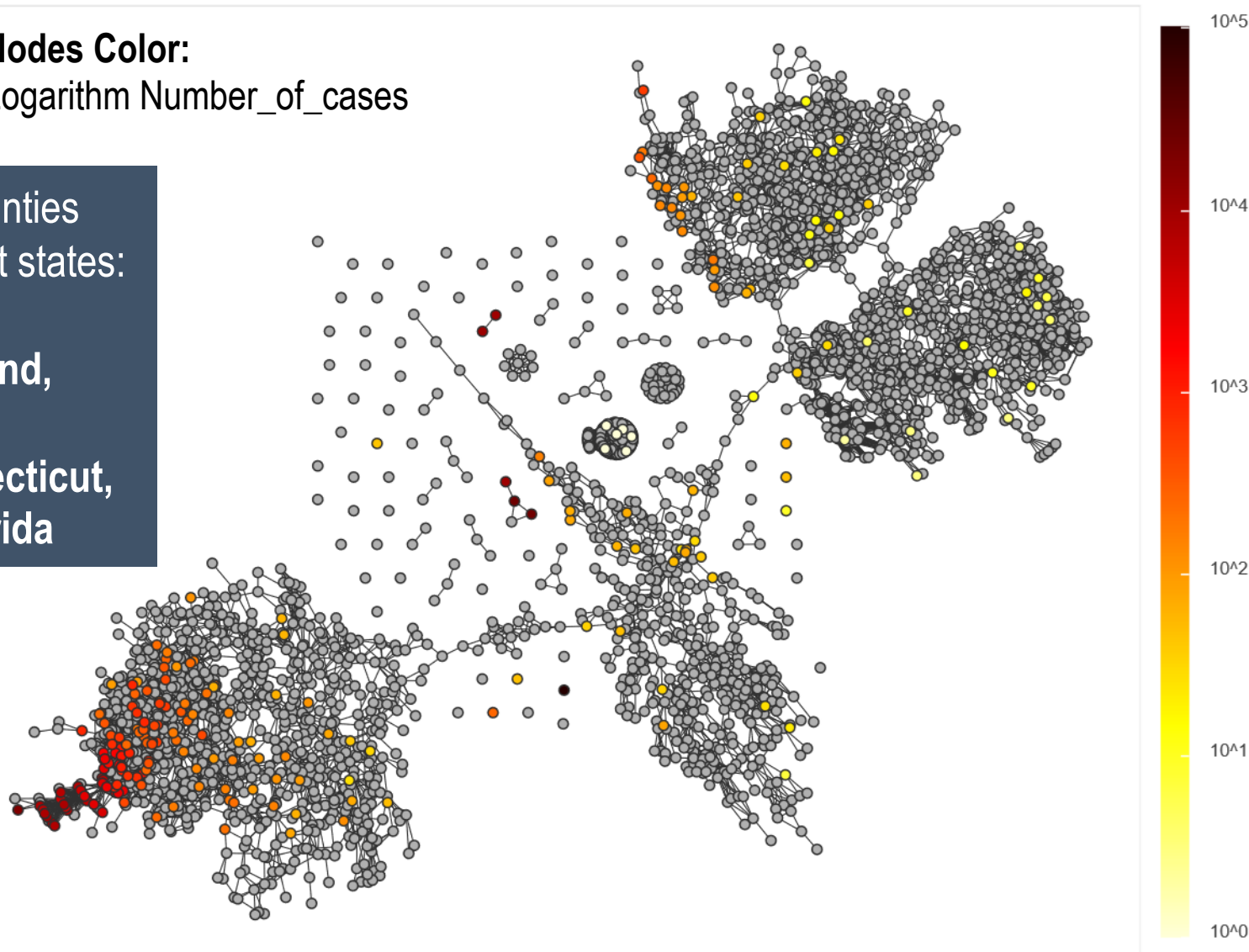


Number of cases (SQL-query East-Coast)

Nodes Color:
Logarithm Number_of_cases

SQL-query highlights the counties
from the following East-Coast states:

**Massachusetts, Rhode Island,
New Hampshire, Vermont,
Maryland, Delaware, Connecticut,
New Jersey, New York, Florida**

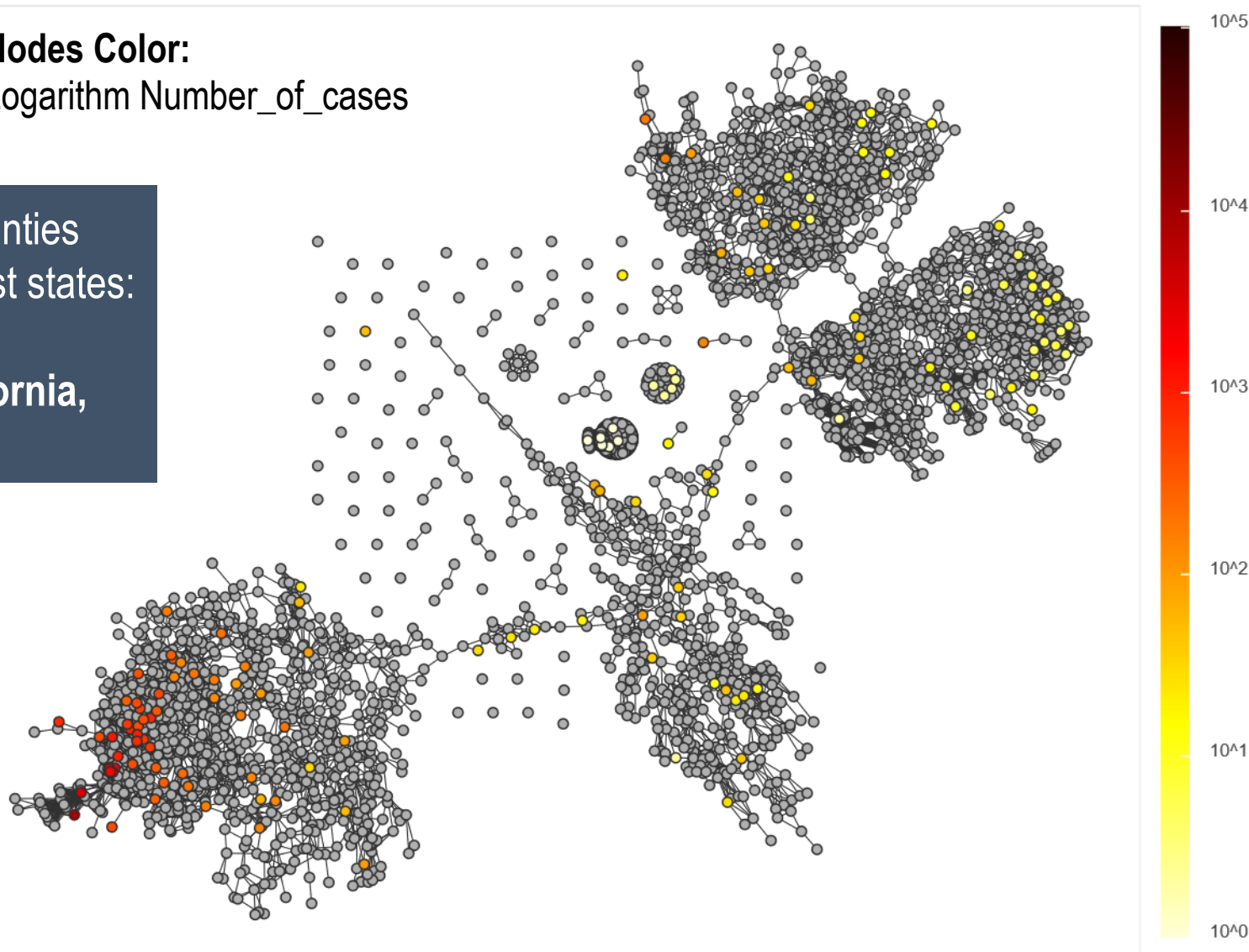


Number of cases (SQL-query West-Coast)

Nodes Color:
Logarithm Number_of_cases

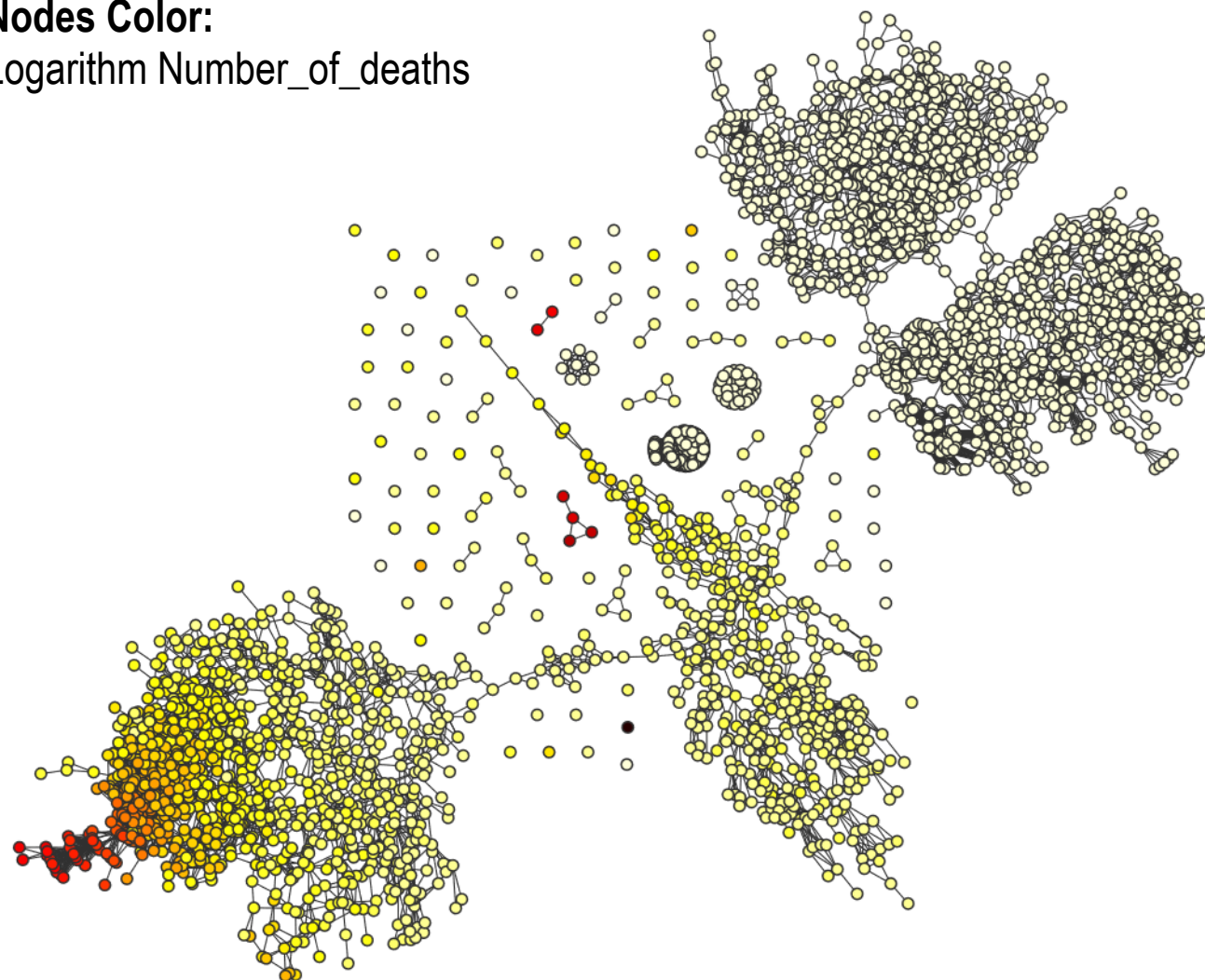
SQL-query highlights the counties
from the following West-Coast states:

**Washington, Oregon, California,
Arizona**



Number of deaths

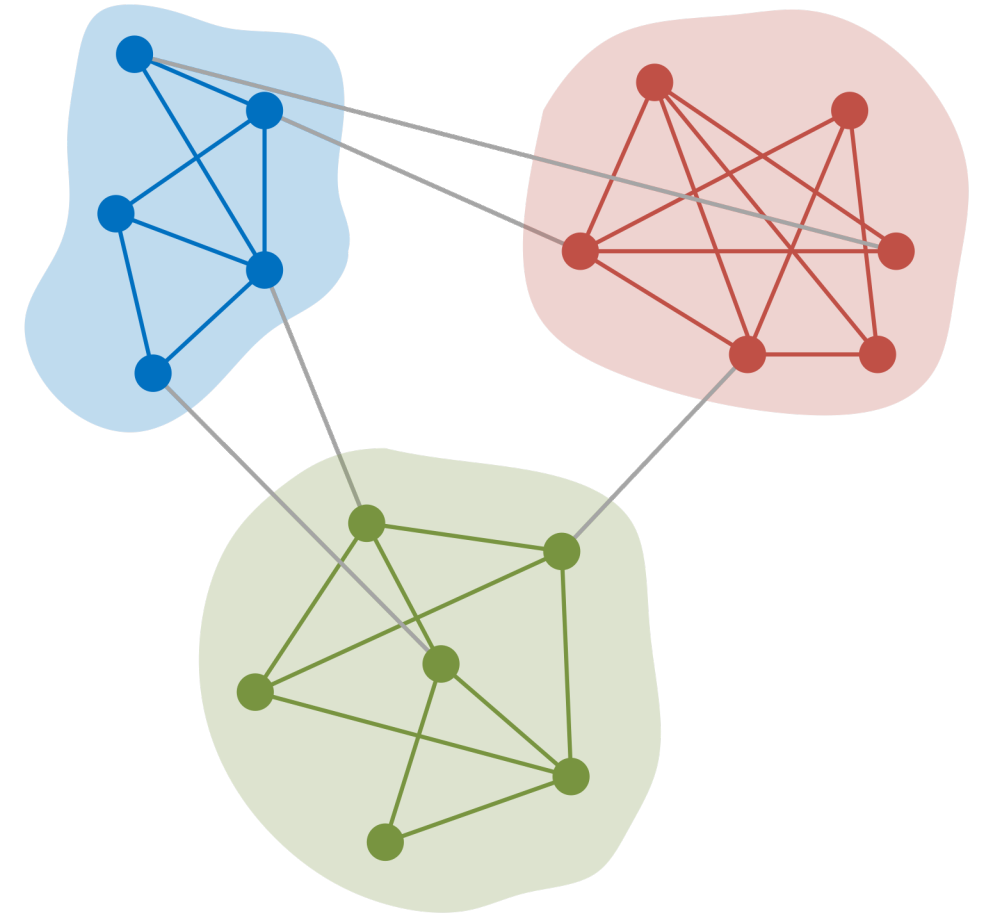
Nodes Color:
Logarithm Number_of_deaths



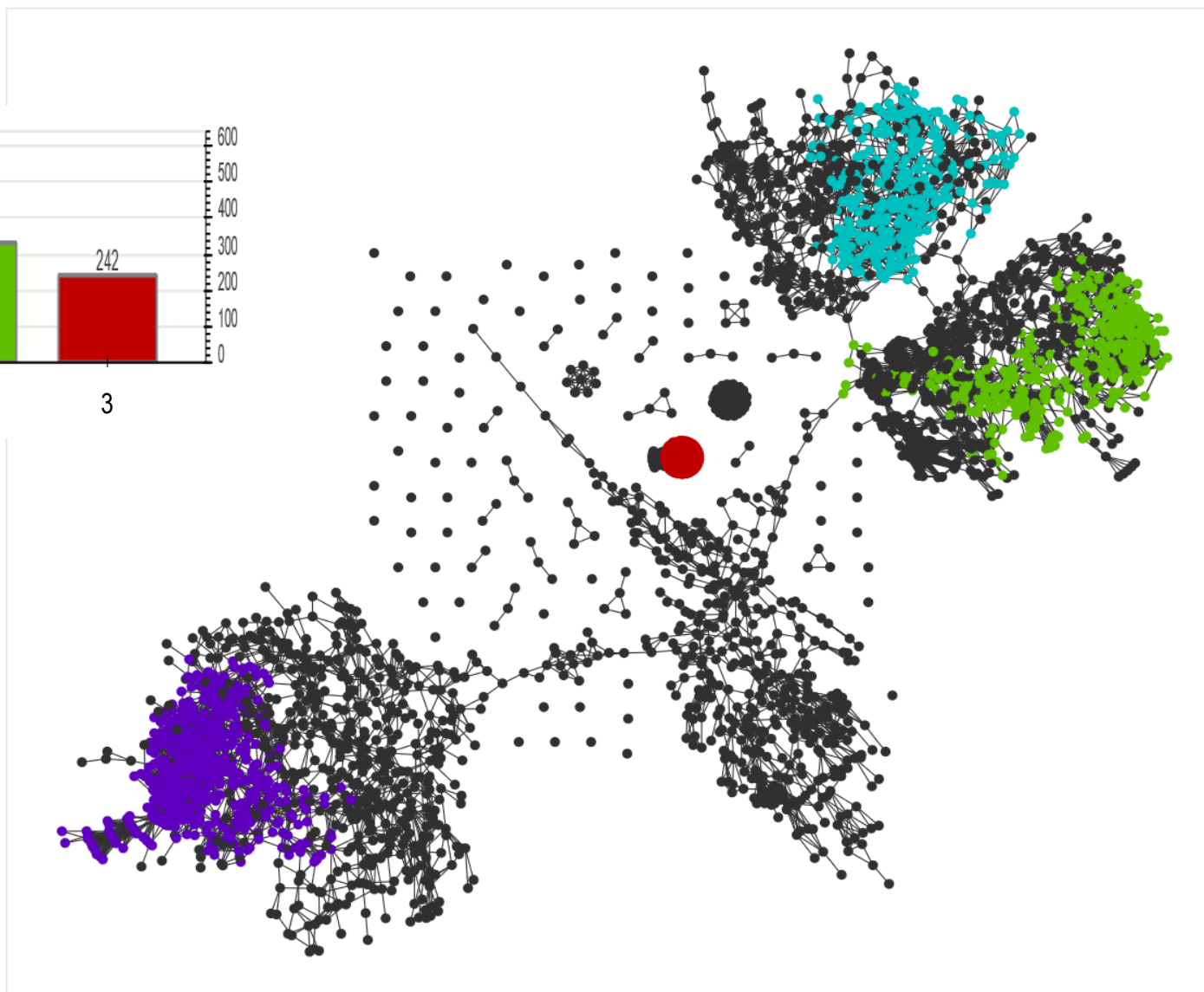
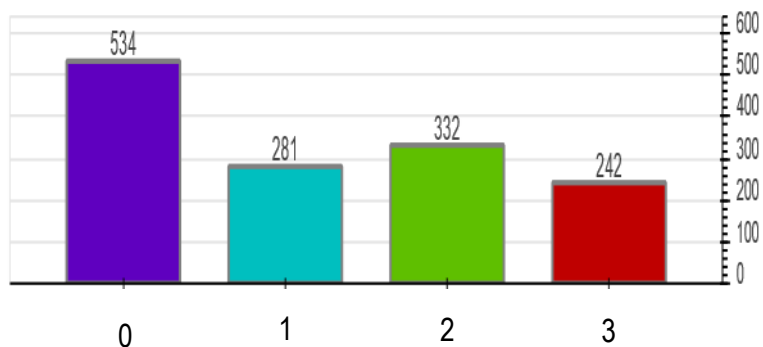
Automatic search for communities using machine learning algorithms

Community is a tightly-knit group of nodes with high density of within-group edges and a lower density of between-group edges

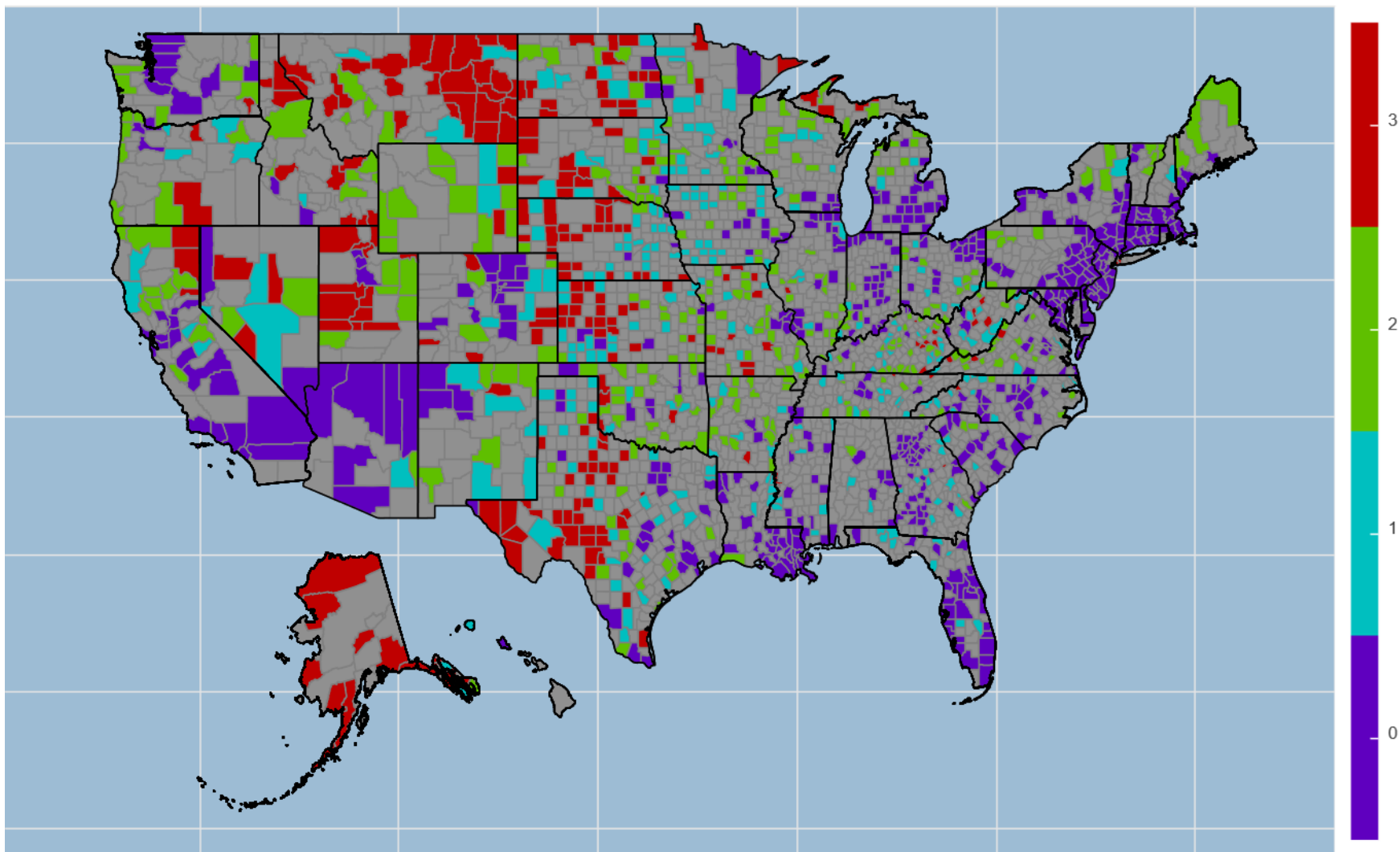
Nodes within communities are likely to **share common properties** and/or **play similar roles** in the graph



Automatic search for communities



Automatic search for communities on the map



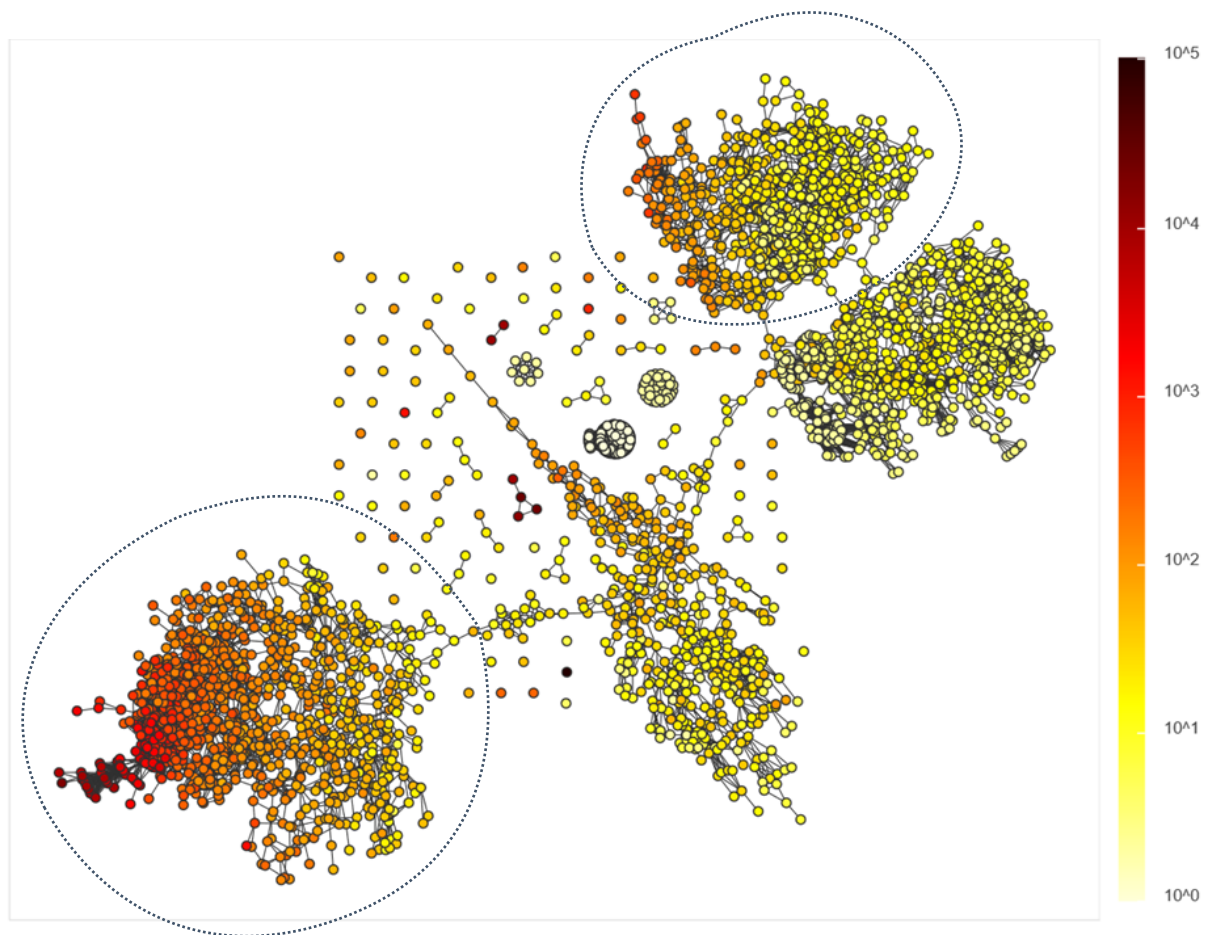


STEP 5.

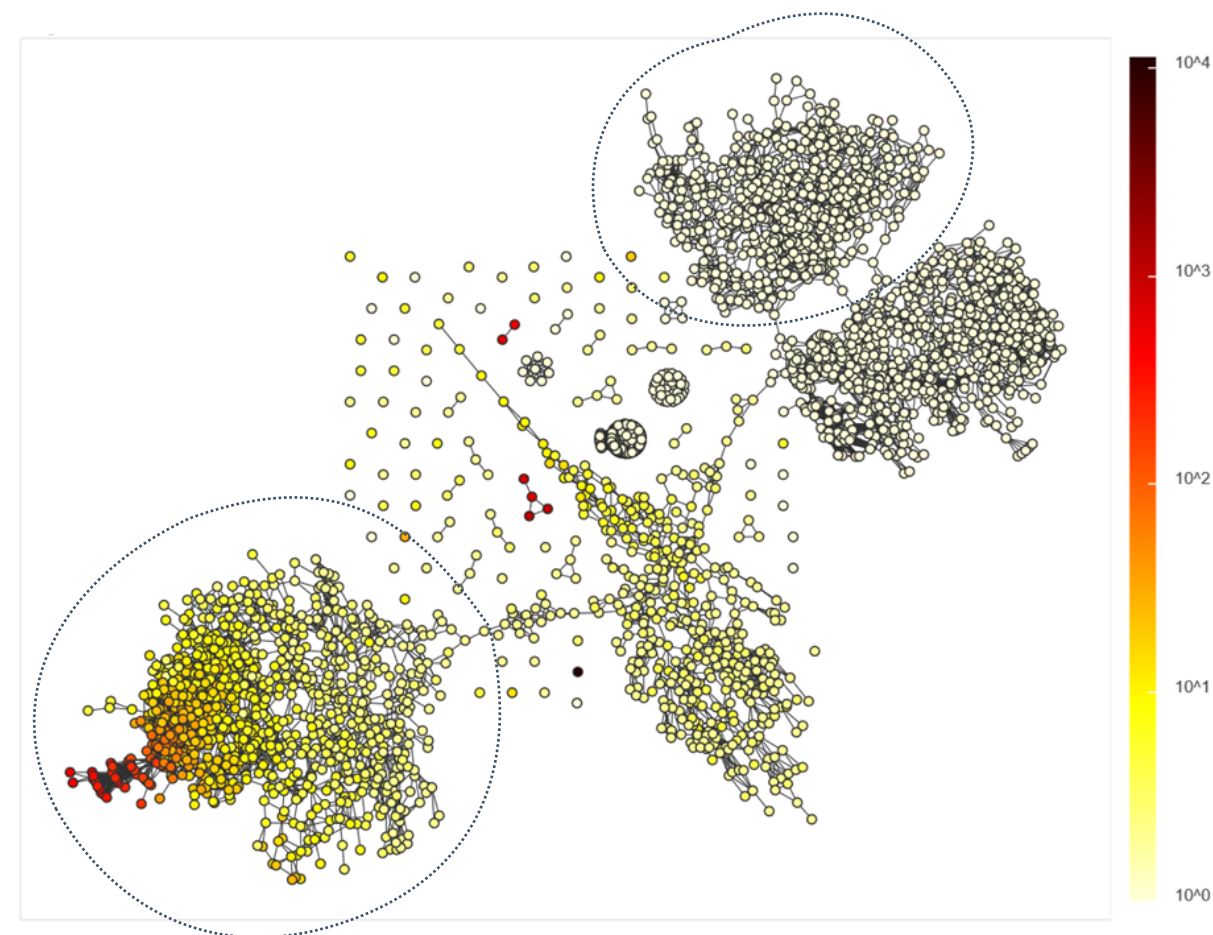
**STATISTICAL ANALYSIS
OF DISCOVERED SIMILARITIES**



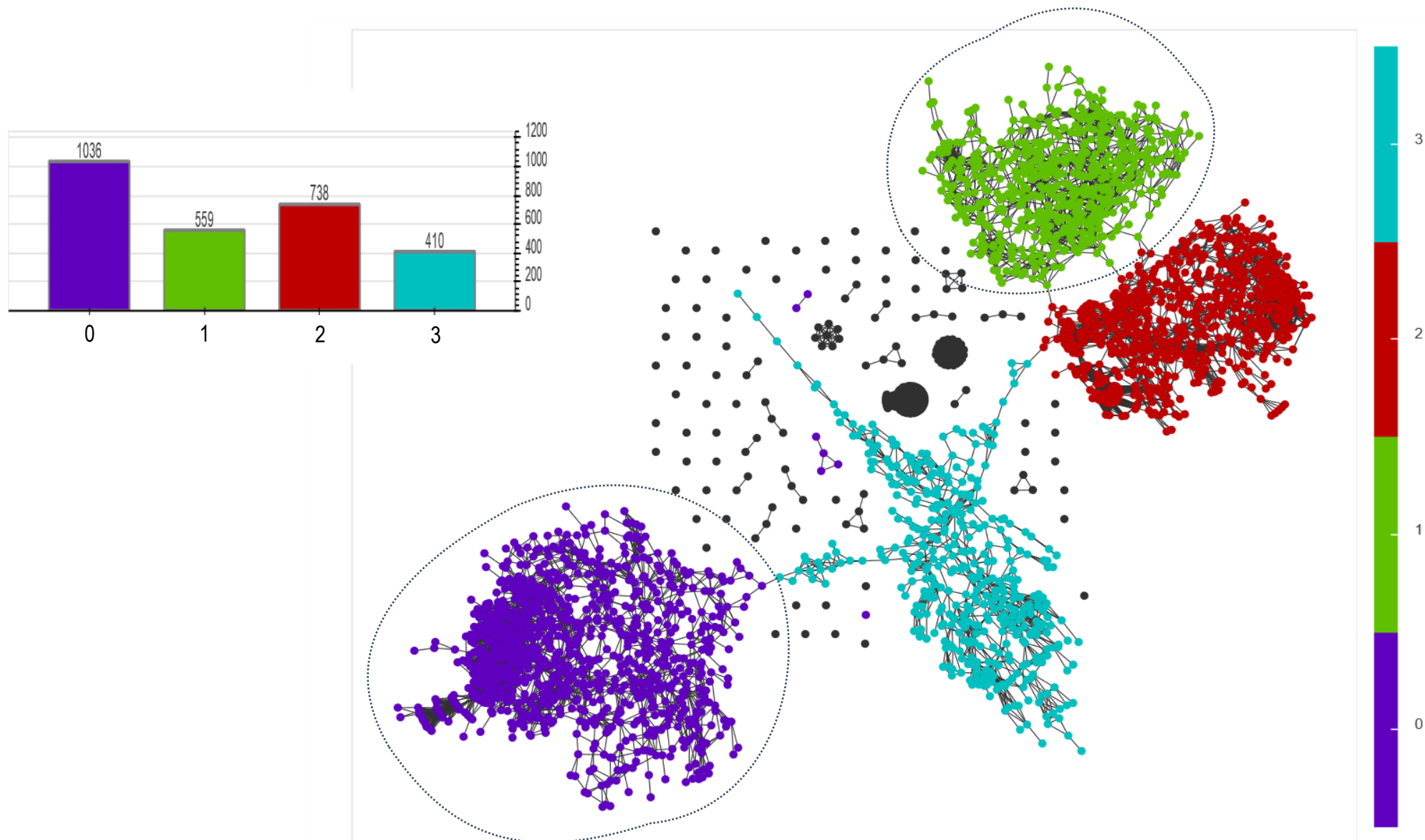
Number of confirmed cases



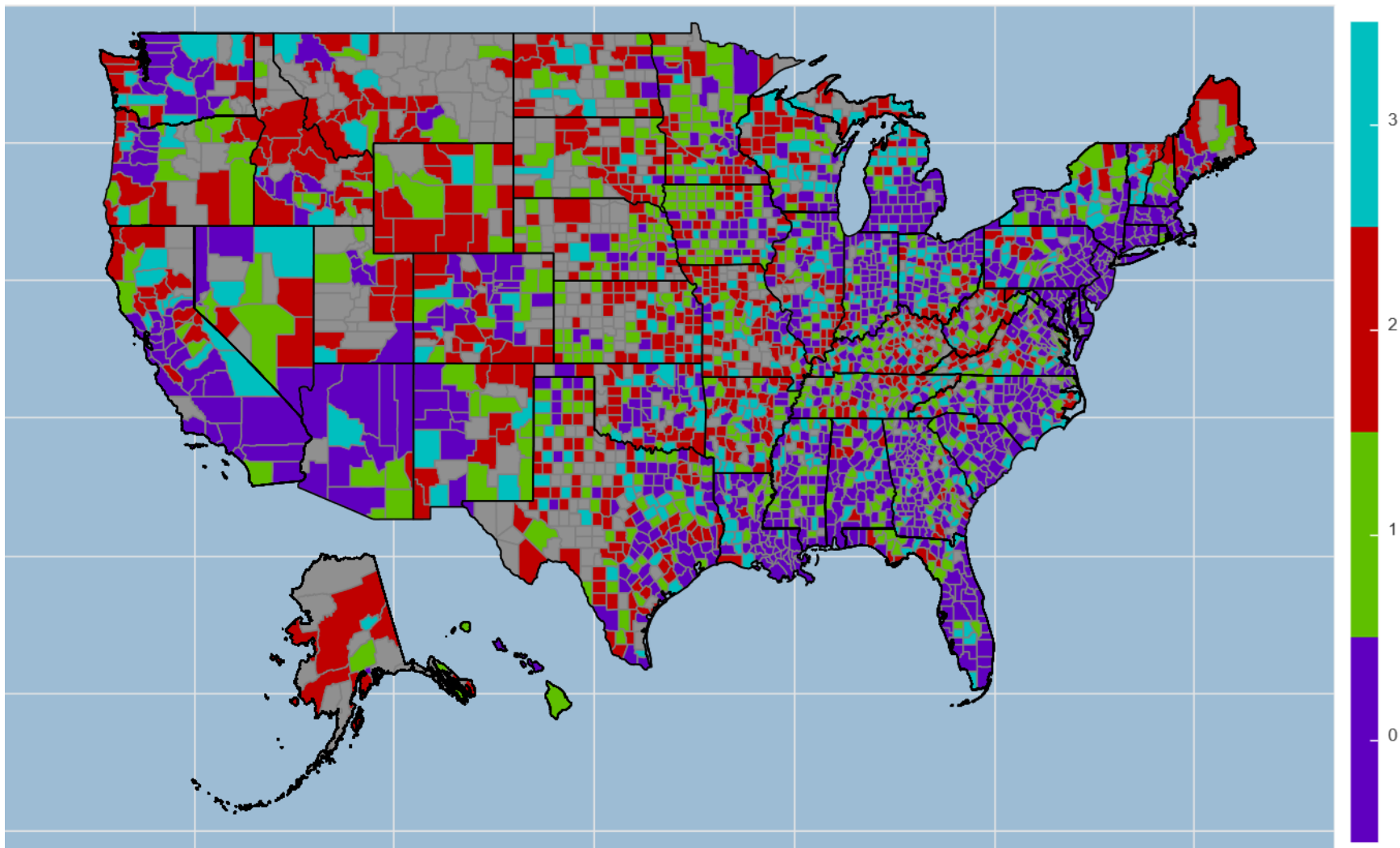
Number of deaths



Major communities

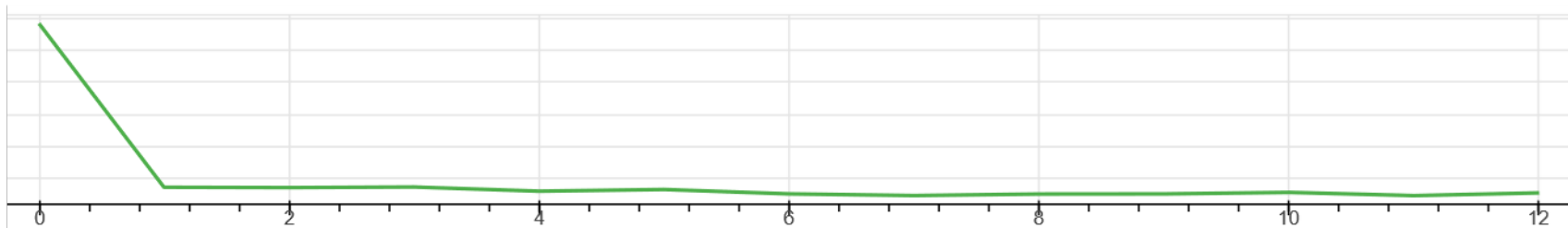


Major communities on the map

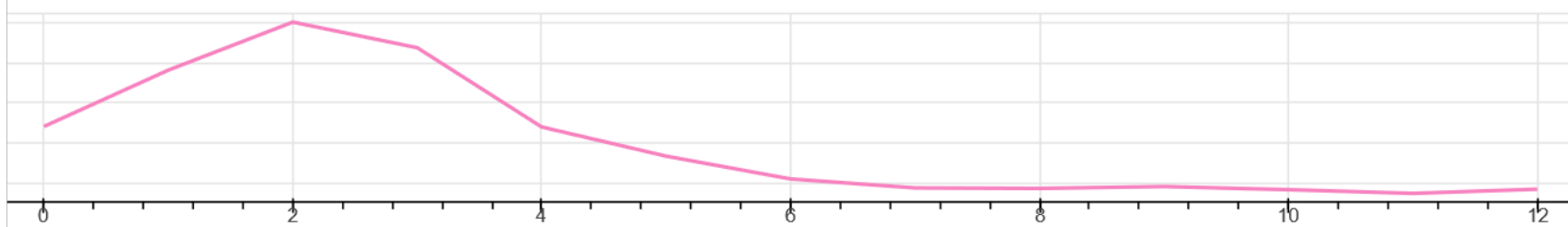


Epidemic curves of confirmed cases

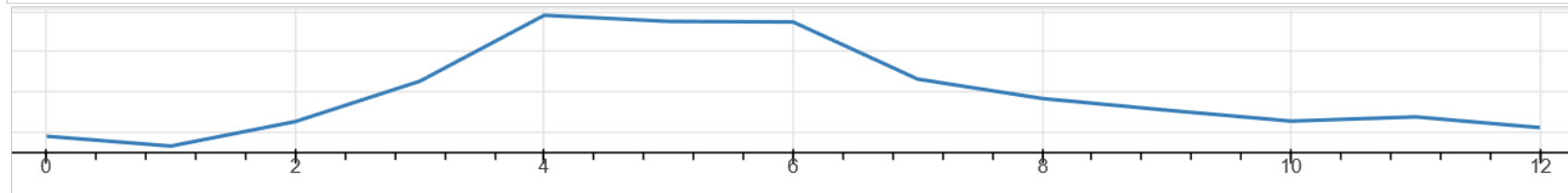
1



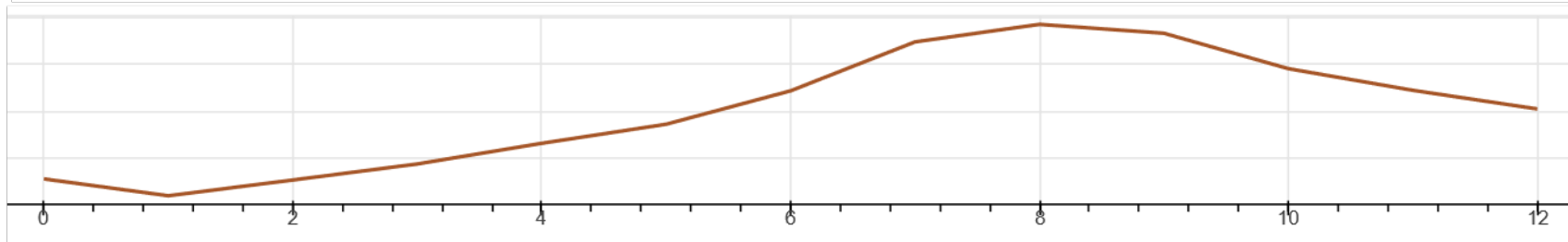
2



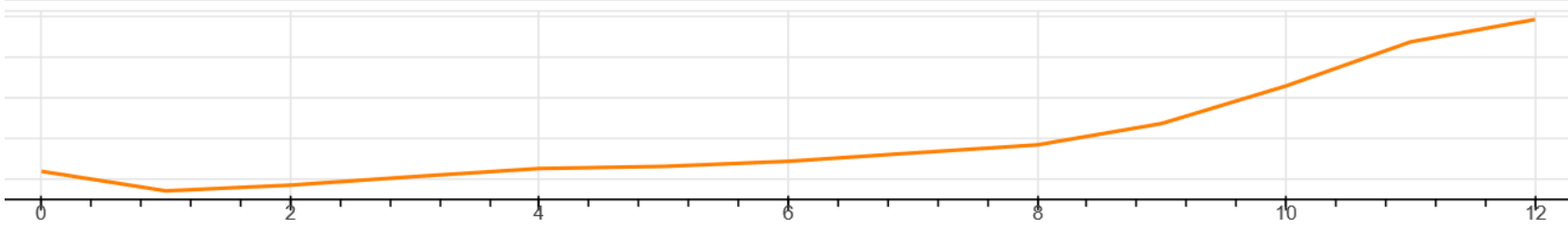
3



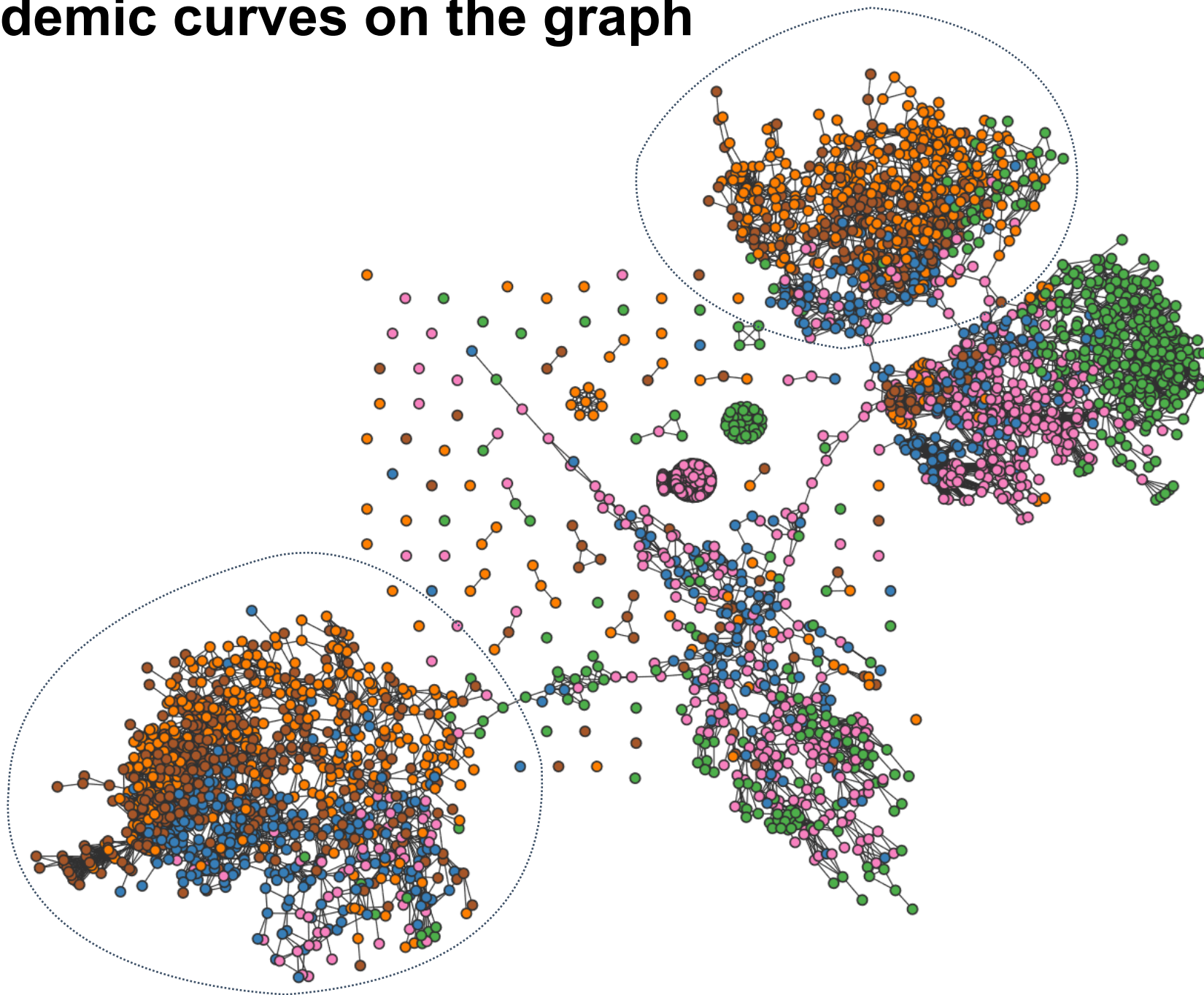
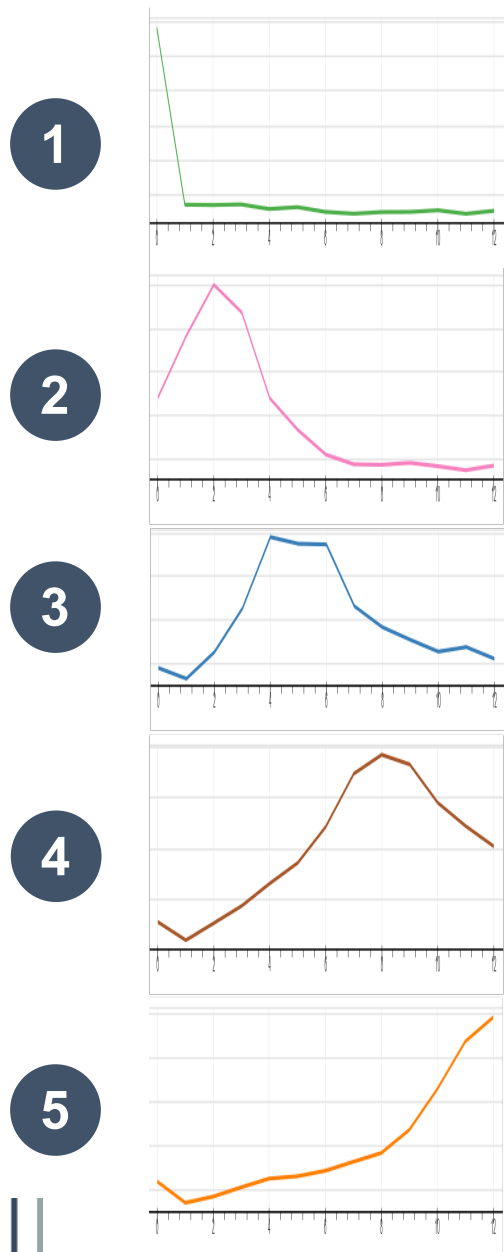
4



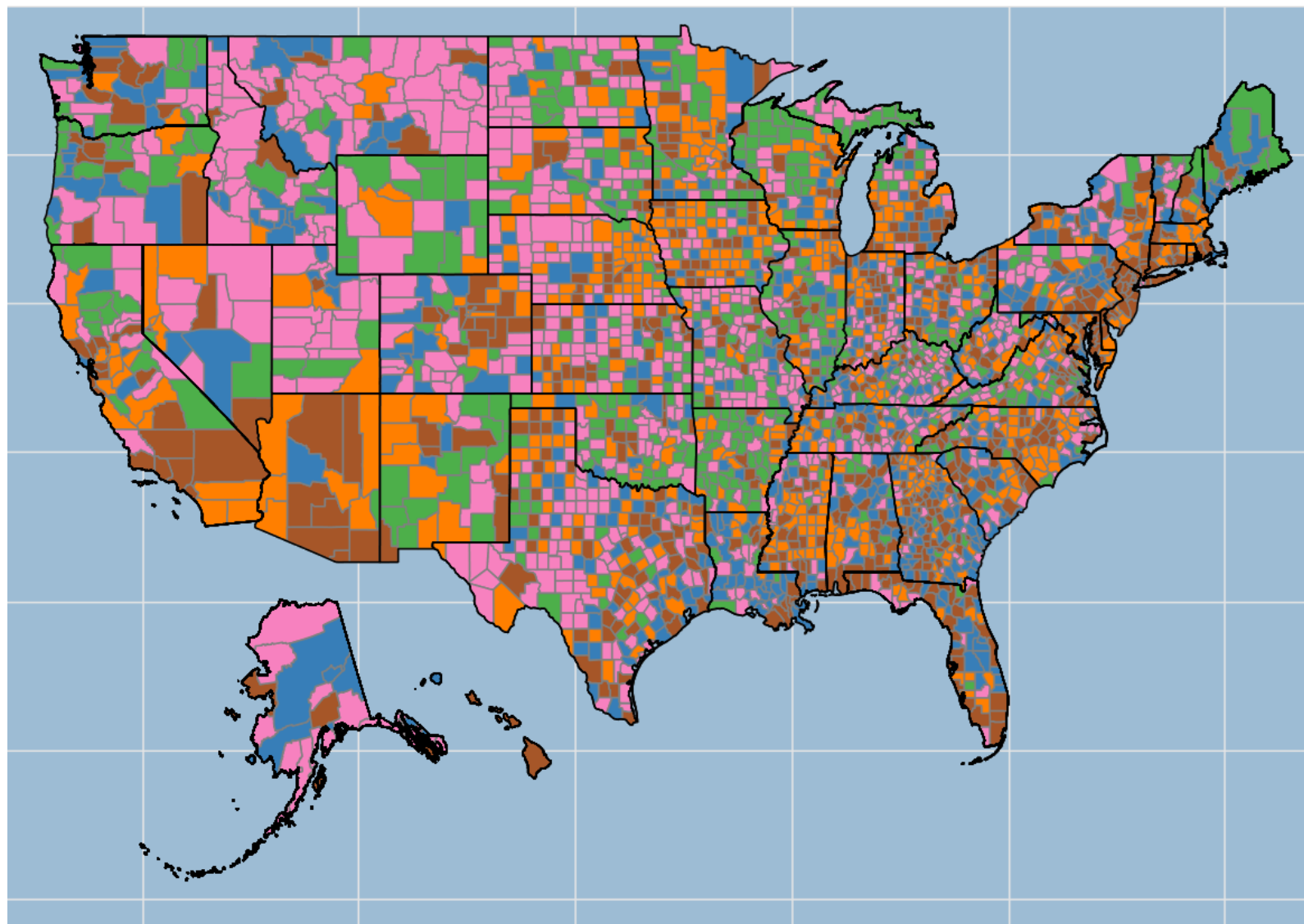
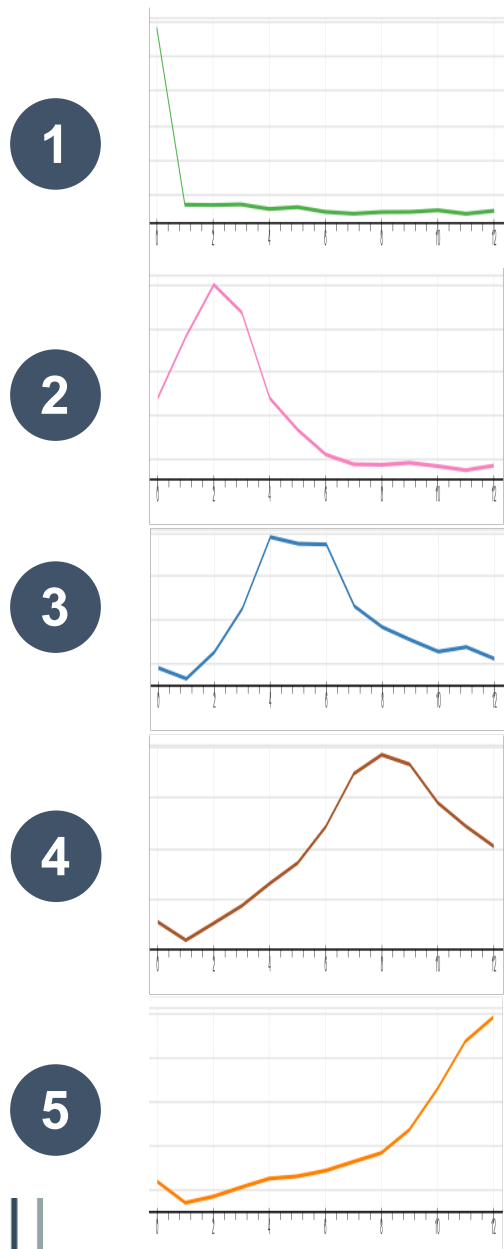
5



Epidemic curves on the graph

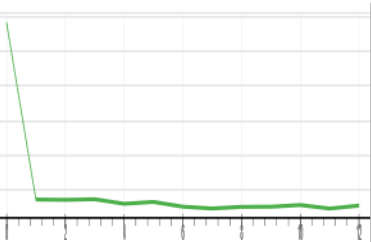
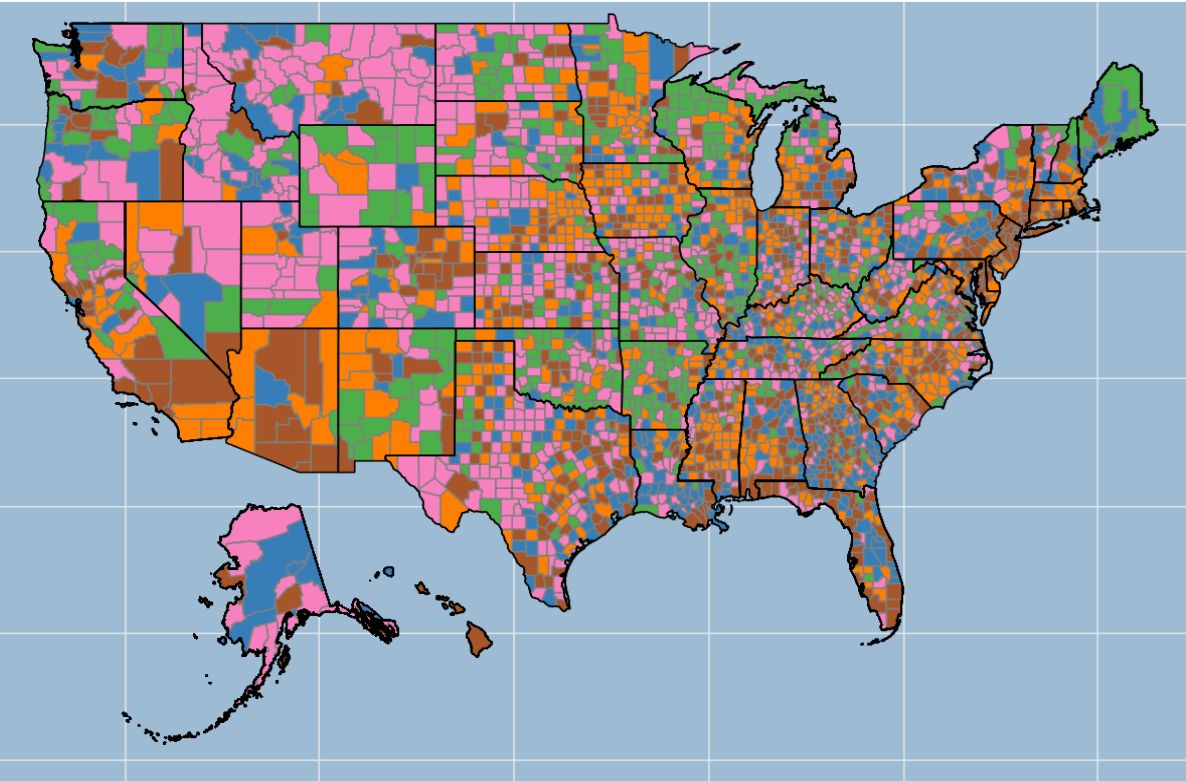


Epidemic curves on the map

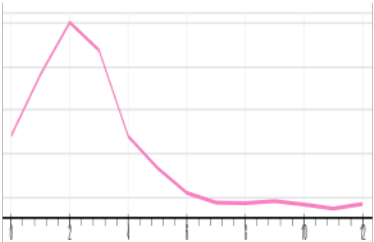




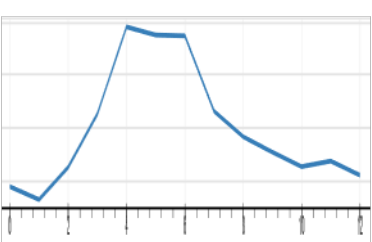
Epidemic curves on the geographical map



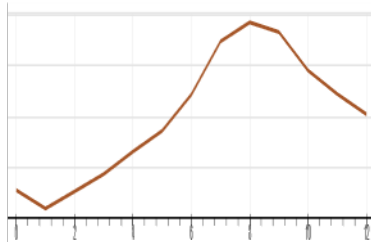
1



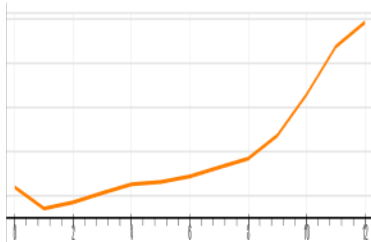
2



3



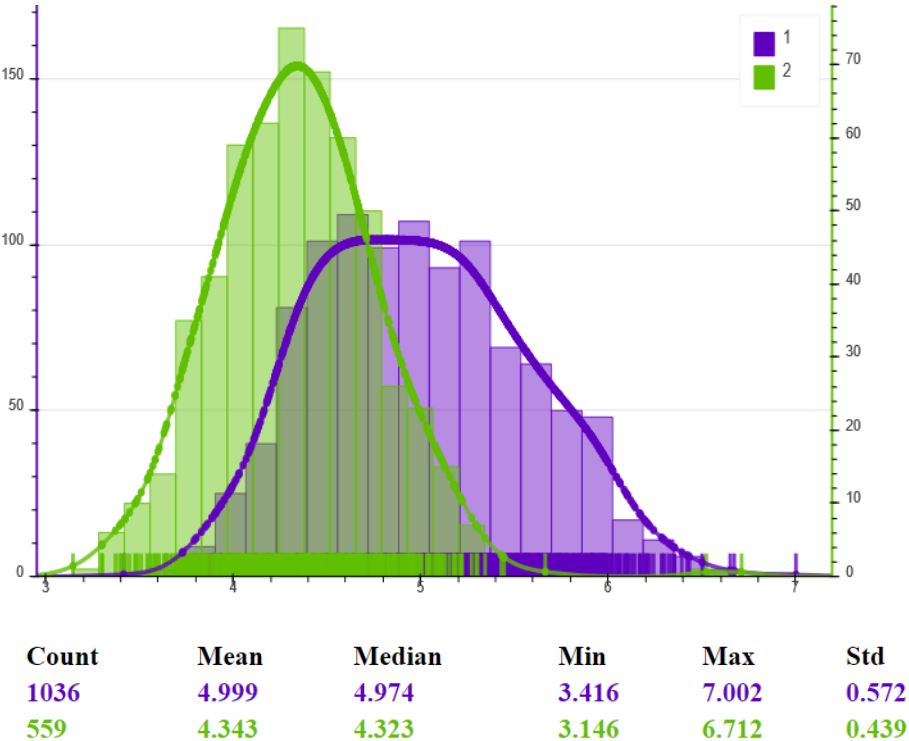
4



5

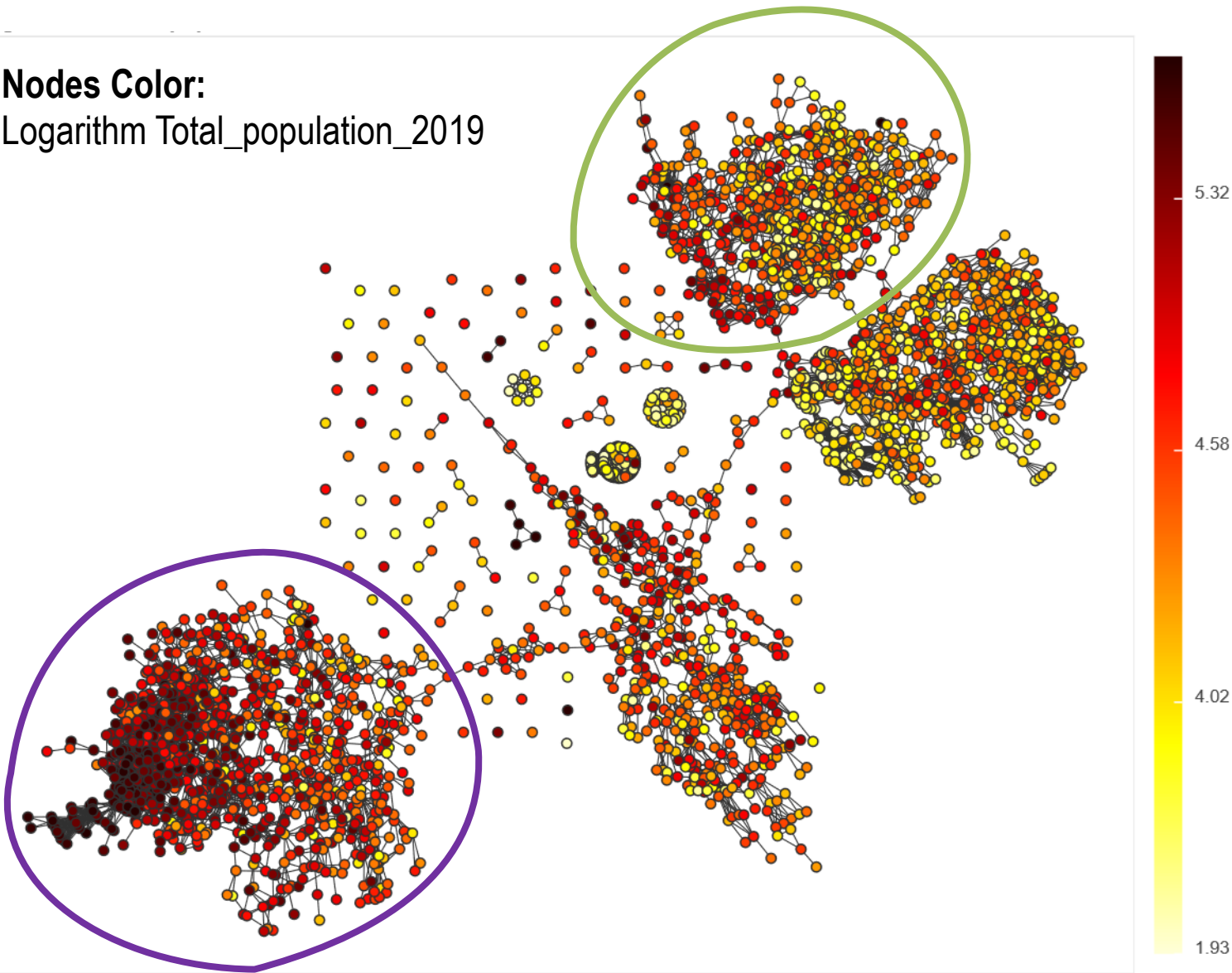


P-Value 1.4e-97



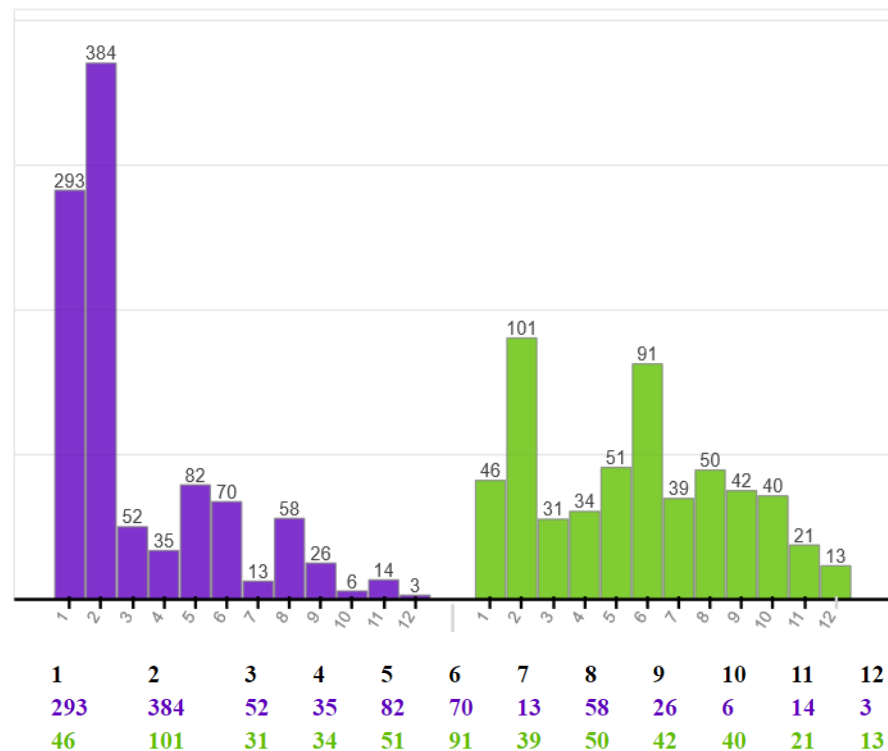
Population

Nodes Color:
Logarithm Total_population_2019

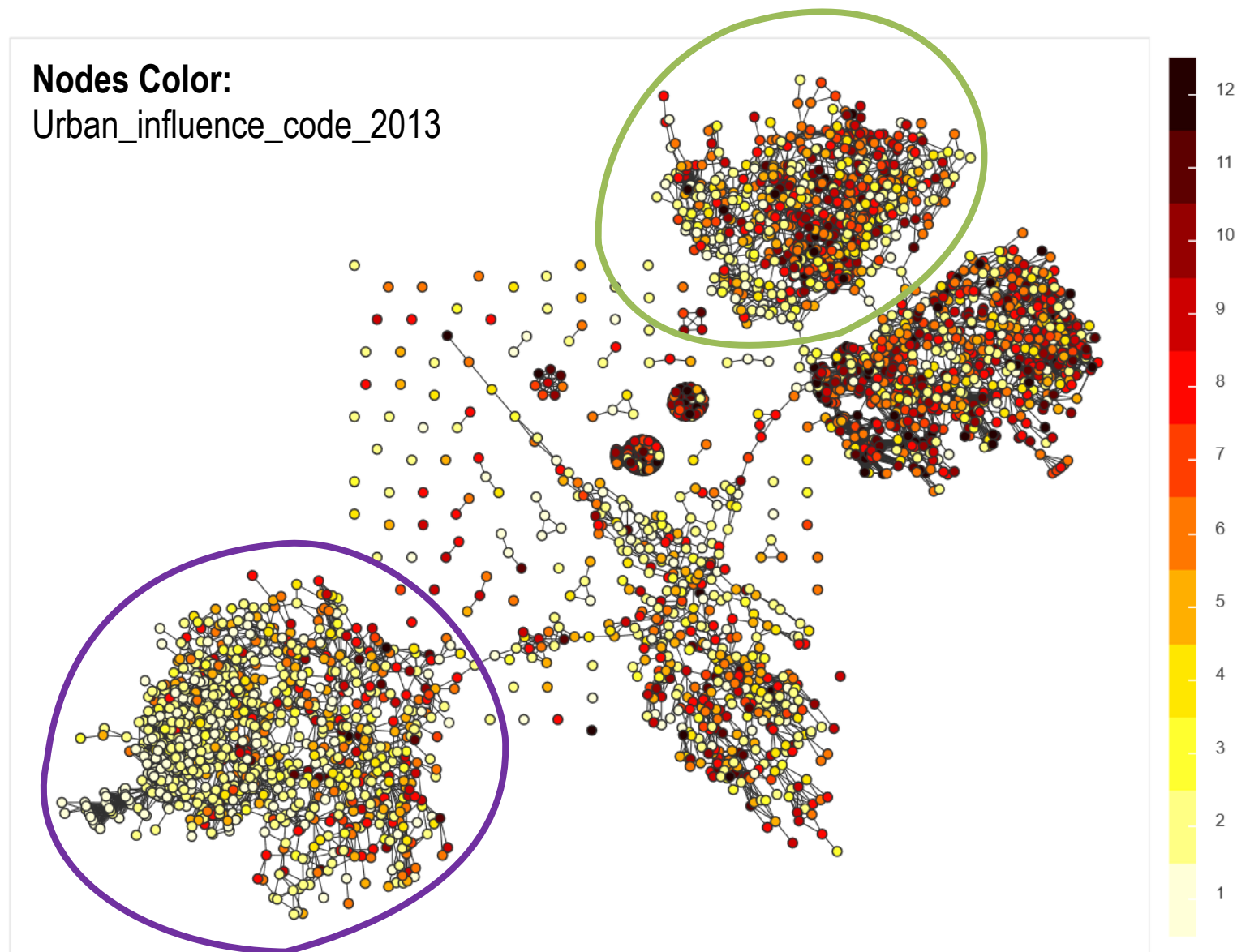


Urban Influence Code

P-Value 1.5e-56

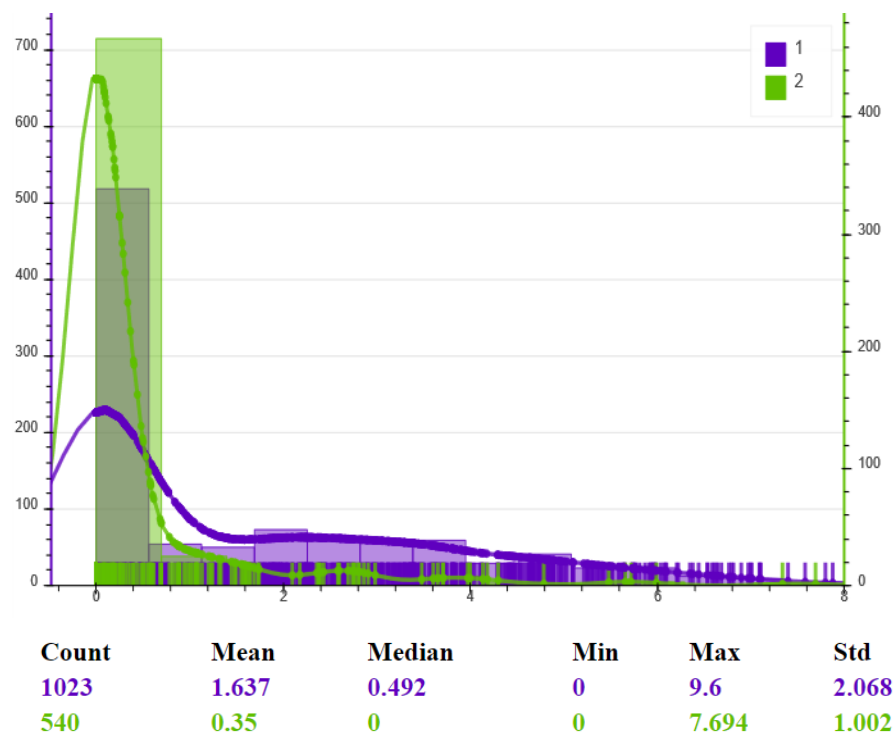


Nodes Color:
Urban_influence_code_2013

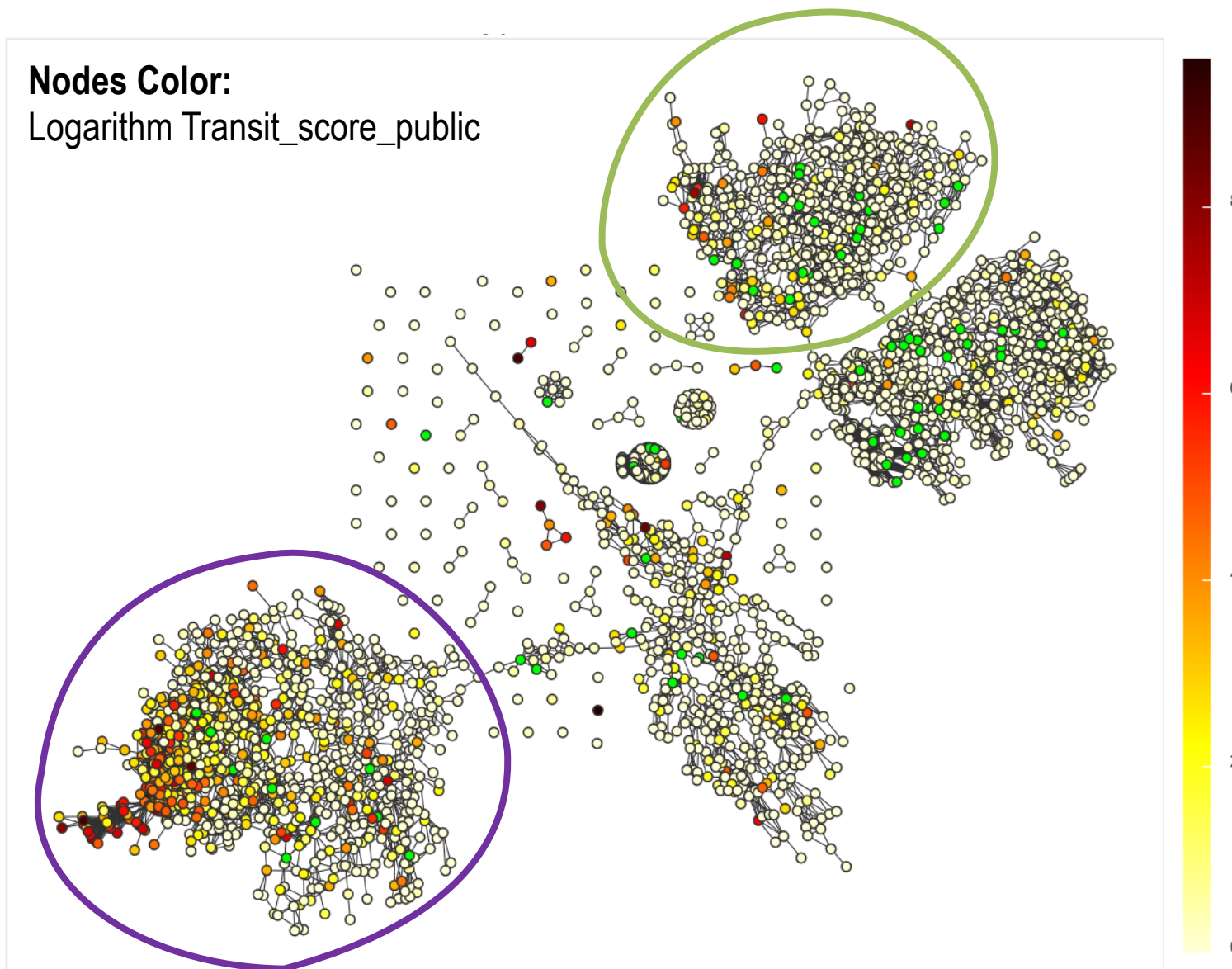


Public transportation

P-Value 3.2e-42

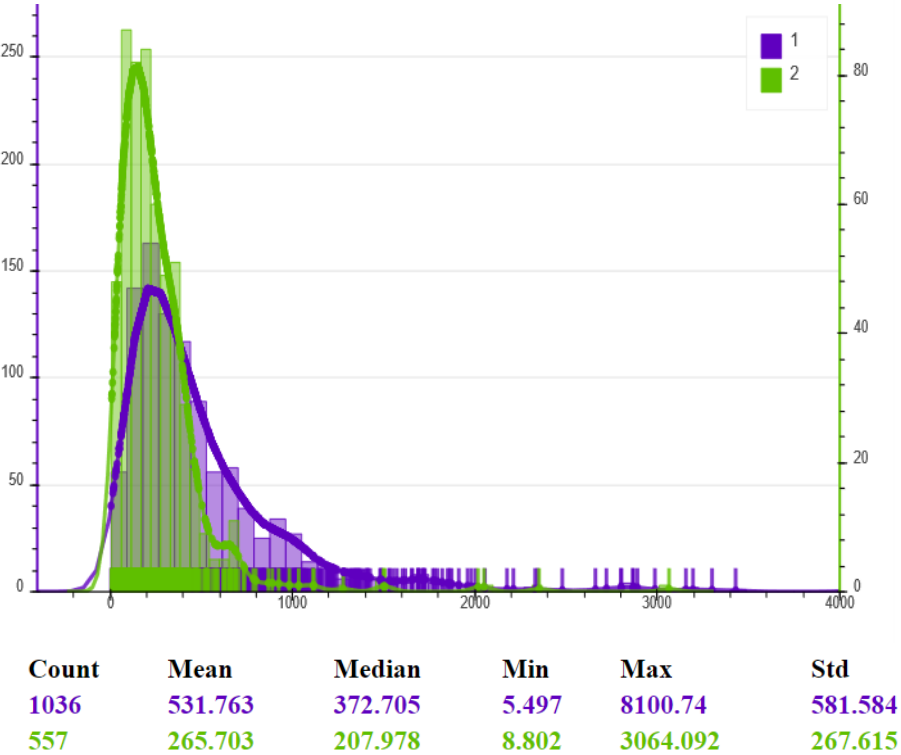


Nodes Color:
Logarithm Transit_score_public

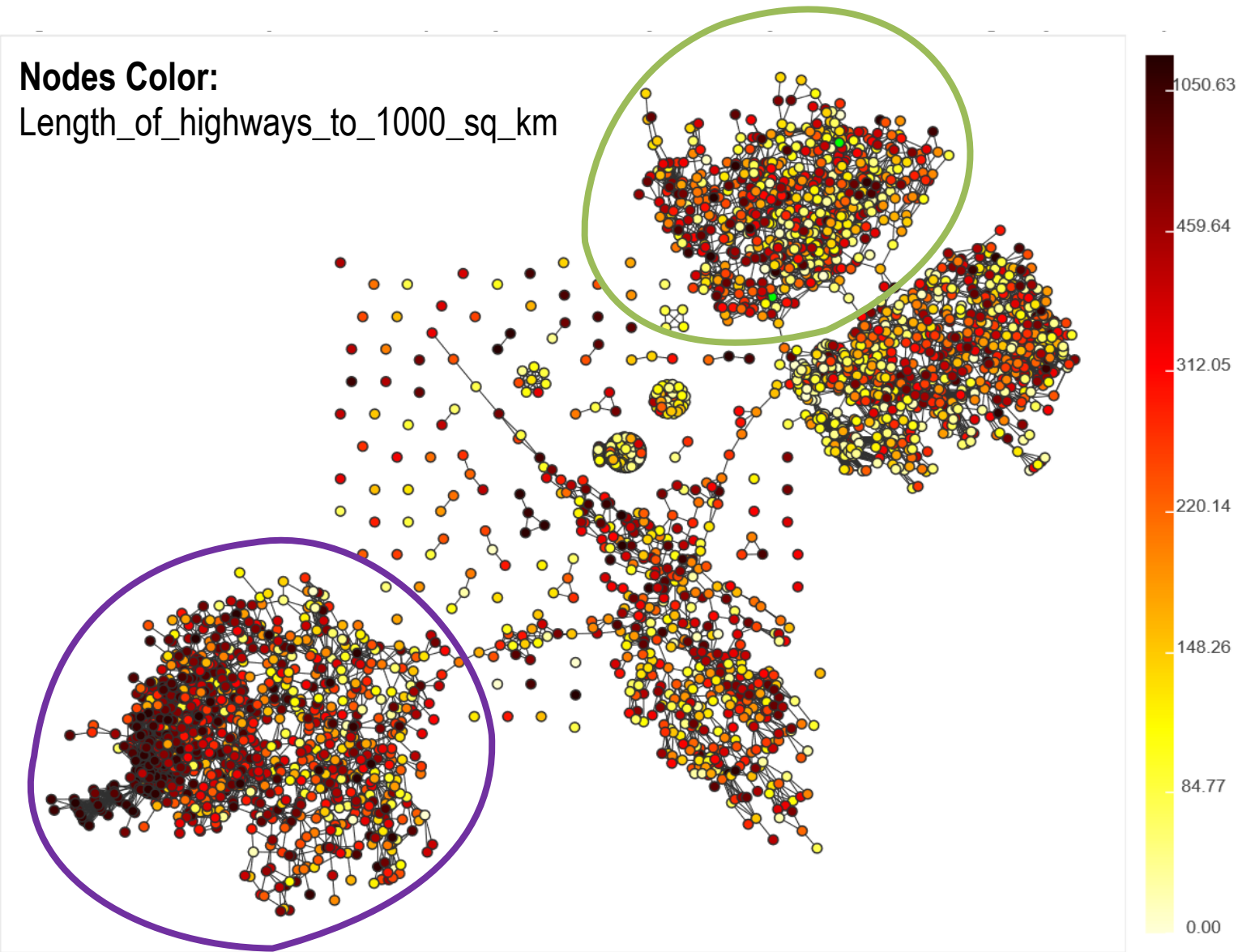


Length of highways

P-Value 9.8e-43

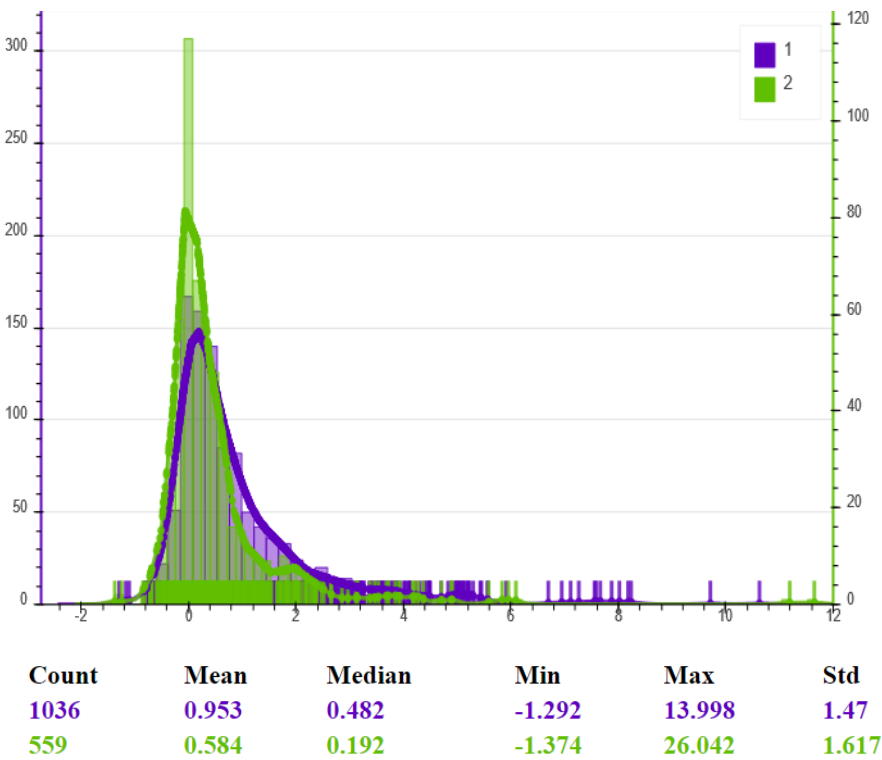


Nodes Color:
Length_of_highways_to_1000_sq_km

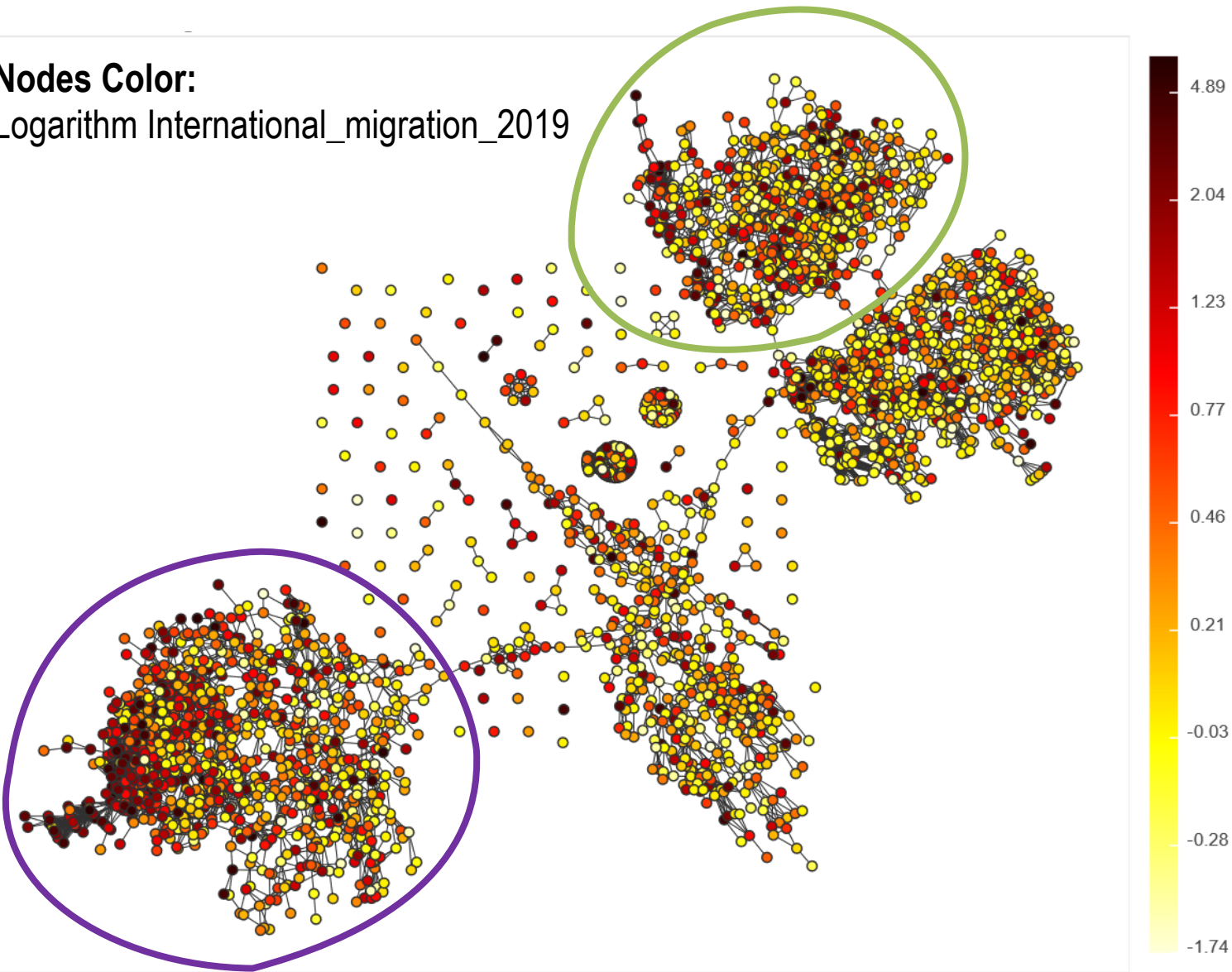


International migration rate

P-Value 1.3e-15

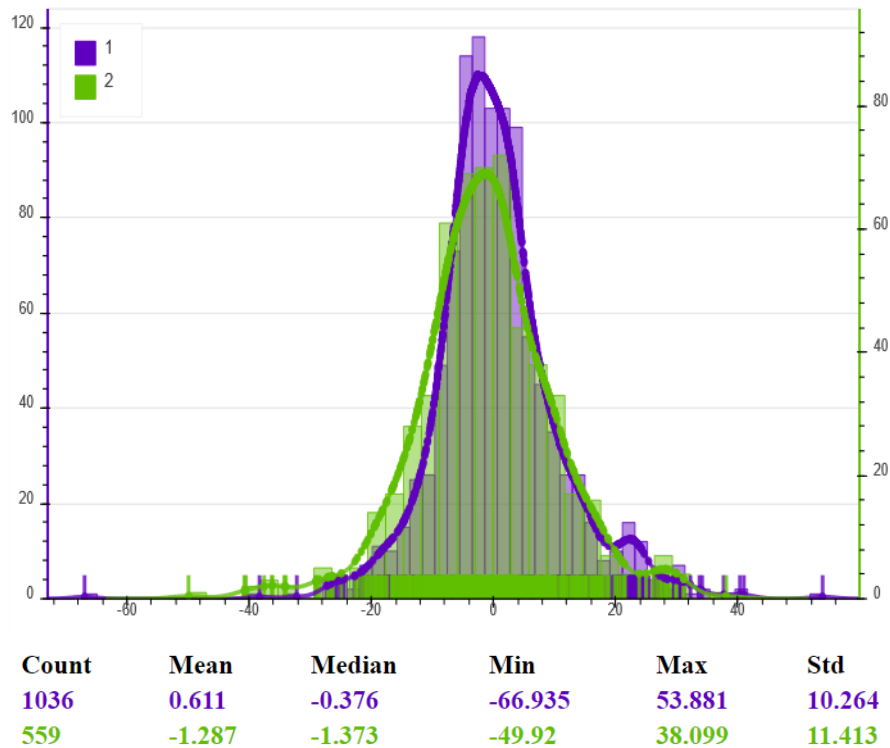


Nodes Color:
Logarithm International_migration_2019

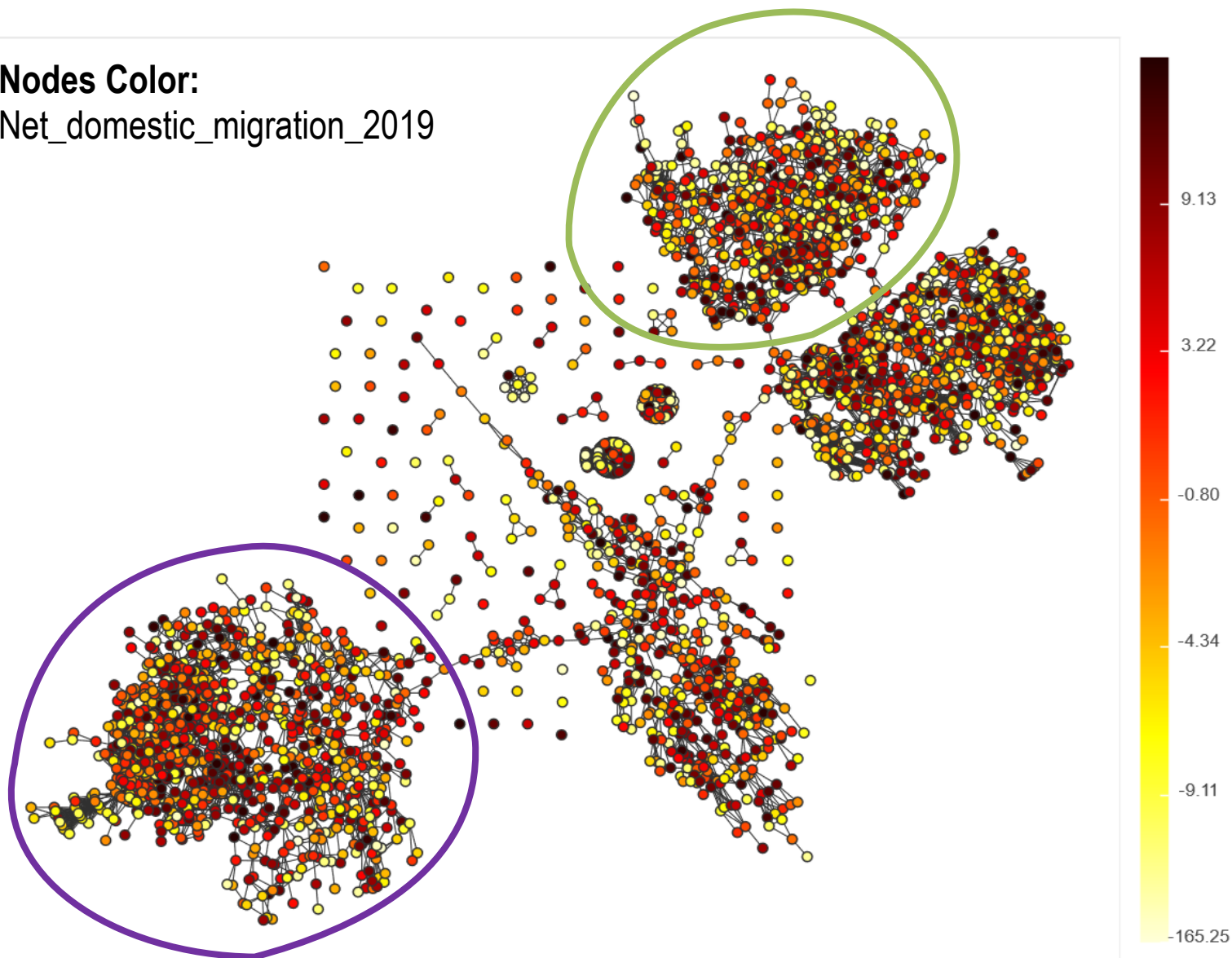


Domestic migration rate

P-Value=0.001

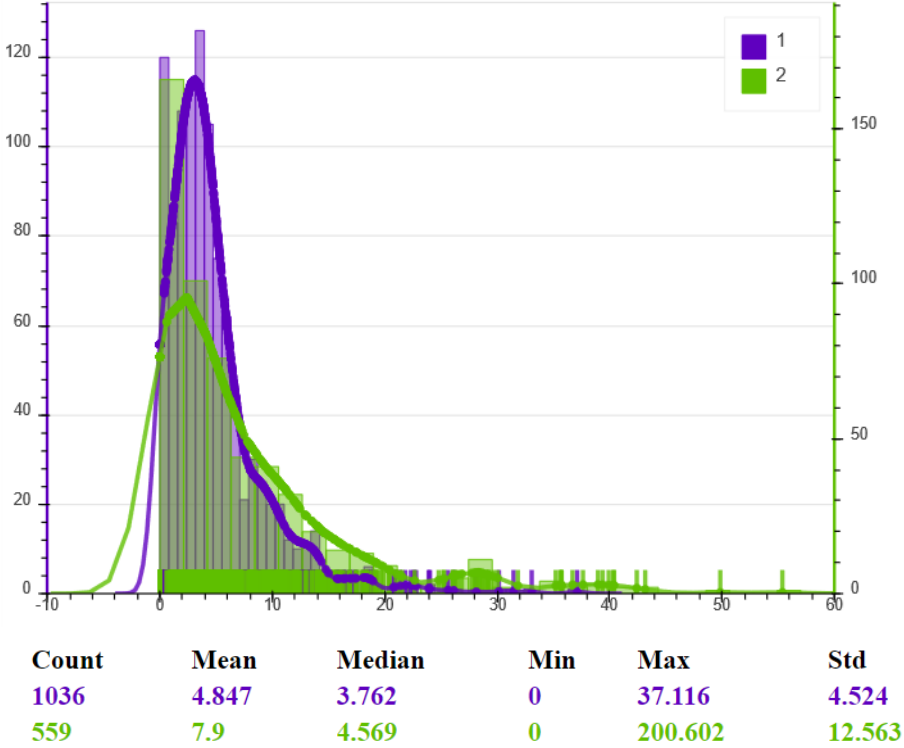


Nodes Color:
Net_domestic_migration_2019

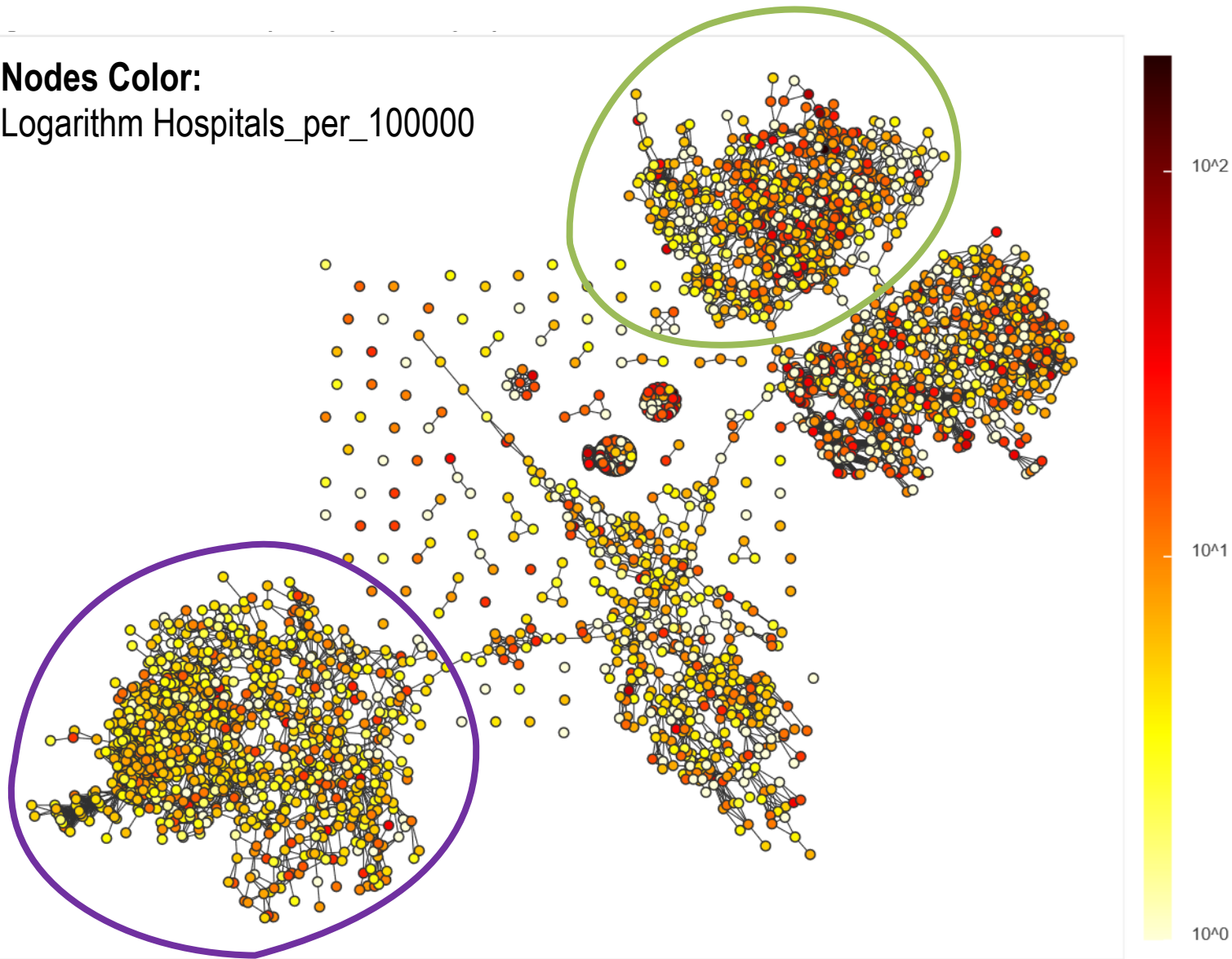


Number of hospitals

P-Value 0.002

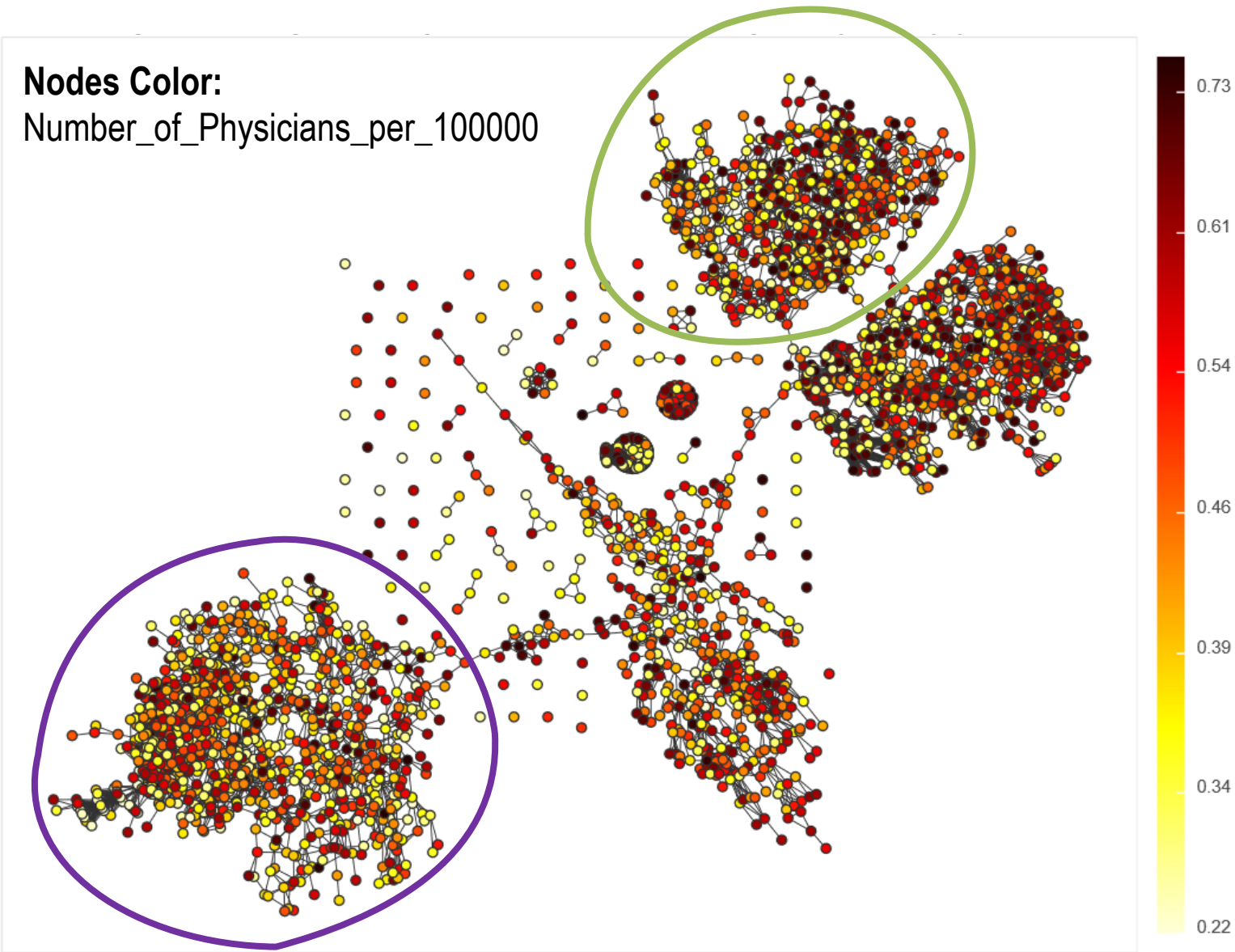
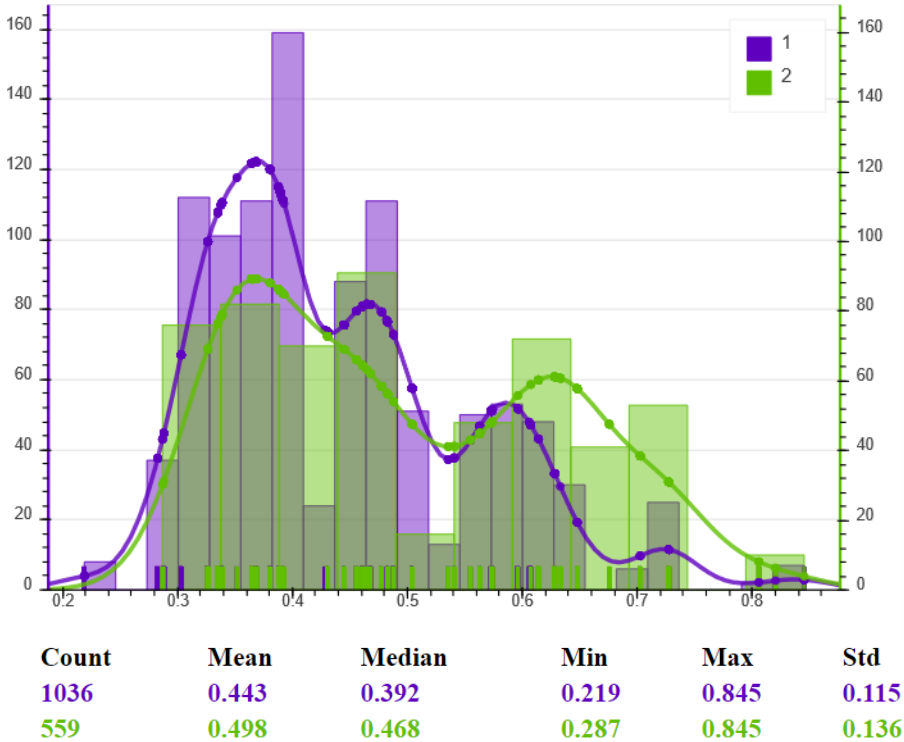


Nodes Color:
Logarithm Hospitals_per_100000



Number of General Practice Physicians

P-Value 1.1e-14



Presentation takeaways:

- We built the model using **Topological Data Analysis** based on:
 - 39-day time series
 - 28 outcomes
 - 250+ predictors (can be expanded after the model is built)
 - 3,142 counties
- Data mining experiment indicated a variety of predictors that could potentially influence the spread of COVID-19 over the US including **populations density**, **healthcare system**, **public transportation**, **migration**, and many other factors
- Please **download the paper** to get more information about the design of the experiment and more findings discovered



THANK YOU

Intego Group, LLC
555 Winderley Place, Ste. 129,
Maitland, FL 32751
Phone: +1 (407) 512-1006
www.intego-group.com/tcdm

