A Multiscale Geospatial Dataset and an Interactive Visualization Dashboard for Computational Epidemiology and Open Scientific Research

Muhammad Usman, *Department of Information and Computer Science, King Fahd University of Petroleum and Minerals, Dhahran, 31261, Saudi Arabia*

Honglu Zhou, Seonghyeon Moon, and Xun Zhang, *Department of Computer Science, Rutgers University, Piscataway, NJ, 08854, USA*

Petros Faloutsos, *Department of Electrical Engineering and Computer Science, York University, Toronto, ON, M3J 1P3, Canada*

Mubbasir Kapadia, *Department of Computer Science, Rutgers University, Piscataway, NJ, 08854, USA*

The coronavirus disease (COVID-19) continued to strike as a highly infectious and fast-spreading disease in 2020 and 2021. As the research community actively responded to this pandemic, we saw the release of many COVID-19-related datasets and visualization dashboards. However, existing resources are insufficient to support multiscale and multifaceted modeling or simulation, which is suggested to be important by the computational epidemiology literature. This work presents a curated multiscale geospatial dataset with an interactive visualization dashboard under the context of COVID-19. This open dataset will allow researchers to conduct numerous projects or analyses relating to COVID-19 or simply geospatial-related scientific studies. The interactive visualization platform enables users to visualize the spread of the disease at different scales (e.g., country level to individual neighborhoods), and allows users to interact with the policies enforced at these scales (e.g., the closure of borders and lockdowns) to observe their impacts on the epidemiology.

In December 2019, a novel respiratory infectious disease called coronavirus (COVID-19) was reported. As a severe infection with high risk that had not been identified in humans yet, we observed a rapid outbreak of this tragic epidemic.[1] Over the past two and a half years, COVID-19 has evolved into a global concern and affected all aspects of society, including daily commutes, college admissions, stocks, economics, people's work style, and even elections.

0272-1716 © 2022 IEEE
Digital Object Identifier 10.1109/MCG.2022.3230444
Date of publication 19 December 2022; date of current version 31 January 2023.

With a common goal to fight this pandemic, scientific researchers from various fields and organizations have been actively collaborating and making extensive efforts. An outstanding effort that cannot be neglected is the release of a number of COVID-19-related data repositories and visualization dashboards, such as the COVID-19 Multilanguage Tweets Dataset[2] and the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center,[3] respectively, which have enabled the offering of new research opportunities and embraced more well-thought-out insights. An open COVID-19 data working group was formed as a global and multiorganizational initiative that aims to enable rapid sharing of trusted and open public health

TABLE 1. Definition of the geolocation node at each scale in our dataset.

Scale	Description	Count
1	Country	206
2	First-order administrative division of a country (i.e., province, state, and equivalence)	1445
3	Administrative division which is smaller than scale 2 node but larger than a functional building (e.g., county, city, town, census area, borough, and parish)	7420
4	Functional building (i.e., airport and port)	2890

TABLE 2. Description of the geolocation node properties.

Index	Category	Name	Timestamped
1	Location geographic	Latitude and longitude	False
2	Population demographic	Population number	False
3	Movement pattern	Mobility trend time series of six space social functions.	True
4	Nonpharmaceutical intervention	Government "stay-at-home" intervention to address COVID-19.	True

data as well as visualization dashboards to advance the response to infectious diseases.[4]

Although the existing data resources have proven to be valuable, we found that they are insufficient to support multiscale and multifaceted modeling of computational disease understanding and simulation. In actuality, the spread of infectious diseases, human movement and behavior, and social and civil infrastructures are closely intertwined. In addition, the hidden mechanism that drives their co-evolution and the resulting phenomena are usually dramatically distinctive at different observation scales. Understanding their interplay from a multiscale perspective is critical for designing public policies and control measures. An easy interpretation for this would be the fact that policies are usually issued at multiple scales, such as national policy or school policy, which directly affects the populations at that corresponding scale and brings varying degrees of effects. For this reason, we believe it is problematic to ignore the differences of scales and model everything together. To this end, we offer an interactive visualization dashboard and a multiscale geospatial dataset by taking into account the geographical and social properties of locations and their connections, and the populations within.

Our dataset and dashboard are built under the context of COVID-19. Starting off from the multiscale geospatial understanding, we consider every geolocation as a node in our dataset. In total, four scales are specifically defined, with the country scale being the coarsest and individual buildings being the finest (see Table 1). Nodes have properties that characterize their location geography, population demographics, nonpharmaceutical government interventions related to COVID-19, and timestamped population movement patterns (see Table 2). Across the scales, the geolocations are connected by a social administrative division denoting the hierarchical containment relationship. Within each scale, the geolocation nodes are connected by two types of transportation, air and land, to enable the consideration of their different movement or spread capacity. Due to issues such as naming redundancy and data inconsistency across the scales, collecting and synchronizing such a dataset is not easy. To enable debugging and gain some nonsuperficial understanding and visual demonstration of the multiscale dataset, we also built an interactive visualization dashboard (see Figure 1). We have developed a website[a] to host the COVID-19 multiscale dataset, descriptions, and visualization dashboard.

In terms of novelty and contributions, we have developed a multiscale geospatial dataset that interlinks location geography, population demographics, nonpharmaceutical government interventions related to COVID-19, and timestamped population movement patterns, all at different scales (e.g., country, state/province, county/city, and individual buildings). Here, in particular, we would like to emphasize the nonnegligible manual efforts to clean the data and the challenges that we have encountered, such as 1) unifying and aligning data from multiple sources, and 2) formulating the

[a][Online]. Available: https://usmannn.github.io/COVID-19-Multiscale-Simulation/

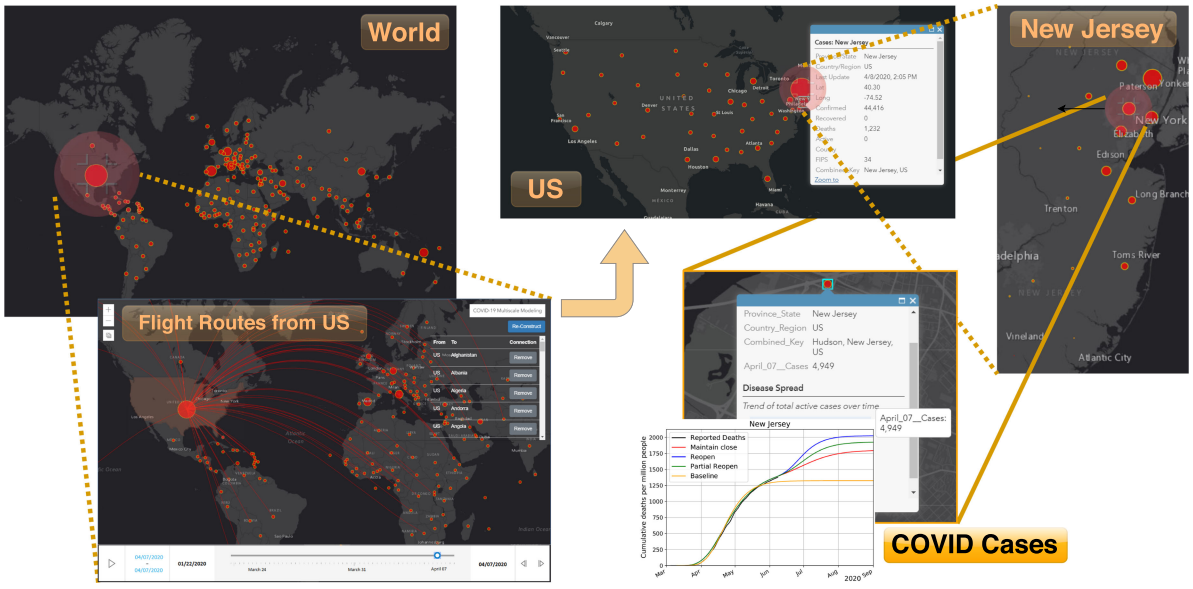

FIGURE 1. Preliminary prototype of multiscale interactive platform with visualization of infection spread in New Jersey. The different components in the figure reflect user interactivity with the system. The red circles represent different nodes, whereas the lines represent the connections between the nodes in the network. The size of the disks relates to disease exposure in these nodes.

novel heterogeneous directed multiscale multigraph, with both static and dynamic properties, that is, information clean and consistent (e.g., issues, such as ambiguity and redundancy, are resolved at the best manual efforts). We have also developed an interactive visualization dashboard. Our platform enables users to visualize the spread of the disease at different scales (e.g., from the country level to individual neighborhoods). Furthermore, it allows users to interact with the policies enforced at different scales in the network (e.g., the closure of borders and lockdowns within communities) to observe their impact on epidemiology.

We, in addition, offer correlation-based analysis to examine our dataset. Some of the salient insights from the data analysis include the following.

- The correlation patterns differ distinctively between the scales (e.g., scale 3 can be way different from scales 1 and 2—see scale descriptions in Table 1), indicating the necessity of exploiting unique modeling strategies for each scale and the possibility of achieving performance gains by encouraging information exchange across scales.
- A nonlinear model might be able to better capture the patterns hidden in our dataset. There exist high nonlinear relationships between population, flight-related features, and the arrival time of COVID-19.
- The later the stay-at-home implementation date is, the higher the infected ratio can be observed at the U.S. state level.
- The COVID-19-related and mobility trend time series share high mutual information, indicating nonindependent relationships.

The rest of this article is organized as follows. We review related work and give an overview of computational epidemiology and visualization dashboards in the "Related Work" section. Dataset descriptions are in the "Multiscale Geospatial Dataset" section. We provide a set of data analysis, such as multiscale feature correlation comparison, in the "Data Analysis" section. The interactive web-based data service (visualization dashboard) is described in the "Interactive Data Visualization Service" section. Implementation language details are presented in the "Implementation Details" section. We highlight the applicability of our dataset for future researchers in the "Data Applicability" section. Finally, the "Conclusion" section concludes this article.

RELATED WORK

In this section, we present a brief background on computational epidemiology and multiscale models of infectious disease, followed by the literature review on the

COVID-19 Visualization Systems (see the "COVID-19 Visualization Systems" section), and on the understanding of the effects of nonpharmaceutical interventions to limit the spread of COVID-19 (see the "Understanding the Effect of Nonpharmaceutical Interventions" section).

Computational epidemiology is a multidisciplinary field that aims to better understand issues central to epidemiology, such as the spread of diseases or the effectiveness of a public health intervention. Prior techniques include rate-based differential equation mathematical models, such as the popular susceptible-infected-recovered (SIR) model[5] and its variants,[6] agent-based modeling,[7] and others.[8] In the area of multiscale models of infectious disease, a large number of published papers focused on modeling the dynamics of infectious diseases and do not address the multispatial scales of these disease systems.[9] Wang et al.[10] considered multiresolution spatial-temporal forecasting. However, the authors only consider the county and the state scale with just two states in the United States.

COVID-19 Visualization Systems

Several efforts have been made to visualize the outbreak of the COVID-19 pandemic. An interactive interface is presented to track COVID-19 cases around the world in real time.[3] Yang et al.[11] proposed a citywide visual analytics system to simulate human mobility and observe the COVID-19 infection status in response. In another system, Samant et al.[12] presented a multilayer network as an interactive dashboard to visualize various aspects of COVID-19 in the United States. Additional literature on visualization platforms and what the proposed interactive visualization solution offers differently is presented in detail in the "Interactive Data Visualization Service" section.

Understanding the Effect of Nonpharmaceutical Interventions

As special control measures are made from the global level, country level, state level, to city level, researchers are attempting to understand their effect on the spread of COVID-19. For example, Chinazzi et al.[13] used a global metapopulation disease transmission model to project the impact of travel limitations on the national and international spread of COVID-19. Their modeling results indicate that sustained 90% travel restrictions to and from Mainland China only modestly affect the epidemic trajectory unless combined with a 50% or higher reduction of transmission in the community. Adiga et al.[14] evaluated the impact of international airline suspensions on the early global spread of COVID-19. Wilder-Smith and Freedman[15] analyzed the pivotal role of interventions, such as isolation, quarantine, social distancing, and community containment, for old-style public health measures in the COVID-19 outbreak. Straightforward and well-thought-out insights are required to be drawn in a more timely manner for policymakers. Moreover, it is not clear how would every unique control measure, taken at different scales ranging from individual to global, influences infectious disease.

MULTISCALE GEOSPATIAL DATASET

In this section, we describe the proposed multiscale geospatial dataset. We first present the overview of the dataset in the "COVID-19 Visualization Systems" section. The data collection process and detailed data description are described in the "Understanding the Effect of Nonpharmaceutical Interventions" section.

Data Overview

Starting off from the multiscale geospatial understanding, we consider every geolocation *in different scales* as a node in the proposed dataset. Overall, four scales are specifically defined, from the coarsest country scale to the finest functional building scale. This is because the unique geographical and social implications of locations, either within or across scales, can lead to drastically different outcome or impact of an infectious disease. Table 1 provides the description and statistics of the geolocations/nodes at each scale. In total, the dataset contains information on 11,961 geolocations/nodes.

To summarize the information provided for the nodes: all of the nodes have associated data describing properties that characterize their location geography. Properties, such as population demographics, are provided for a portion of the nodes. To support case study, nonpharmaceutical government interventions related to COVID-19 and movement patterns during that period of time for the nodes are also provided as *time-varying* node properties. See Table 2 for an enumeration of the types and names of node properties provided in the dataset.

Intuitively, the geolocation nodes are connected by the administrative division *across the scales*, denoting the hierarchical containment relationship. For example, the scale 1 "US" node contains (is connected by) 56 scale 2 state-level (or equivalent) nodes. *Within each scale*, the geolocation nodes are connected by different types of transportation, specifically, by air, land, and sea. See Table 3 for node connections that connect nodes at each given scale.

TABLE 3. Description of the geolocation node *connection within a given scale*.

Index	Category	Name	Attributed
5	Air	International flight	True
6	Air	Domestic flight	True
7	Land	Sharing border	False

As a result, our dataset formulates a multiscale multigraph[b] with 1) heterogeneous type of nodes and bidirected edges, and 2) both static and dynamic attributes of nodes and edges.

The major data preprocessing stages are as follows:

1) collecting the available information sources;
2) identifying all geolocations of all scales from all sources;
3) forming the graph nodes by removing geolocation duplicates and resolving issues, such as ambiguity and naming redundancy;
4) adding node connections (flight edges and border edges);
5) populating node and edge attributes.

We purposely kept the focus of our data collection to 2020, which is known to be the high time of the COVID-19 pandemic, with a diverse set of demographic and travel policy interventions in different countries. This does not make our multimodel solution (as in our suggestion discussed in the "Conclusion" section driven by our data analysis) or the visualization platform unusable for a different time or a year. The collected dataset, including visualization, could be found in our web-based dashboard, and the dashboard will be described in the "Interactive Data Visualization Service" section.

Data Collection and Description

In this section, we describe the data sources, our method for data processing, as well as any assumption made.[c]

Geolocation Identity and Connection by Scale

Types of geolocations are demonstrated in Table 1. The countries are identified by their unique ISO2 code, and airports are identified by the unique IATA code. The scales in the geolocation name are separated by "/" to avoid ambiguity since child node of two countries may have the same name (e.g., "AU/New South Wales/Bathurst/Bathurst Airport" and "CA/New Brunswick/Bathurst/Bathurst Airport"). Since a geolocation may have multiple name aliases, node redundancy has been resolved at the best effort (e.g., "CN/Macau" and "CN/Macao" are redundant. In our dataset, only "CN/Macau" can be found).

For the United States, we have the most fine-grained information. Totally, we have 56 scale 2 locations for the USA: 50 states of USA plus six special cases: District of Columbia (a federal district), American Samoa, Guam, Northern Mariana Islands, Virgin Islands, and Puerto Rico. It is worth mentioning that the New York City actually contains five counties: New York County (Manhattan), Kings County (Brooklyn), Bronx County (The Bronx), Richmond County (Staten Island), and Queens County (Queens). The five children counties of the New York City is not included in our dataset under the reason that the characteristics or dynamics of the New York City have been paid more attention. Any properties of the node "US/New York/New York City" is an aggregation of its five counties. If you want to reduce space, in my opinion, the above "New York City" related description can be deleted, i.e., from "It is worth mentioning that" to "an aggregation of its five counties."

We identify the geolocation nodes and their basic geographic attributes, i.e., latitude and longitude, using all of the available data resources of geolocation property or connection (more details in the following subsections). For example, the airport nodes are obtained from the flight data, and the multiscale parents of an airport are obtained from the data resources of airport property.[16]

Geolocation Connection by Air

As demonstrated in Table 3, we utilize international flight data and domestic flight data of counties to connect geolocations at a given scale by air. We obtain the flight data from OpenFlight[16] and Bureau of Transportation Statistics (BTS) from the U.S. Department of Transportation.[17] Every flight is a directed edge connecting two airports, e.g., the edge "(CN/Guangdong/Guangzhou/CAN, US/California/Los Angeles/LAX)" denotes the air connection from the "CN/ Guangdong/Guangzhou/CAN" airport to the "US/California/Los Angeles/LAX" airport. It also denotes the air connection from "CN/Guangdong/

[b]In graph theory, a multigraph is a graph that is permitted to have multiple edges, that is, edges that have the same end nodes.
[c]The COVID-19 infection data are obtained from the Coronavirus Visual Dashboard operated by the Johns Hopkins University.[3]

Guangzhou" to "US/California/Los Angeles" at scale 3, "CN/Guangdong" to "US/California" at scale 2, and "CN" to "US" at scale 1. This characteristic is one of the reasons why the graph formulated by our dataset is a multigraph.

Geolocation Connection by Land

The edge, geolocation connection by land, indicates that the two geolocations are sharing borders. We provide border edges that connect worldwide countries, border edges that connect the U.S. scale 2 nodes (i.e., states), and border edges that connect the U.S. counties. The border edges are also bidirected because we believe the capacity of the border edges are asymmetric. Note that a geolocation node is guaranteed to have no self-loops (i.e., it does not have an edge pointing to itself).

Daily Passenger Number for Flights

In order to gain an understanding of how flights affect the population movement between the regions, we obtain daily passenger number for every pair of airports that have flights. To achieve that, T-100 Segment (all carriers) data from the BTS (the U.S. Department of Transportation)[17] are used. The BTS provides T-100 Segment data annually that contain nonstop segment data by aircraft type, transported passengers, transported freight and mail in pounds, available capacity (i.e., seats), origin airport information, destination airport information, aircraft hours, and distance, up to the maximum monthly scale. For the realistic simulation, very recent data are preferable but March 2019 data are chosen and March 2020 is estimated to be the same as the March 2019 because 2020 data were not provided when our data were first collected and the virus had a serious impact on the USA during March 2020. Of the data provided, only origin airport code, destination airport code, and passenger number are used, and the monthly transported passengers data are divided by 30 to estimate daily passenger number for every pair of airports that have flights. Since the passenger number provided for flights that are specific for freight and mail transportation (16.5% flights) is 0, we further add 1 for these flights to weight the flights properly. To summarize, these data indicate, on the normal days, the average daily passenger number for every pair of airports that have flights.

Geolocation Population

The region population data are from the U.S. Census Bureau,[18] United Nations (UN), Department of Economic and Social Affairs, and Population Division. The U.S. Census Bureau provides census statistics data every year up to city-level population for the United States. Original data have census survey data for 2010 and estimated population from 2011 to 2018. The 2018 estimated population is selected for simulation. For the other country-level populations, the UN's world population prospect 2019 data are utilized. This provides the estimation of the world population by the following basis from 1950 to 2020: location, single calendar year, single year of age, five-year age group, and gender.

Geolocated Government Nonpharmaceutical Intervention on COVID-19

To demonstrate an example of interesting applications that our dataset can provide, we collect data on the "stay-at-home" policy that the U.S. state governments have issued to address the COVID-19 pandemic with the corresponding effective time period (start date and end date). We obtain the data from The New York Times. A text description of the policy as well as its web link is also provided in our dataset. Note that not all states have both start date and end date. Some states only have a start date, but no end date was released till the point of the first collection of our dataset.

Geolocation Movement Pattern

To demonstrate the movement capability of the geolocations, we enrich our dataset with Google COVID-19 Community Mobility Reports.[19] This provides time-varying movement trend patterns for some geolocations (starting from "Feb. 15, 2020"). Each mobility trend is a set of time series presented by a community that is classified by space social functions (i.e., places, such as grocery stores and parks, within a geographic area) and highlights the percent change in people's visits. Specifically, these data show how visits and length of stay at different places change compared with the normal days. Google calculated these changes based on the Google Map data from users who have opted-in to Location History for their Google account. The six social functions are as follows:

1) *grocery and pharmacy*—places such as grocery markets, food warehouses, drug stores, and pharmacies;
2) *parks*—places such as parks, public beaches, marinas, plazas, and public gardens;
3) *transit stations*—places such as public transport hubs such as subway, bus, and train stations;
4) *retail and recreation*—places such as restaurants, cafes, shopping centers, museums, libraries, and theaters;
5) *residential*—places of residence;
6) *workplaces*—places of work.

TABLE 4. Features calculated from node static properties at each scale for Pearson and Spearman's correlation analysis.

Name	Description
POP	The size of population.
NUM_FLI	The number of flight edges that are connected to the node.
WEI_FLI	The total weight (daily passenger number) on the flight edges that are connected to the node.
NUM_NEI	The number of border-adjacent neighboring nodes at the same scale.
NUM_NEIF	The total number of flights that the border-adjacent neighboring nodes at the same scale have.
DIS_CNA	The Euclidean distance from coordinates (latitude and longitude) of this node to the coordinates of China.
FLI_CNA	Whether this node has direct flight from China.
ARRV	The arrival time of COVID-19 of this node, calculated as the number of days since 31 December 2020, which is the official date that China announced the first case of COVID-19.

DATA ANALYSIS

We examine our data and conduct the following analysis:

1) multiscale feature correlation comparison on node static properties (see the "Multiscale Correlation Analysis" section), 2) visualization of the relationship between the "stay-at-home" policy implementation start date with respect to the number of infected people on that day (see the "Stay-at-Home Order Analysis" section), and 3) Analysis of the dynamic node properties from an information theory viewpoint (see the "Dynamic Node Property Analysis" section).

Multiscale Correlation Analysis

Multiple types of features can be calculated from our dataset. Inspired from Adiga et al.[14] we calculate features based upon the node static properties, as given in Table 4 from scale 1 to 3, and conduct Pearson correlation analysis in Figure 2 and Spearman's correlation analysis in Figure 3.

From Figure 2, we could observe the following.

1) The number of flights and flight weights are highly and linearly correlated across all scales, with Pearson correlation coefficient (r) above 0.92.
2) At the scale 1, arrival time of COVID-19 is negatively linearly correlated with whether the country has direct flight from China, with $r = -0.54$.
3) At the scale 2, population of the state is highly and positively linearly correlated with the number of flights ($r = 0.86$) and the total flight weights ($r = 0.92$). Population of the state is also positively linearly correlated with whether the state has direct flight from China ($r = 0.62$). In addition, whether a state has direct flight from China is positively and linearly correlated with the total flight weight of this state ($r = 0.57$). On the other hand, the arrival time of COVID-19 of a state is negatively correlated with the total flight weight of this state ($r = -0.51$).
4) At the scale 3, only the number of flights and the total flight weights are highly correlated with whether having direct flight from China.
5) When we take all of the three scales into consideration, the linear correlation pattern is less obvious than looking at each scale individually, where the state scale, i.e. scale 2, has the most obvious correlation patterns. Interestingly, when we consider all scales together, distance from China becomes the most highly correlated feature for the arrival time of COVID-19, rather than population at the county or city scale, and flight-related features (especially whether having direct flight from China) at the country and state scale.

The patterns shown in Figure 2 can be summarized as the following: 1) the correlation patterns differ distinctively among the scales, which indicates the necessity of exploiting unique modeling strategies for each scale and the possibility to achieve performance gains by letting the scales aid each other; 2) as one potential use case of our dataset, the arrival time of COVID-19 could be predictable from whether having flight from China at the country scale and could be predictable from the total flight weight at the state scale. Other features, such as population, do matter and may play roles at different level of importance at each of the scale.

Pearson correlation indicates the hidden linear relationship between the features, whereas the Spearman's correlation indicates the existence of nonlinear relationship. Compared with Pearson correlation, the Spearman's correlation coefficients are overall higher (see Figure 3), indicating that a nonlinear model might be able to better capture the patterns hidden in our dataset. The number of flights and flight weights are

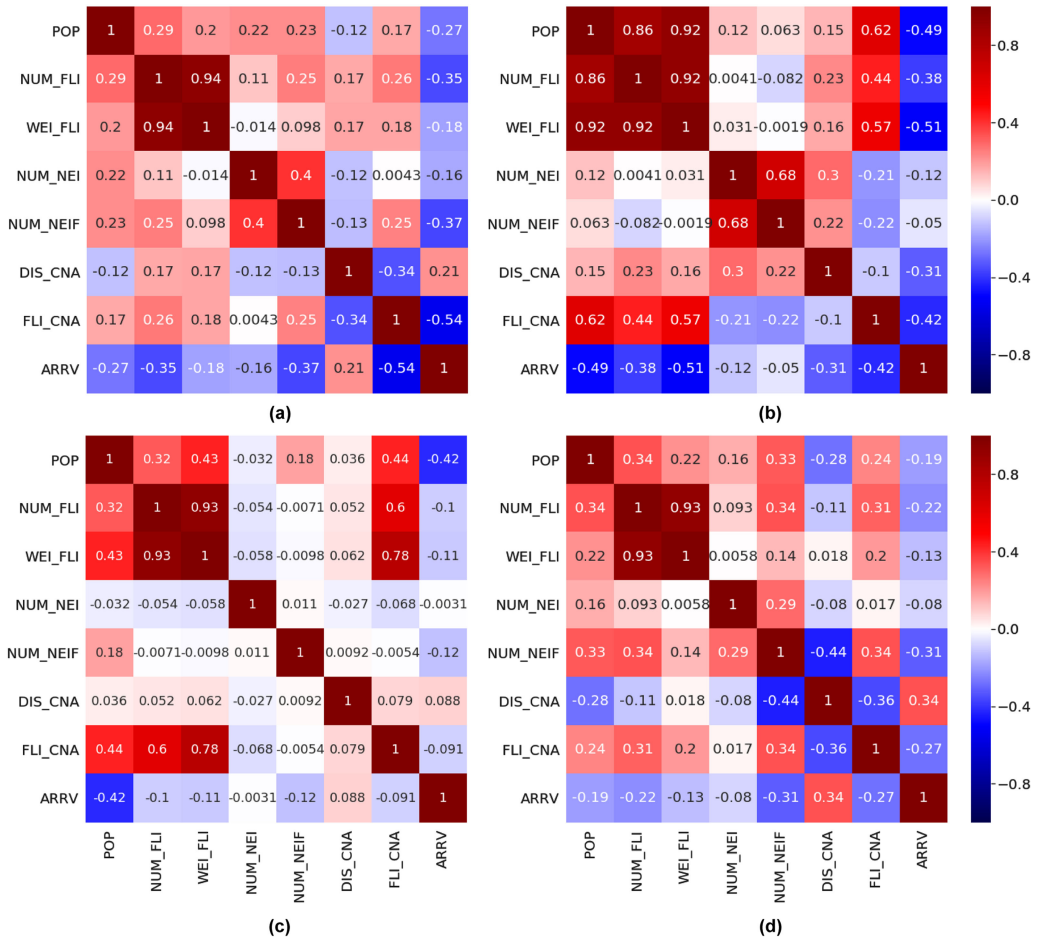

FIGURE 2. Pearson correlation analysis on features extracted from the static properties provided by our dataset. (a) Scale 1: Country. (b) Scale 2: State. (c) Scale 3: County. (d) Scales 1–3.

still highly correlated across all scales, however, when using the Spearman's correlation.

1) At the scale 1, population is highly correlated with the number of flights of this country and the number of border-adjacent neighbors that the country has. The number of flight is highly correlated with whether the country has flight from China. The arrival time of COVID-19 is highly negatively and nonlinearly correlated with the number of flights, and the total number of flights that this country's border-adjacent neighbor has.
2) At the scale 2, arrival time of COVID-19 is positively correlated with population, highly and negatively correlated with the flight-related features (number of flights and total flight weights). Population of the state is also highly and positively correlated with flight-related features. In addition, the number of border-adjacent neighbors is positively highly correlated with the total flights that these neighbors have.
3) Correlation pattern is much less obvious at the scale 3. Only population is strongly negatively correlated with the arrival time of COVID-19 (−0.72 correlation coefficient).

In summary: 1) the correlation patterns still differ distinctively among the scales, and the scale 3 is way different from the other two scales, and 2) there exist high nonlinear relationships between population, flight-related features and the arrival time of COVID-19.

Stay-At-Home Order Analysis

Based upon the features in Table 5 that we computed from our collected dataset, we conduct analysis on the stay-at-home order in Figure 5. Specifically,

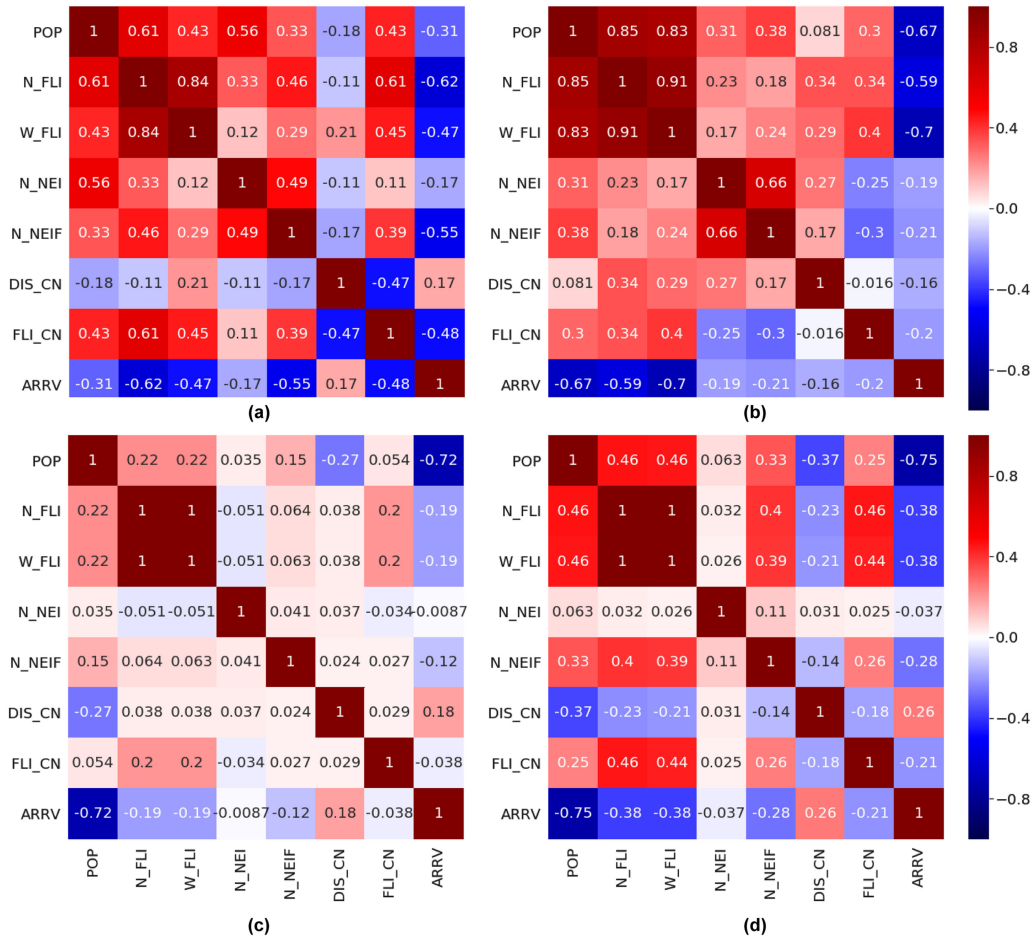

FIGURE 3. Spearman's correlation coefficient analysis on features extracted from the static properties provided by our dataset. (a) Scale 1: Country. (b) Scale 2: State. (c) Scale 3: County. (d) Scales 1–3.

we plot the relationship between the stay-at-home policy implementation start date (*x*-axis) with respect to the ratio of COVID-19 infected population on that day (*y*-axis) at the U.S. state level in Figure 5(a). As the figure shows, the later the implementation date is, the higher infected ratio can be observed, except for New York and the District of Columbia as two outliers, which suggests abnormal high rate of infection for these two states. In Figure 5(b), Pearson correlation analysis is conducted with features in Table 5; meanwhile, in Figure 5(c), Spearman's correlation analysis is conducted. We could observe high and positive correlation between population and the number of infected people in the state. High and positive correlation could also be observed between infected ratio and the number of infected people, as well as infected ratio and the days since China announced the first case of COVID-19 on the stay-at-home policy implementation day.

Dynamic Node Property Analysis

We compute and plot the mutual information in Figure 6(a) and the conditional entropy in Figure 6(b) between eight time series of different types of node dynamic properties that are provided by our collected dataset. Among the eight time series, six of them are mobility trend classified by space social functions and two of them are COVID-19-related time series, i.e., the daily accumulated infected people and the daily net infected people. On both plots, the diagonal axis shows the entropy of the corresponding time series. To interpret the plots, for example, the entropy of "Retail" is 5.3, as shown in Figure 6(a) and (b). The mutual information between "Retail" and "Grocery" is 3.8, as shown in Figure 6(a). The conditional entropy, "Retail" conditioned on "Grocery" is 1.5. Entropy of "Retail" is the sum of the mutual information between "Retail" and "Grocery" and the conditional entropy, i.e., "Retail" conditioned on "Grocery" (basically $H(X) = I(X;Y) + H(X|Y)$).

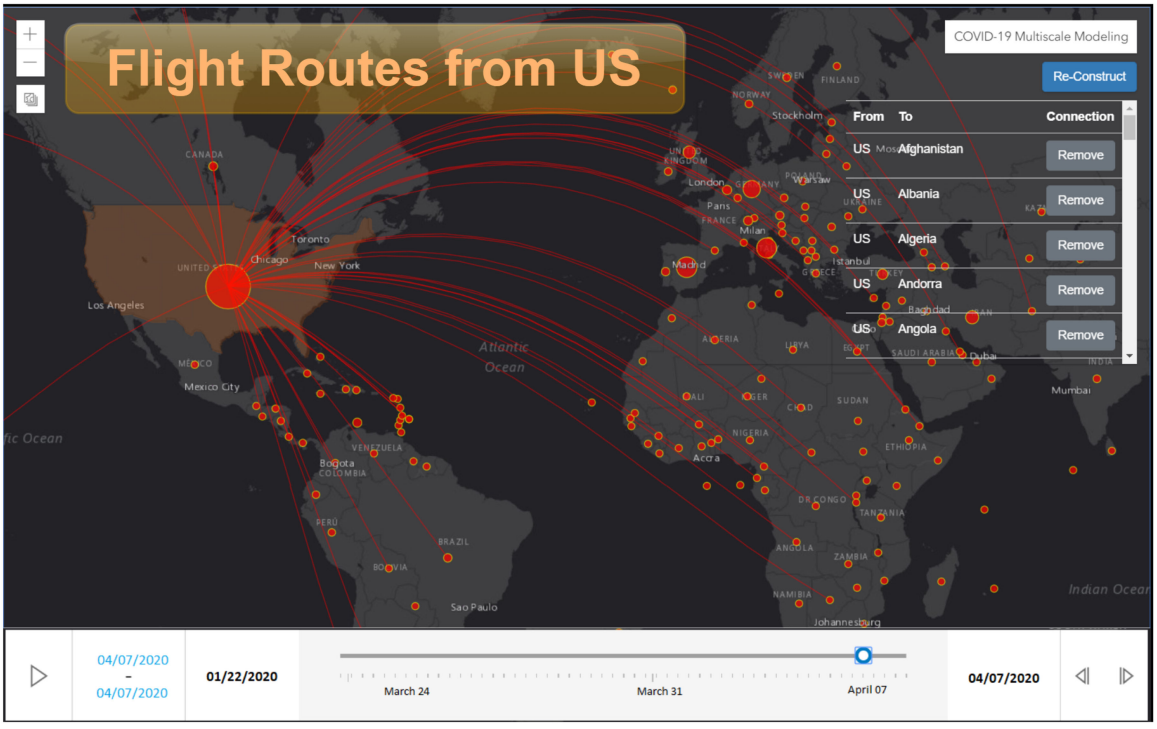

FIGURE 4. Outward flight routes from the US node to other countries. The scrollable area on the right-hand side shows pairwise flight connections between the USA and other countries. In front of every row, there is a button to remove that particular connection from the multiscale graph. The red circular disks are the nodes with varying sizes reflecting the number of active COVID-19 cases in that node (i.e., the bigger the disk size, the higher the number of active COVID-19 cases). The slider at the bottom allows moving between days to visualize the COVID-19 spread for a particular node or all the nodes.

TABLE 5. Features that are calculated to conduct the Pearson and Spearman's correlation analysis on the "stay-at-home" order.

Name	Description
Infected	The number of COVID-19 infected people on the stay-at-home policy implementation day.
Population	The size of population.
Delta_Days	On the stay-at-home policy implementation day, the number of days since 31 December 2020, which is the official date that China announced the first case of COVID-19.
Infected_Ratio	Infected divided by population.

Mutual information is a quantity that measures a relationship between two random variables that are sampled simultaneously. In particular, it measures how much information is communicated, on average, in one random variable about another. High mutual information indicates a large reduction in uncertainty, low mutual information indicates a small reduction, and zero mutual information between two random variables means the variables are independent. As we can see from the plots, the eight time series share high mutual information, indicating nonindependent relationships.

INTERACTIVE DATA VISUALIZATION SERVICE

We present an interactive web-based dashboard for visualization and downloading the multiscale geospatial dataset (visit this website for more details[d]). The platform enables the users to expand to different levels of abstraction from the scales of continent level, country,

[d][Online]. Available: https://usmannn.github.io/COVID-19-Multiscale-Simulation/

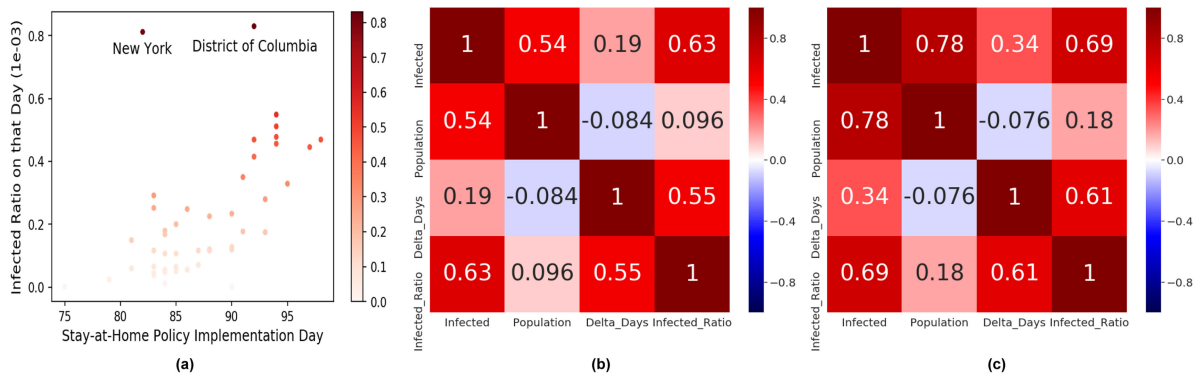

FIGURE 5. "Stay-at-home" order analysis. (a) Stay-at-home versus infected ratio. (b) Pearson correlation. (c) Spearman's correlation.

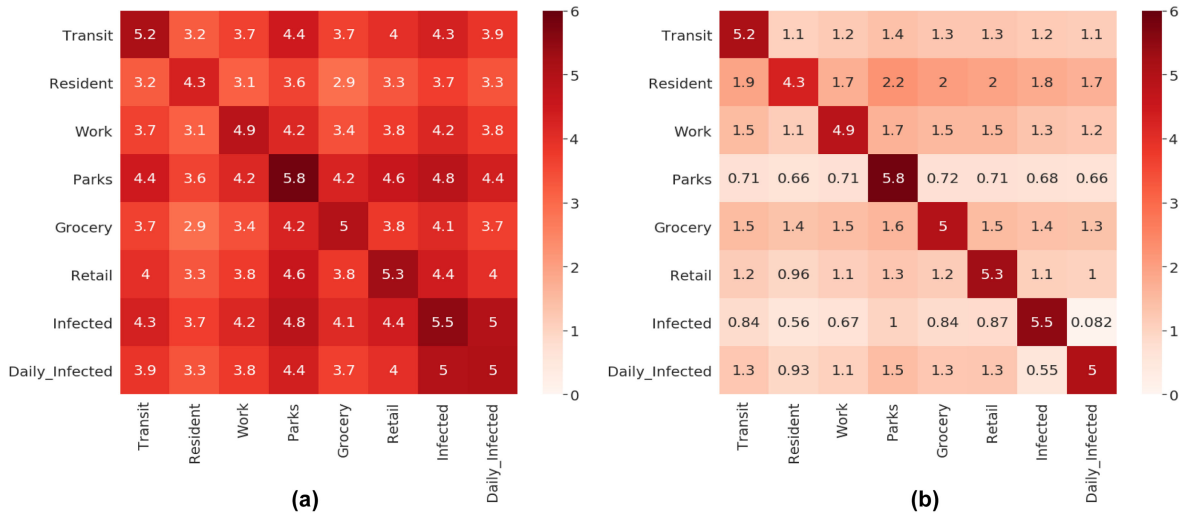

FIGURE 6. (a) Mutual information and (b) conditional entropy between the eight time series of different types of node dynamic properties.

cities, and provinces to individual neighborhoods, to interactively visualize the impact of COVID-19.

Similar efforts have been reported in the literature. For example, an extensive data visualization repository is presented for exploratory data analysis of COVID-19 pandemic.[20] An interactive dashboard is developed to track COVID-19 within U.S. states, counties, and cities.[21] These works, however, mostly perform at a single level of abstraction (e.g., country scale), and do not take into account any mobility data to model the flow of information between and within different scales, and solely rely on the estimated values to implicitly incorporate the transmission. They also restrict users to interact with different policies integrated into the modeling of the disease spread to visualize how the different policies may or may not impact the epidemiology.

In contrast, our platform enables the users to interact with different policies (e.g., closure of the borders and lockdowns), which are modeled into the multiscale network, to visualize their impact on the spread or control of the disease. More specifically, the platform allows to:

1) add/remove connections between different nodes (e.g., edges in the multiscale network);
2) add/remove certain geolocations (e.g., nodes in the multiscale network);
3) update the disease transmission rate on each connection (e.g., the width of the edges between nodes);
4) visualize temporal epidemiology trends;
5) move between different levels of abstraction (e.g., to different scales—country, state, city,

etc.) to visualize how a policy affects the spread or control of the disease at different scales;
6) download the multiscale geospatial dataset of the epidemiology for the user-explored policies.

Figure 1 shows an overview of our multiscale interactive web-based platform. The different images in the figure are resulted based on the user interactivity with the system. In the start (e.g., default view), the platform shows all the country scale nodes. A user can then interactively select one or more nodes to visualize their connections with other nodes, update the transmission rate of the disease on the selected edges, add or remove the nodes and/or edges from the multiscale network, and can reconstruct the graph based on the latest changes to observe their impact on the disease spread. Furthermore, the user can move between scales (e.g., from the country level to states with the country) to visualize how a policy change that is done at a higher scale may have impacted the trends (e.g., spread of the disease) in the lower scales.

Figure 4 shows a specific use case where a user has hovered over and clicked the US node. By doing so, the visualization platform has shown the active outward flight routes from the USA to other countries on different continents. For every active route (i.e., edge in the multiscale graph), the dashboard allows the user to remove it (e.g., enforce a border closure policy for a particular country) and then reconstruct the multiscale graph by hovering and clicking the reconstruct button. This would now allow the user to observe the implications after enforcing, for example, the border closure policy for one or more countries on the spread of COVID-19. The slider at the bottom enables the user to go back and forth between days and months to visualize the spread of COVID-19 over time, with or without the presence of any enforced policies.

IMPLEMENTATION DETAILS

We selected Python as the main implementation language for data processing and data analysis because of its advantages, such as versatility, extensibility, and convenient accessibility. HTML5 was used for the interactive visualization dashboard. The packages Pandas and Numpy were used to process data files (CSV and JSON), and Scipy and Numpy were used for data analysis. The computational resource we use for the collection and analysis of the dataset, as well as the development of the dashboard, is a single machine with 256 GB memory, and 48 CPU cores at 2.30 GHz.

DATA APPLICABILITY

In this section, we describe some applications and exemplary research questions (RQ) that future researchers can conduct and answer using our dataset.

RQ 1. Model "propagation" between regions: Our dataset could be used for scientific research to model and simulate "propagation" between regions at any given level. Ultimately, "propagation" is affected by the action (e.g., movement) of people. For example, joining our dataset with the data of COVID-19 cases, "propagation" would be spread of disease. The multiscale geospatial graph that our dataset provides can be considered as additional structural inputs for propagation prediction. The RQ is, how would movement between the regions affect the spread of disease of every region?

RQ 2. Identify interesting population clusters: Since our dataset provides rich population demographic information, researchers can use our dataset to identify interesting population clusters that may disclose some hidden but useful facts. For example, a region with higher than average elderly populations may be at high risk for COVID-19.

RQ 3. Predict the hospitalization needs of a geolocation: The RQ, would and how a region's properties determine the medical and hospital care needs, can be raised.

RQ 4. Understand the multiscale impact of policy: What if the New York City strictly banned air travel, would and how this affect other regions? RQ, such as would and how a region's policy affect another region within or across scales, can be raised.

RQ 5. Understand the impact of a region to another region: For example, the mobility trend of the New York state shows people tend to stay-at-home more. Would and how this affect the future mobility trend of the New York state's adjacent region such as New Jersey?

RQ 6. Capture and model the complex relationship of regions across scales: For example, would and how the New York State affect a county in New Jersey that is closest to New York? Would and how the total number of COVID-19 cases in the USA affect a county in Alaska?

RQ 7. Evaluation for graph-based model: Since our database functionality provides a graph with 1) heterogeneous type of nodes and bidirected edges and 2) both static and dynamic attributes of nodes and edges. Novel graph-based models, such as Graph Convolutional Neural Network, can be evaluated on our dataset.

RQ 8. Discover (geo)spatio-social patterns: Since our dataset provides a snapshot of the population

landscape of each geolocation, and the region can be country, province, county, city, and town, our dataset is useful for applications to discover (geo)spatio-social patterns. For example, determining vote populations for election campaign, understanding bilateral trade flows between regions, forecasting air quality by jointly modeling the impact from population, transportation route planning, hotel site selection, cross-city mobile traffic volume forecasting, etc. It is worth reminding that transportation route planning and hotel site selection are similar to the Traveling Salesman Problem (TSP). TSP's multiscale nature makes it a challenging graph task that requires reasoning about both local node neighborhoods as well as global graph structure. Our dataset accords with TSP's multiscale nature.

CONCLUSION

The emergence and rapid outbreak of COVID-19, unfortunately, has become a severe global concern. Researchers are actively taking responses to fight against this pandemic, thus giving rise to a number of related datasets and visualization dashboards. To address the limitation of the existing resources and enable more comprehensive research on computational epidemiology, we offer a multiscale geospatial dataset and an interactive visualization dashboard. Our analysis on the curated geospatial dataset indicates the necessity of multiscale modeling and provides valuable insights, such as the existence of high nonlinear relationships between population, flight-related features, and the arrival time of COVID-19. We formulate our dataset as a multiscale multigraph with heterogeneous type of nodes and bidirected edges and with both static and dynamic attributes residing on both nodes and edges. Because of this general formulation, our dataset have broad applications. Further, to aid research on computational epidemiology, our dataset jointly considers location geography, population demographics, nonpharmaceutical government intervention (i.e., stay-at-home policy), and timestamped population movement patterns. As our dataset provides a broad set of information, certain attributes are present only for partial nodes; augmenting the dataset to address this limitation can be an avenue worth exploring. In addition, using the collected dataset to answer various exemplary questions described in the "Data Applicability" section can be interesting future work directions.

ACKNOWLEDGMENTS

This publication is based upon work supported by King Fahd University of Petroleum & Minerals (KFUPM). Author(s) at KFUPM acknowledge the Interdisciplinary Research Center for Intelligent Secure Systems for the support. This work was also supported in part by NSF awards: IIS-1703883, IIS-1955404, IIS-1955365, RETTL-2119265, and EAGER-2122119, and in part by the U.S. Department of Homeland Security under Grant 22STESE00001 01 01.

REFERENCES

1. WHO, "COVID-19 situation reports," 2020. Accessed: Jun. 2022. [Online]. Available: https://rb.gy/qxukpf
2. E. Chen, K. Lerman, and E. Ferrara, "COVID-19: The first public coronavirus twitter dataset," 2020, *arXiv:2003.07372*.
3. E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *Lancet Infect. Dis.*, vol. 20, no. 5, pp. 533–534, 2020.
4. B. Xu et al., "Epidemiological data from the COVID-19 outbreak, real-time case information," *Sci. Data*, vol. 7, no. 106, pp. 1–6, 2020.
5. H. W. Hethcote, "The mathematics of infectious diseases," *SIAM Rev.*, vol. 42, no. 4, pp. 599–653, 2000.
6. C. C. Ku, T.-C. Ng, and H.-H. Lin, "Epidemiological benchmarks of the COVID-19 outbreak control in China after Wuhan's lockdown: A modelling study with an empirical approach," 2020. Accessed: Jun. 2022. [Online]. Available: https://ssrn.com/abstract=3544127
7. P. Patlolla, V. Gunupudi, A. R. Mikler, and R. T. Jacob, "Agent-based simulation tools in computational epidemiology," in *Proc. Int. Workshop Innov. Internet Community Syst.*, 2004, pp. 212–223.
8. P. F. Gorder, "Computational epidemiology," *Comput. Sci. Eng.*, vol. 12, no. 1, pp. 4–6, 2009.
9. W. Garira and F. Chirove, "A general method for multiscale modelling of vector-borne disease systems," *Interface Focus*, vol. 10, no. 1, 2020, Art. no. 20190047.
10. L. Wang, J. Chen, and M. Marathe, "Tdefsi: Theory guided deep learning based epidemic forecasting with synthetic information," *ACM Trans. Spatial Algorithms Syst.*, vol. 6, no. 3, 2020, Art. no. 15.
11. C. Yang et al., "EpiMob: Interactive visual analytics of citywide human mobility restrictions for epidemic control," *IEEE Trans. Vis. Comput. Graphics*, early access, Apr. 6, 2022, doi: 10.1109/TVCG.2022.3165385.
12. K. Samant, E. Memeti, A. Santra, E. Karim, and S. Chakravarthy, "Cowiz: Interactive COVID-19 visualization based on multilayer network analysis," in *Proc. IEEE 37th Int. Conf. Data Eng.*, 2021, pp. 2665–2668.
13. M. Chinazzi et al., "The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak," *Science*, vol. 368, no. 6489, pp. 395–400, 2020.

14. A. Adiga et al., "Evaluating the impact of international airline suspensions on the early global spread of COVID-19," *Medrxiv*, 2020.
15. A. Wilder-Smith and D. Freedman, "Isolation, quarantine, social distancing and community containment: Pivotal role for old-style public health measures in the novel coronavirus (2019-nCoV) outbreak," *J. Travel Med.*, 2020.
16. OpenFlight, "Openflight: Airport, airline and route data," 2017. Accessed: Jun. 2022. [Online]. Available: https://openflights.org/data.html
17. USDOT, "Bureau of transportation statistics," 2019. Accessed: Jun. 2022. [Online]. Available: https://www.transtats.bts.gov/databases.asp?Mode_ID=1&Mode_Desc=Aviation&Subject_ID2=0
18. USCB, "City and town population totals: 2010-2018," 2019. Accessed: Jun. 2022. [Online]. Available: https://rb.gy/x90vz4
19. Google, "Google COVID-19 community mobility reports," 2019. Accessed: Jun. 2022. [Online]. Available: https://www.google.com/covid19/mobility
20. S. K. Dey, M. M. Rahman, U. R. Siddiqi, and A. Howlader, "Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach," *J. Med. Virol.*, vol. 92, no. 6, pp. 632–638, 2020.
21. B. D. Wissel et al., "An interactive online dashboard for tracking COVID-19 in US counties, cities, and states in real time," *J. Amer. Med. Inform. Assoc.*, vol. 27, no. 7, pp. 1121–1125, 2020.

MUHAMMAD USMAN is an assistant professor in the Department of Information and Computer Science, King Fahd University of Petroleum and Minerals, Dhahran, 31261, Saudi Arabia, where he is also with the Center for Intelligent Secure Systems. His research interests include crowd simulation and modeling, human-centered AI, crowd behavior dynamics, architectural design analysis and optimization, spatial visualizations, and virtual reality. Usman received his Ph.D. degree in computer science from York University, Toronto, ON, Canada. He is the corresponding author of this article. Contact him at muhammad.usman@kfupm.edu.sa.

HONGLU ZHOU is a Ph.D. student in the Computer Science Department, Rutgers University (since 2017 Fall), Piscataway, NJ, 08854, USA, under the supervision of Prof. Mubbasir Kapadia. She has done internships at NEC Laboratories America and Google (YouTube). Her research interests include graph representation learning, crowd dynamics, and computer vision. Zhou received her Bachelor of Engineering degree in computer science and technology from the Communication University of China. Contact her at hz289@scarletmail.rutgers.edu.

SEONGHYEON MOON is a Ph.D. candidate at Rutgers University, Piscataway, NJ, 08854, USA, in computer science. He is a member of the Intelligent Visual Interfaces Lab supervised by Dr. Mubbasir Kapadia. His research interests include crowd simulation, crowd analysis, object segmentation, and object tracking. Moon received his M.Sc. degree in mechanical engineering from the Gwangju Institute of Science and Technology. Contact him at sm2062@cs.rutgers.edu.

XUN ZHANG is a Ph.D. candidate in the Department of Computer Science, Rutgers University, Piscataway, NJ, 08854, USA. His research interests include knowledge-powered semantics, virtual environments, and crowd narratives. Contact him at xz348@scarletmail.rutgers.edu.

PETROS FALOUTSOS is a professor at the Department of Electrical Engineering and Computer Science, York University, Toronto, ON, M3J 1P3, Canada, and an affiliate scientist at the Toronto Rehabilitation Institute, University Health Network, Toronto, ON M5G 2A2. Faloutsos received his Ph.D. degree in computer science from the University of Toronto, Toronto. Contact him at pfal@cse.yorku.ca.

MUBBASIR KAPADIA is the director of the Intelligent Visual Interfaces Lab and an associate professor in the Computer Science Department, Rutgers University, Piscataway, NJ, 08854, USA. His research lies at the intersection of artificial intelligence, visual computing, and human–computer interaction, with a mission to develop intelligent visual interfaces to empower content creation for human-aware architectural design, digital storytelling, and serious games. Kapadia received his Ph.D. degree in computer science from the University of California. His research is funded by DARPA and NSF, and through generous support from industrial partners including Disney Research, Autodesk Research, Adobe Research, and Unity Labs. He is a member of the IEEE. Contact him at mk1353@cs.rutgers.edu.