α-Satellite: An AI-driven System and Benchmark Datasets for Dynamic COVID-19 Risk Assessment in the United States

Yanfang Ye, Shifu Hou, Yujie Fan, Yiming Zhang, Yiyue Qian,
Shiyu Sun, Qian Peng, Mingxuan Ju, Wei Song, and Kenneth Loparo

***Abstract*— The fast evolving and deadly outbreak of coronavirus disease (COVID-19) has posed grand challenges to human society. To slow the spread of virus infections and better respond for community mitigation, by advancing capabilities of artificial intelligence (AI) and leveraging the large-scale and up-to-date data generated from heterogeneous sources (e.g., disease related data, demographic, mobility and social media data), in this work, we propose and develop an AI-driven system (named α-*Satellite*), *as an initial offering*, to provide dynamic COVID-19 risk assessment in the United States. More specifically, given a point of interest (POI), the system will automatically provide risk indices associated with it in a hierarchical manner (e.g., state, county, POI) to enable people to select appropriate actions for protection while minimizing disruptions to daily life. To comprehensively evaluate our system for dynamic COVID-19 risk assessment, we first conduct a set of empirical studies; and then we validate it based on a real-world dataset consisting of 5,060 annotated POIs, which achieves the area of under curve (AUC) of 0.9202. As of June 18, 2020, α-*Satellite* has had 56,980 users. Based on the feedback from its large-scale users, we perform further analysis and have three key findings: i) people from more severe regions (i.e., with larger numbers of COVID-19 cases) have stronger interests using our system to assist with actionable information; ii) users are more concerned about their nearby areas in terms of COVID-19 risks; iii) the user feedback about their perceptions towards COVID-19 risks of their query POIs indicate the challenge of public concerns about the safety versus its negative effects on society and the economy. Our system and generated datasets have been made publicly accessible via our website[1].**

***Index Terms*— AI System, Heterogeneous Data, Dynamic COVID-19 Risk Assessment, Community Mitigation.**

I. Introduction

Coronavirus disease (COVID-19) [1] is an infectious disease caused by a new virus that had not been previously identified in humans; this respiratory illness (with symptoms such as a cough, fever and pneumonia) was first identified during an investigation into an outbreak in Wuhan, China in December 2019 and is now rapidly spreading globally. The novel coronavirus and its deadly outbreak have posed grand challenges to human society. As of June 18, 2020, there have been 2,212,968 cases and 119,638 reported deaths in the U.S.; and the World Health Organization (WHO) characterized COVID-19 - that has infected more than 8,410,000 people with more than 450,000 deaths in at least 188 countries - a global pandemic.

It is believed that the novel virus which causes COVID-19 emerged from an animal source, but it is now rapidly spreading from person-to-person through various forms of contact. According to the Centers for Disease Control and Prevention (CDC) [2], the coronavirus seems to be spreading easily and sustainably in the community - i.e., *community transmission* which means people have been infected with the virus in an area, including some who are not sure how or where they became infected. An example of community transmission that caused the outbreak of COVID-19 in King county at Washington (WA) state is shown in Fig.1.

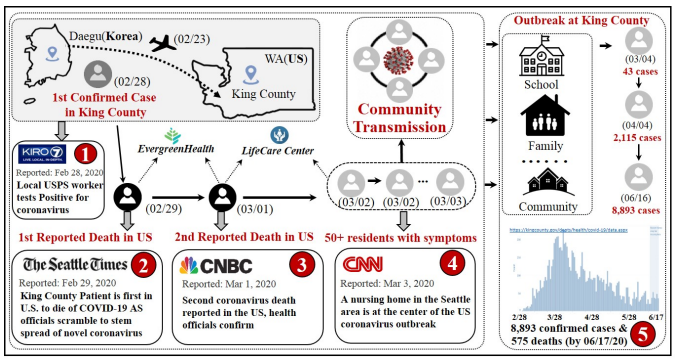


Fig. 1. An example of community transmission that caused an outbreak.

The challenge with community transmission is that carriers are often asymptomatic and unaware that they are infected and through their movements within the community they spread the disease. According to the CDC, before a vaccine or drug becomes widely available, *community mitigation*, which is a set of actions that persons and communities can take to help slow the spread of respiratory virus infections, is the most readily available interventions to help slow transmission of the

This work is partially supported by the NSF under grants IIS-2027127, IIS-1951504, CNS-2034470, CNS-1940859, CNS-1946327, CNS-1814825 and OAC-1940855, and by the DoJ/NIJ under grant NIJ 2018-75-CX-0032. The authors would also like to thank the strong support from Google for the use of Google Maps Platform.

The authors are with the Department of Computer and Data Sciences, Department of Electrical, Computer and Systems Engineering, Case Western Reserve University, Cleveland, OH, USA, (Corresponding Author: Yanfang Ye; e-mail: yanfang.ye@case.edu).

[1]https://COVID-19.yes-lab.org/

virus in communities [3]. A growing number of areas reporting community transmission would represent a significant turn for the worse in the battle against the novel coronavirus; this points to **an urgent need** for expanded surveillance so we can better understand the spread of COVID-19 and thus better respond with actionable strategies for community mitigation.

Unlike the 1918 influenza pandemic [4] where the global scope and devastating impacts were only determined well after the fact, COVID-19 history is being written daily, if not hourly, and if the right types of data can be acquired and analyzed there is the potential to improve self awareness of the risk to the population and develop proactive interventions. Realizing the true potential of real-time surveillance, with this opportunity comes the challenge: the available data are uncertain and incomplete while we need to provide actionable strategies with caution and rigor - i.e., enabling people to select appropriate actions for protection while minimize disruptions to daily life to the extent possible. To address this challenge, leveraging our long-term experiences in combating and mitigating widespread malware attacks using AI-driven techniques [5]–[8], in this work, we propose and develop *an AI-driven system to provide dynamic COVID-19 risk assessment at the first attempt* to help combat the fast evolving pandemic by using large-scale and up-to-date data generated from heterogeneous sources. More specifically, given a point of interest (POI), the developed system will automatically provide risk indices associated with it in a hierarchical manner (e.g., state, county, POI) to assist people with actionable information for community mitigation.

The framework of our proposed and developed system (named α-*Satellite*) is shown in Fig.2. In α-*Satellite*, (1) we first develop a set of tools to collect and preprocess the large-scale and up-to-date data related to COVID-19 from multiple sources; and then (2) we construct an attributed heterogeneous information network (AHIN) to model the collected multi-source data in a comprehensive way; (3) based on the constructed AHIN, to address the challenge of limited data that might be available for learning (e.g., social media data to learn public perceptions towards COVID-19 in a given area might not be sufficient), we propose a conditional generative adversarial net (cGAN) to gain the public perceptions towards COVID-19 in each given area; finally (4) we utilize meta-path based schemes to model both vertical and horizontal information associated with a given area, and devise a novel heterogeneous graph auto-encoder (GAE) to aggregate information from its neighborhood areas to estimate the risk of the given area in a hierarchical manner. The major contributions of our work can be summerized as followings:

- *Novel heterogeneous graph architecture*. To provide dynamic COVID-19 risk assessment for any given area (i.e., POI), we collect the large-scale and up-to-date data from multiple sources: i) disease related data (i.e., up-to-date county-based coronavirus related data); ii) demographic data from the United States Census Bureau; iii) mobility data that estimates how busy an area is in terms of traffic density; and iv) social media (i.e., Reddit) data. To model the multi-source data in a comprehensive manner, in this work, we present a novel heterogeneous graph architecture, i.e., AHIN, for abstract representation.
- *AHIN enrichment by cGAN*. In the constructed AHIN, there might be missing values of attributed features (e.g., limited social media data to learn public perceptions towards COVID-19 for a given area). To address this issue, we propose a cGAN for synthetic data generation for public perception learning to enrich the AHIN.
- *Heterogeneous GAE for dynamic COVID-19 risk assessment*. Based on the enriched AHIN, for any given area, we propose an innovative heterogeneous GAE to integrate both vertical information (i.e., information associated with its related city, county and state) and horizontal information (i.e., information from its neighborhood areas) for dynamic COVID-19 risk assessment.
- *The developed system and generated benchmark datasets have been made publicly accessible*. We first evaluate our system α-*Satellite* through a set of empirical studies; and then we validate it based on a real-world dataset consisting of 5,060 annotated POIs, which achieves the area of under curve (AUC) of 0.9202. As of June 18, it has had 56,980 users with the feedback from 7,348 of them. Based on the analysis of its large-scale user feedback, we have three key findings: i) people from more severe regions (i.e., with more COVID-19 cases) have stronger interests using our system for actionable information; ii) users are more concerned about their nearby areas in terms of COVID-19 risks; iii) user feedback about their perceptions towards COVID-19 risks of their query POIs indicate the challenge of public concerns of safety versus its negative effects on society and the economy.

II. Related Work

There have been many works on using AI and machine learning techniques to help combat COVID-19. In the biomedical domain, based on the image data (e.g., computed tomography (CT) and X-ray scans), extensive deep learning-based approaches [9]–[18] have been proposed to assist with COVID-19 diagnosis, prognosis and treatment. In pharmaceutical research area, there have been ample research studies [19]–[24] to investigate COVID-19 pharmaceuticals. For example, Google DeepMind [19] applies the proposed protein structure prediction system (i.e., AlphaFold) to predict the structures of several proteins associated with COVID-19 based on the corresponding amino acid sequences. Another research direction is to utilize social media and/or bibliometric data to help combat COVID-19 [25]–[30]. For example, [25] proposes to analyze Twitter data to understand the perceptions of COVID-19 outbreak across time and countries. Although the results are encouraging, the studies of using computational models to combat COVID-19 in the U.S. are scarce and there has no work on dynamic COVID-19 risk assessment for any given POI to assist with community mitigation by far. ***To meet this urgent need and to bridge the research gap***, in this work, by advancing capabilities of AI and leveraging the large-scale and up-to-date data generated from heterogeneous sources, we propose and develop an AI-driven system, named α-*Satellite*, to provide dynamic COVID-19 risk assessment in a hierarchical manner *at the first attempt* to help combat the fast evolving COVID-19 pandemic.

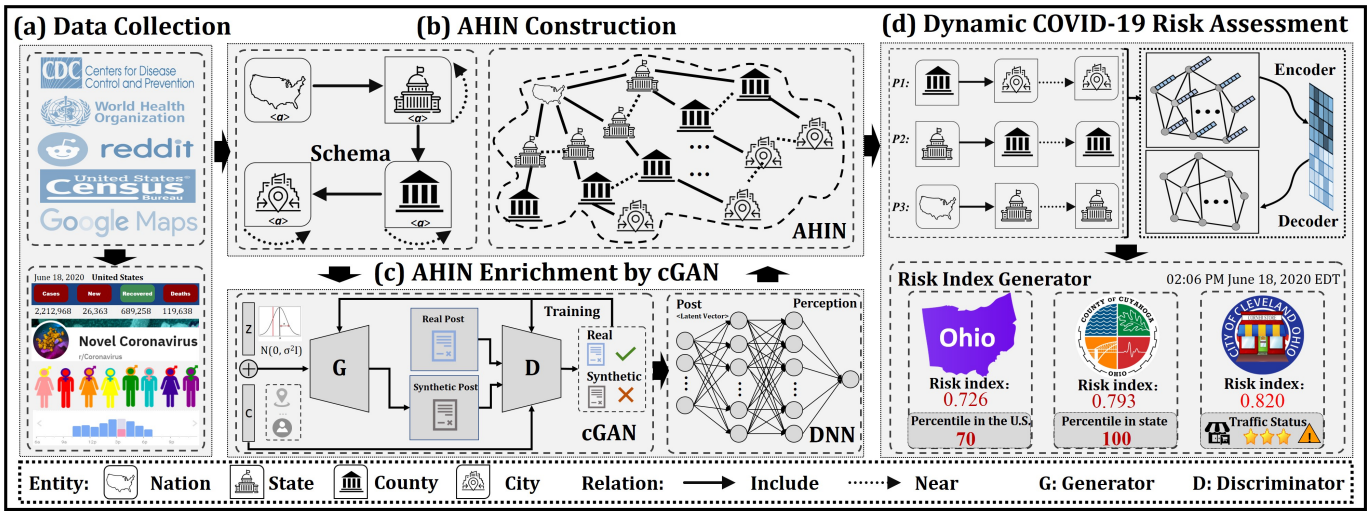


Fig. 2. System architecture of α-*Satellite* for dynamic COVID-19 risk assessment. In α-*Satellite*, (a) we first collect and preprocess large-scale and up-to-date data from heterogeneous sources; and then (b) we construct an AHIN to model the multi-source data in a comprehensive way; later (c) we devise a cGAN to enhance the public perceptions towards COVID-19 to enrich the constructed AHIN; finally (d) we utilize meta-path based schemes to model both vertical and horizontal information associated with a given area, and devise heterogeneous GAE to aggregate information from its neighborhood areas to estimate the risk of the given area in a hierarchical manner.

III. Proposed Method

In this section, we will introduce our proposed method for dynamic COVID-19 risk assessment in detail, which is integrated in our developed system α-*Satellite*.

A. AHIN Built from Heterogeneous Sources

Realizing the true potential of real-time surveillance requires identifying the proper data sources, based on which we can devise models to extract meaningful and actionable information for community mitigation. Since relying on a single data source for estimation and prediction often results in unsatisfactory performance, we develop a set of tools to collect and parse the large-scale and up-to-date data related to COVID-19 from multiple sources. We describe the collected data and their representations in detail below.

A1: disease related data. We collect the up-to-date county-based coronavirus related data including the numbers of confirmed cases, new cases, deaths and the fatality rate, from i) official public health organizations such as WHO, CDC, state and county government websites, and ii) digital media with nearly real-time updates of COVID-19 (e.g., 1point3acres [31]). For a given area, its related COVID-19 data will be represented by a numeric feature vector $\mathbf{a}_1$. For example, as of June 18, 2020, Cuyahoga county at Ohio (OH) state has had 5,336 cases, 65 new cases, 319 deaths and 6.0% fatality rate, which can be represented as $\mathbf{a_1} =< 5336, 65, 319, 0.060 >$.

A2: demographic data. The United States Census Bureau provides the demographic data including basic population, business, and geography statistics for all states and counties, and for cities and towns with more than 5,000 people. The demographic information may contribute to the risk assessment of an associated area: for example, as older adults may be at higher risk for more serious complications from COVID-19 [32], [33], the age distribution of a given area can be considered as an important input. In this work, given a specific area, we mainly consider its associated city's (or town's) demographic data, including the estimated population, population density (i.e., number of people per square kilometer), age distribution (i.e., percentage of people over 65 year-old), gender distribution (i.e., percentage of females), median individual income, and education (i.e., percentage with degrees of college or above). For example, given Cleveland at OH, its obtained demographic data are: Cleveland with population of 383,793, population density of 13,227, 13.5% people over 65 year-old, 51.8% females, median individual income of 18,387 and 46.1% population above 25-year old with degrees of college or above, which will be represented as $\mathbf{a}_2 =< 383793, 13227, 0.135, 0.518, 18387, 0.461 >$.

A3: mobility data. Given a specific area (either user input or automatic positioning), a mobility measure that estimates how busy the area is in terms of traffic density will be retained from location service providers (i.e., Google Maps), which is represented by five degree levels [1,5] (the larger the busier).

A4: social media data. Users in social media are likely to discuss and share their experiences of COVID-19, which may contribute complementary knowledge such as public perceptions towards COVID-19. In this work, we initialize our efforts with the focus on Reddit, as it provides the platform for scientific discussion of dynamic policies, announcements, symptoms and events of COVID-19. In particular, we consider i) three subreddits with general discussion (i.e., r/Coronavirus, r/COVID19 and r/CoronavirusUS); ii) four region-based subreddits (i.e., r/CoronavirusMidwest, r/CoronavirusSouth, r/CoronavirusSouthEast and r/CoronavirusWest); and iii) 48 state-based subreddits (i.e., Washington, D.C. and 47 states). To analyze public perceptions towards COVID-19 for a given area (note that all users are anonymized for analysis using hash values of usernames), we first exploit Stanford Named Entity Recognizer [34] to extract the location-based information (e.g., county, city), and then utilize tools such as NLTK [35] to

conduct sentiment analysis (i.e., negative, neutral or positive). More specifically, negative indicates less aware or pessimistic of COVID-19, and vice versa. For example, with the analysis of a user post in subreddit of r/CoronaVirusPA on March 14, 2020: *"I live in Montgomery County, PA and everyone here is acting like there's nothing going on."*, the location-related information of Montgomery county and Pennsylvania state (i.e., PA) can be extracted, and a public perception towards COVID-19 in Montgomery county at PA can be learned (i.e., negative indicating less aware of COVID-19). Another example post of "*As coronavirus spreads, northwest Louisiana prepares for its arrival*" indicates a positive signal. After performing the sentiment analysis based on the Reddit posts associated with a given area, the public perceptions towards COVID-19 in this area will be represented by a normalized value (i.e., [0,1], the larger value the more aware or optimistic).

After extracting the above features, we concatenate and normalize them as an attributed feature vector a attached to each given area for representation, i.e., $\mathbf{a_1} = \mathbf{a_1} \oplus \mathbf{a_2} \oplus \mathbf{a_1 3} \oplus \mathbf{a_4}$. We zero-pad the elements if the data are not available.

To comprehensively describe a given area for dynamic COVID-19 risk assessment, besides the above extracted attributed features, we further consider higher-level semantics and the rich relations among different areas.

R1: administrative affiliation. According to the severity of COVID-19, the available resources and impacts to the residents, different states may have different policies, strategies and orders responding to COVID-19. Accordingly, given an area, we extract its administrative affiliation in a hierarchical manner. Particularly, we acquire the *state-include-county* and *county-include-city* relations from City-to-County Finder [36].

R2: geospatial relation. We also consider the geospatial relations between a given area and its neighborhood areas. More specifically, given an area, we retain its k-nearest neighbors at the same hierarchical level by calculating the euclidean distances based on their global positioning system (GPS) coordinates obtained from Google Maps and Wikipedia [37].

Given the rich semantics and complex relations extracted above, it is important to model them in a proper way so that different relations can be better and easier handled. To solve this problem, we introduce AHIN to model them, which is able to be composed of different types of entities associated with attributed features and different types of relations.

Definition 1: ***Attributed Heterogeneous Information Network (AHIN)*** [38]: Let $\mathcal{T} = \{T_1, ..., T_m\}$ be a set of m entity types, $\mathcal{X}_i$ be the set of entities of type T_i and A_i be the set of attributes defined for entities of type T_i. An AHIN is defined as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ with an entity type mapping ϕ: $\mathcal{V} \rightarrow \mathcal{T}$ and a relation type mapping ψ: $\mathcal{E} \rightarrow \mathcal{R}$, where $\mathcal{V} = \bigcup_{i=1}^{m} \mathcal{X}_i$ denotes the entity set and $\mathcal{E}$ is the relation set, $\mathcal{T}$ denotes the entity type set and $\mathcal{R}$ is the relation type set, $\mathcal{A} = \bigcup_{i=1}^{m} A_i$, and $|\mathcal{T}|+|\mathcal{R}| > 2$. ***Network Schema*** [38]: The network schema of an AHIN $\mathcal{G}$ is a meta-template for $\mathcal{G}$, denoted as a directed graph $\mathcal{T}_\mathcal{G} = (\mathcal{T}, \mathcal{R})$ with nodes as entity types from $\mathcal{T}$ and edges as relation types from $\mathcal{R}$.

In this work, we have four types of entities (i.e., nation, state, county and city, $|\mathcal{T}| = 4$), two types of relations (i.e., *R1* and *R2*, $|\mathcal{R}| = 2$), and each entity is attached with an attributed feature vector $\mathbf{a}$ as described above. Based on the definitions, the network schema of AHIN is shown in Fig.3.

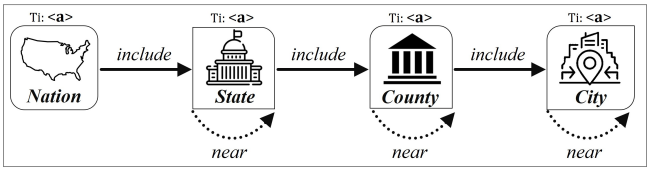


Fig. 3. The designed network schema of AHIN in our work.

B. AHIN Enrichment by cGAN

Although the constructed AHIN can model the complex and rich relations among different entities attached with attributed features, it faces a challenge that there may be missing values of attributed features attached to entities in the AHIN because of limited data that might be available for learning. More specifically, given an area, there may not be sufficient social media (i.e., Reddit in this work) data to learn the public perceptions towards COVID-19 in this area. For example, for the state of Vermont, as of June 18, 2020, in its corresponding subreddit r/CoronavirusVT, there only have been 19 posts by 15 users discussing the virus. To address this issue, we propose to exploit cGANs [39] for synthetic (virtual) data generation for public perception learning to enrich the AHIN.

Different from traditional GANs [40], a cGAN is a conditional model extended from GANs, where both the generator and discriminator are conditioned on some extra information. Here, we exploit cGAN for synthetic post vector generation. In our designed cGAN, given an area where Reddit data are limited or not available, the condition composes of: the disease related feature vector in this area $\mathbf{a}_1$, its related demographic feature vector $\mathbf{a}_2$ and its GPS coordinate denoted as $\mathbf{o}$.

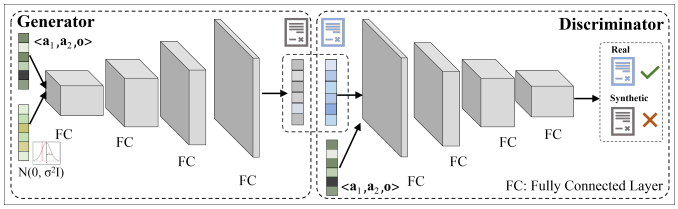


Fig. 4. The devised cGAN for synthetic post latent vector generation.

As shown in Fig.4, the generator in the devised cGAN aims to incorporate the prior noise $p_z(\mathbf{z})$ with conditions of $\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{o}$ as the inputs to generate synthetic posts represented by latent vectors; in the discriminator, real post representations obtained by using *doc2vec* [41] or generated synthetic post latent vectors along with conditions of $\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{o}$ are fed to a discriminative function. Both generator and discriminator could be a non-linear mapping function, such as a multi-layer perceptron (MLP). The generator and discriminator play the adversarial minimax game formulated as the following minimax problem:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{p}\sim p_{data}(\mathbf{p})}[\log D(\mathbf{p}|\mathbf{a}_1,\mathbf{a}_2,\mathbf{o})] + \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}[\log(1 - D(G(\mathbf{z}|\mathbf{a}_1,\mathbf{a}_2,\mathbf{o})))]. \quad (1)$$

The generator and discriminator are trained simultaneously: adjusting parameters for generator to minimize $\log(1 - D(G(\mathbf{z}|\mathbf{a}_1, \mathbf{a}_2, \mathbf{o})))$ while adjusting parameters for discriminator to maximize the probability of assigning the correct labels to both training examples and generated samples.

After applying cGAN for synthetic post latent vector generation, we exploit a deep neural network (DNN) with five fully-connected layers and one softmax layer to learn the public perceptions towards COVID-19 in this area. More specifically, we first use *doc2vec* to obtain the representations of real posts collected from Reddit and feed them to train the DNN model; and then given a generated synthetic post latent vector, we use the trained model to gain its related perception of COVID-19.

C. Dynamic COVID-19 Risk Assessment

To estimate the COVID-19 risk of a given area, it may not be sufficient if only considering its vertical information (e.g., information associated with its related city, county and state); the horizontal information (i.e., information from its neighborhood areas) will also be important inputs. To comprehensively integrate both vertical and horizontal information, we propose to exploit the concept of meta-path [42] to formulate the relatedness among different areas in the constructed AHIN.

Definition 2: ***Meta-path*** [42]: A meta-path $\mathcal{P}$ is a path defined on the network schema $\mathcal{T}_\mathcal{G} = (\mathcal{T}, \mathcal{R})$, and is denoted in the form of $T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} ... \xrightarrow{R_L} T_{L+1}$, which defines a composite relation $R = R_1 \cdot R_2 \cdot \ldots \cdot R_L$ between types T_1 and T_{L+1}, where $\cdot$ denotes relation composition operator, and L is the length of $\mathcal{P}$.

Based on the above definition, Fig.5.(a) shows our designed meta-paths (i.e., *P1-P3*). For example, *P1* of $county \xrightarrow{include} city \xrightarrow{near} city$ denotes that, to estimate the risk of a specific city, we not only consider the city itself, but also the information from its related county and nearby cities.

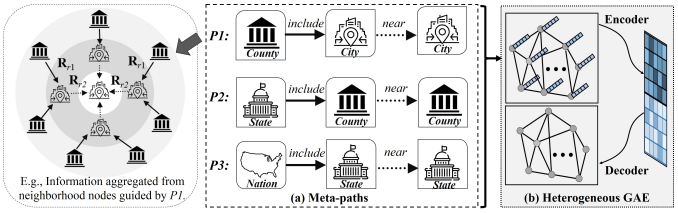


Fig. 5. The designed meta-paths and proposed heterogeneous GAE.

Given a node (i.e., area) in the constructed AHIN, guided by its corresponding meta-path scheme (i.e., city level guided by *P1*, county level guided by *P2*, and state level guided by *P3*), to aggregate the information propagated from its neighborhood nodes, we propose a heterogeneous graph auto-encoder (GAE) model to achieve this goal. The designed heterogeneous GAE model consists of an encoder and a decoder: the encoder aims at encoding meta-path based propagation to a latent representation, and the decoder will reconstruct the topological information from the representation.

Encoder. We here exploit attentive mechanism [43]–[45] to devise the encoder: it will first iteratively search the meta-path based neighbors $\mathcal{N}(v)$ for each node v, and then each node will attentively aggregate information from its neighbors. To learn the importance of information from neighborhood nodes, we first present each relation type $r \in \mathcal{R}$ in the constructed AHIN by $\mathbf{R}_r \in \mathbb{R}^{|\mathbf{a}| \times |\mathbf{a}|}$, where $|\mathbf{a}|$ denotes the dimension of the attributed feature vector; and then the attentive weight β of node u (the neighbor of v) indicates the relevance of these two nodes measured in terms of the space $\mathbf{R}_r$, that is,

$$\beta_r(v, u) = \mathbf{a}_v^T \mathbf{R}_r \mathbf{a}_u, \tag{2}$$

where $\mathbf{a}_v$ and $\mathbf{a}_u$ are the attributed feature vectors attached to node v and u. We further normalize the weights across all the neighbors of v by applying softmax function:

$$\widetilde{\beta}_r(v, u) = \frac{\exp(\beta_r(v, u))}{\sum_{u' \in \mathcal{N}(v)} \exp(\beta_r(v, u'))}. \tag{3}$$

Then, the neighbors' representations can be formulated as the linear combination:

$$\mathbf{a}_{\mathcal{N}(v)} = \sum_{u \in \mathcal{N}(v)} \widetilde{\beta}_r(v, u) \mathbf{a}_u, \tag{4}$$

where the weight $\widetilde{\beta}_r(v, u)$ denotes the information propagated from u to v in terms of relation r. Finally, we aggregate v's representation $\mathbf{a}_v$ and its neighbors' representations $\mathbf{a}_{\mathcal{N}(v)}$ by:

$$\mathbf{a}_v = avg(\mathbf{a}_v + \mathbf{a}_{\mathcal{N}(v)}). \tag{5}$$

Decoder. The decoder is used to reconstruct the network topological structure [46]: based on the latent representations generated from the encoder, the decoder is trained to predict whether there is a link between two nodes in the constructed AHIN. More specifically, the objective is to minimize the following reconstruction loss:

$$\mathcal{L} = -\sum_{v \in \mathcal{V}} \sum_{u \in \mathcal{V}} \mathcal{E}_{v,u} \log(\sigma(\mathbf{a}_u^\top \mathbf{a}_v)), \tag{6}$$

where $\mathcal{E}_{v,u}$ denotes the link between node v and u in AHIN, $\sigma(x) = 1/(1 + e^x)$ is the sigmoid function. We then perform stochastic gradient descent for the training.

To this end, leveraging latent representations learned from the heterogeneous GAE, the risk index of a given area is calculated as:

$$Idx(v) = \sum_{i=1}^{|\mathbf{a}|} \gamma_i \mathbf{a}_v(i), \tag{7}$$

where γ_i is the adjustable parameter that can be specified by human experts, which denotes the importance of i-th element in $\mathbf{a}_v$ (e.g., the case numbers, population density, age distribution, mobility measure, etc.) in the rapidly changing situation. More specifically, during different phases, the importance of different factors could be different. For example, compared with the stage of issuing stay-at-home order, the factor of mobility measure may overweight each of the other individual elements for dynamic risk assessment of a given POI after reopening. The risk index $Idx(v)$ will be normalized in the range of [0,1] (i.e., the larger value the higher risk).

IV. Experimental Results and Analysis

To meet the critical need to act promptly and deliberately in this rapidly changing situation, we have deployed our system α-*Satellite* for public test (https://COVID-19.yes-lab.org). Given a POI (either user input or automatic positioning), the developed system will automatically provide dynamic COVID-19 risk indices associated with it in a hierarchical manner (e.g., state, county, POI) to enable people to select appropriate actions for protection while minimizing disruptions to daily life. After we launched our system for public test, as of June 18, ***α-Satellite has had 56,850 users***. We describe our publicized benchmark datasets as well as the experimental results and analysis based on the large number of user feedback below.

A. Benchmark Datasets for Public Use

As described in Section III-A, we have collected and parsed the large-scale and up-to-date data related to COVID-19 from multiple sources. We describe each dataset in detail below, which has been made publicly available via our website.

DB_1: disease related dataset. We have collected the up-to-date county-based coronavirus related data including the numbers of confirmed cases, new cases, deaths and the fatality rate, from official public health organizations (e.g., WHO, CDC, state and county government) and digital media with nearly real-time updates (e.g., 1point3acres). It includes the data from 50 states, Washington, D.C., Puerto Rico and 3,208 counties on a daily basis from Feb. 28, 2020 to date.

DB_2: demographic and mobility dataset. We parse the demographic data collected from the United States Census Bureau in a hierarchical manner: for each city, county or state in the U.S., the data includes its estimated population, population density (e.g., number of people per square kilometer), age, gender, income and education distributions. We have made the demographic and mobility dataset publicly available including the information of estimated population, population density, and GPS coordinates for 31,140 cities, 3,208 counties, 50 states as well as Washington, D.C. and Puerto Rico.

DB_3: social media data from Reddit. In this work, we initialize our efforts on social media data with the focus of COVID-19 public perception analysis on Reddit. In particular, we have collected and analyzed 48 state-based subreddits (i.e., Washington, D.C. and 47 states). By the date, we have crawled and automatically analyzed 59,170 posts by 17,539 users on Reddit associated with 593,365 comments by 65,337 users on the discussion of COVID-19 from Feb. 17, 2020 to date. Along with these data, this publicized dataset also includes the sentiment analysis result of each post and comment.

DB_4: constructed AHIN. Based on our designed AHIN network schema (shown in Fig.3), the constructed AHIN has 34,401 nodes (i.e., 1 node with type of nation, 52 nodes with type of state, 3,208 nodes with type of county, 31,140 nodes with type of city) and 103,243 edges (i.e., 34,400 edges with relation type of *R1* and 68,843 edges with relation type of *R2*).

B. Evaluation of COVID-19 Risk Assessment

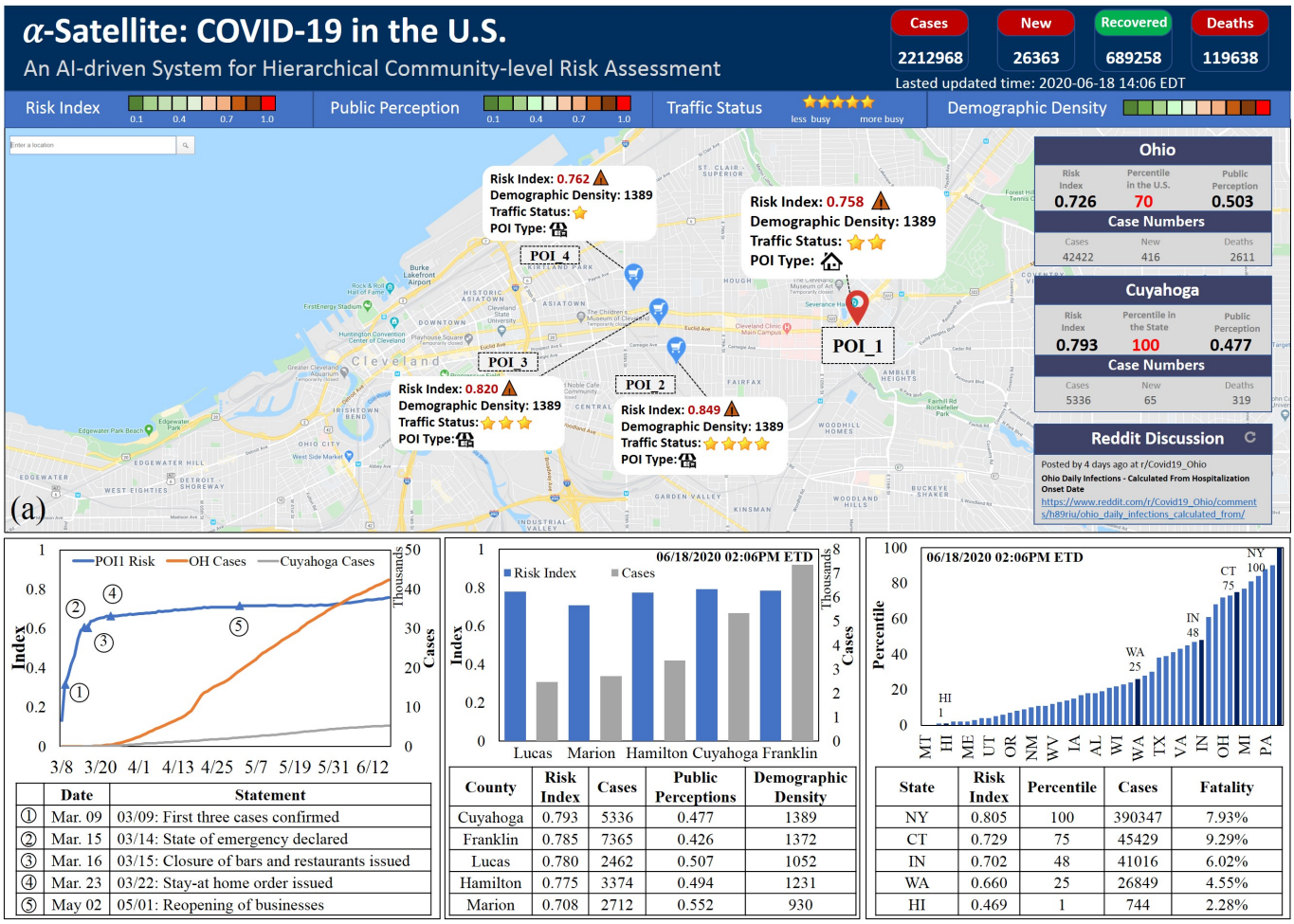
In this section, we comprehensively evaluate the performance of our developed system α-*Satellite* for dynamic COVID-19 risk assessment through a set of studies.

Study 1: risk index of a given area. Given a POI (either user input or automatic positioning by Google Maps), the developed system will automatically provide its related risk index (i.e., ranging from [0,1], the larger number indicates higher risk and vice versa) along with the public perceptions towards COVID-19 in this area (i.e., ranging from [0,1], the larger value denotes more aware or optimistic and vice versa), demographic density (i.e., the number of people per square kilometer in its related county), and traffic status (i.e., ranging from [1,5], the larger the heavier traffic and vice versa). Fig.6.(a) shows an example: given the POI of 10900 Euclid Ave, Cleveland, OH 44106 (denoted as POI_1), the risk index provided by the system was 0.758 indicating relatively high risk (i.e., demographic density of 1,389, and traffic status of 2) at 2:06pm EDT on June 18, 2020. Meanwhile, the risk indices of its corresponding county and state are also shown in a hierarchical manner: Cuyahoga county with risk index of 0.793, risk percentile of 100 in the state denoting highest risk among all the counties in OH, and public perception of 0.477; OH state with risk index of 0.726, risk percentile of 70 in the country denoting above medium-level of risk in the U.S., and public perception of 0.503. The provided risk indices of a given area could enable people for actionable information.

Study 2: comparisons of risk indices on different dates. In this study, given the same area of POI_1, we examine how the generated risk indices change over time. Fig.6.(b) shows the comparison results on different dates at the time of 2:06pm EDT, from which we have the following observations: (1) in general, its risk indexes increased over days from Mar. 8, 2020 (i.e., 0.131) to June 18, 2020 (i.e., 0.758), as the confirmed cases in its related county (i.e., Cuyahoga county) and its related state (i.e., OH) continued to grow; (2) after the first three case were confirmed in Cuyahoga county at OH on Mar. 9, there was a sharp rise of risk index compared with March 8 (from 0.131 to 0.314); (3) the risk growth rates relatively slowed down after the public health and executive orders were issued in responses to COVID-19: the government declared a state of emergency on Mar. 14, ordered Ohio bars and restaurants to close on Mar. 15 and issued a stay-at-home order on Mar. 22; (4) there has not yet dramatic growth of risks after the reopening of businesses since May 1 till mid-June.

Study 3: comparisons of risk indices at different areas. In this study, given the same time, we examine how the generated risk indices change over areas. When a user inputs the POIs in the search bar such as "*grocery stores near me*", the system will display the nearby grocery stores using Google Maps application programming interface (API) and automatically provide the associated indices. For example, using the same time in the first study, Fig.6.(a) shows nearby grocery stores of POI_1 and their related risk indices, from which we can see that the indices of nearby areas (i.e., POI_2-4) might vary due to multiple factors such as traffic statuses, POI types, etc.

Study 4: comparisons of different counties and states. In this study, we compare the indices of different counties and states given the same time. Using the time in the first study, Fig.6.(c)-(d) show examples of comparisons. More specifically, at county-level, using OH state as an example, we choose the counties with top five largest numbers of confirmed



	Date	Statement
①	Mar. 09	03/09: First three cases confirmed
②	Mar. 15	03/14: State of emergency declared
③	Mar. 16	03/15: Closure of bars and restaurants issued
④	Mar. 23	03/22: Stay-at home order issued
⑤	May 02	05/01: Reopening of businesses

(b) Comparisons of different dates (POI_1).

County	Risk Index	Cases	Public Perceptions	Demographic Density
Cuyahoga	0.793	5336	0.477	1389
Franklin	0.785	7365	0.426	1372
Lucas	0.780	2462	0.507	1052
Hamilton	0.775	3374	0.494	1231
Marion	0.708	2712	0.552	930

(c) Comparisons of different counties in OH.

State	Risk Index	Percentile	Cases	Fatality
NY	0.805	100	390347	7.93%
CT	0.729	75	45429	9.29%
IN	0.702	48	41016	6.02%
WA	0.660	25	26849	4.55%
HI	0.469	1	744	2.28%

(d) Comparisons of different states in US.

Fig. 6. Risk index of a given area (i.e., POI_1) and comparisons of the risk indices on different dates, in different counties and states in the US.

cases on June 18 for comparisons and Fig.6.(c) shows the risk indices are related with multiple factors (e.g., public perceptions, demographic distributions, fatality rates, information from nearby counties) rather than the case numbers only. Fig.6.(d) shows the risk percentiles of all states, whose comparisons also demonstrate the similar conclusion.

Study 5: systematical evaluation of α-*Satellite* for dynamic risk assessment. In this study, we systematically evaluate the performance of our system for dynamic risk assessment. After we launched our system for public test, we have asked a group of users (e.g., professors, students and staff in the university, editors, clinicians and company employees in OH) to use our system and annotate their query POIs (i.e., labeled as either relatively low risk (denoted as RL-risk) or relatively high risk (denoted as RH-risk)). As of June 18, we got 5,535 annotated POIs; by excluding the ones with conflicted annotations, we finally obtained 5,060 annotated POIs to build the ground-truth (i.e., 3,312 POIs labeled as RL-risk and 1,748 RH-risk). In the experiments, we empirically set the threshold ζ as 0.650 (i.e., if the risk index $Idx(v) \leq \zeta$, then the POI will be marked as RL-risk; otherwise, RH-risk). Based on the real-world dataset consisting of 5,060 annotated POIs, we use the widely-used metrics of accuracy (ACC), F1 measure and the area under curve (i.e., AUC) to quantitatively validate its performance. We first investigate the effectiveness of each extracted feature ($\mathbf{a_1}$: disease related data, $\mathbf{a_2}$: demographic data, $\mathbf{a_3}$: mobility data, $\mathbf{a_4}$: social media data) and the cGAN module. From Table I, we can see that: (1) adding each type of feature helps the performance of α-*Satellite*; (2) incorporating cGAN module into α-*Satellite* yields better results (ID5 vs. ID4), since cGAN enriches the AHIN by enhancing robust latent representations for accessing public perceptions towards COVID-19. Fig.7.(a) plots the training losses of generator and discriminator in cGAN, which demonstrates the training stability of cGAN. To further evaluate the system, we also compare it with the long short-term memory (LSTM) network. For LSTM, we consider the past 14-day data of a given POI as the input to train a prediction model (i.e., with five fully connected layers and one softmax layer). The results in Table I (ID5 vs. ID6) show that α-*Satellite* outperforms LSTM, which achieves an impressive AUC of 0.9202 (as shown in Fig.7.(b)). The reason behind this is that, for any given POI, α-*Satellite* not only considers its vertical information but also aggregates the information from its neighborhood areas (i.e., horizontal information).

During the public test of the system, we have receiving a number of good feedback from users in terms of the ease of use and its utility for COVID-19 risk assessment, such as: "*Thanks for putting together this tool. It's much needed and I hope will help curb transmission here in NEO.*""*I am on the Executive Leadership team of a group of 225 dental practices across the United States. ... I would like to get access to test your tool, as this could be a valuable tool for our clinicians.*" The experimental results and user feedback both demonstrate the effectiveness of our system for COVID-19 risk assessment.

TABLE I
SYSTEMATICAL EVALUATION OF α-SATELLITE.

Method	ID	Setting	ACC	F1	AUC
α-*Satellite*	1	$\mathbf{a_1}$	0.8098	0.7690	0.8435
	2	$\mathbf{a_1}$-$\mathbf{a_2}$	0.8368	0.7935	0.8638
	3	$\mathbf{a_1}$-$\mathbf{a_3}$	0.8646	0.8252	0.8875
	4	$\mathbf{a_1}$-$\mathbf{a_4}$	0.8934	0.8585	0.9110
	5	$\mathbf{a_1}$-$\mathbf{a_4}$, cGAN	**0.9120**	**0.8797**	**0.9202**
LSTM	6	$\mathbf{a_1}$-$\mathbf{a_4}$, cGAN	0.8705	0.8320	0.8919

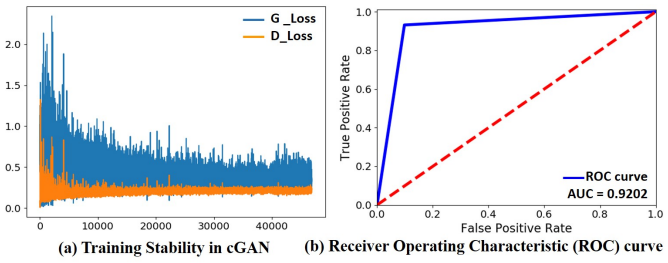
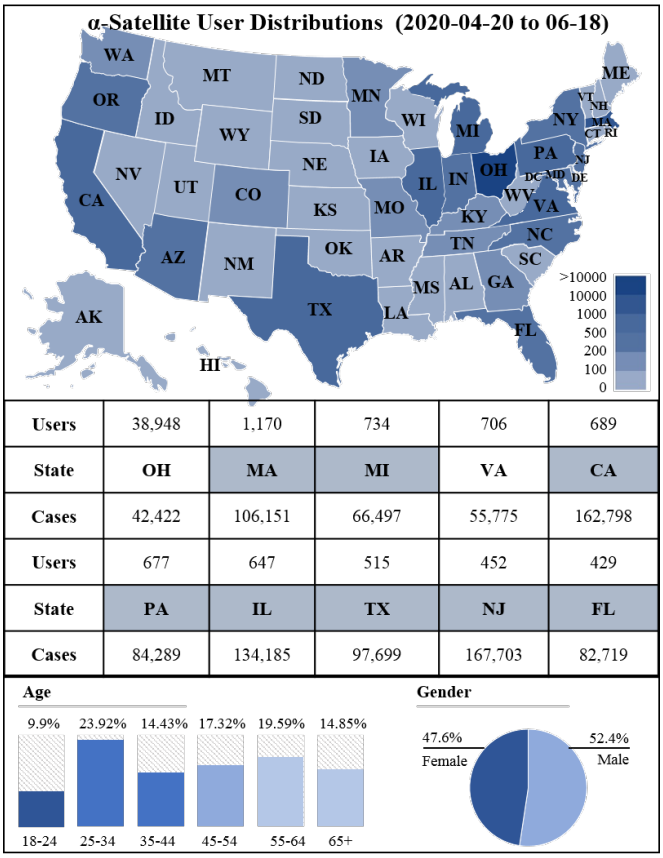


Fig. 7. The evaluation of α-Satellite for COVID-19 risk assessment.

C. Analysis of User Queries and Feedback

After we launched our system α-*Satellite* for public test, it has had 56,980 users as of June 18, 2020. In this section, based on Google Analytics platform and zip codes of user query POIs (i.e., all the data are anonymized), we perform further analysis of user queries and feedback.

Analysis 1: user distribution. In this study, based on Google Analytics platform, we analyze the distribution of 50,514 users from the U.S. who have visited our system. Fig.8 illustrates the geo- and demographic distributions of the users, from which we have following observations: (1) The system has attracted the users across all the states in the country; the top group of users are 25-34 years old (i.e., 23.92%) followed by the group of 55-64 years old (i.e., 19.59%), while males are relatively more than females. (2) The state of OH has largest number of users (i.e., 38,948 users accounting for 77.10%), which may be because people know our system mainly through local media releases. (3) The top ten states with largest numbers of users are listed in the table, eight out of which (as highlighted in the table) are the ones with largest numbers of COVID-19 cases. We further analyze the correlation between user and COVID-19 case distributions. Fig.9 shows the more severe regions with larger numbers of COVID-19 cases (both at state and county levels) the more α-*Satellite* users. The observation indicates that people from more severe regions (i.e., with larger numbers of COVID-19 cases) might have stronger interests using our system to assist with actionable information.



Users	38,948	1,170	734	706	689
State	**OH**	**MA**	**MI**	**VA**	**CA**
Cases	42,422	106,151	66,497	55,775	162,798
Users	677	647	515	452	429
State	**PA**	**IL**	**TX**	**NJ**	**FL**
Cases	84,289	134,185	97,699	167,703	82,719

Fig. 8. The geo- and demographic distributions of α-*Satellite* users.

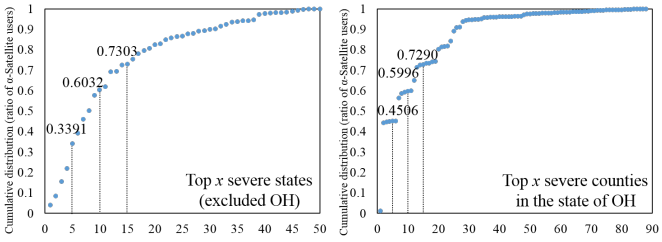


Fig. 9. The correlation between user and COVID-19 case distributions.

Analysis 2: user query POI distribution. After we launched the system for public test, among 50,514 users from the U.S., we got the feedback from 7,348 users in terms of their perceptions towards 9,048 query POIs. We further analyze the distributions of these POIs queried by users. From Fig.10.(a), we can see that 6,182 users (84.13%) only query one POI while 1,136 users (15.87%) query more than one POI. Based on the zip codes of query POIs, Fig.10.(b) shows that users are more interested in or concerned about their nearby areas (i.e., the overlaps indicate the percentages of users query the POIs within the same regions): 95.55% users query the POIs in the same states where they are; taking users in OH as an example, 96.86% of them query the POIs in the state of

OH. We also perform the analysis at county level: 90.91% users in Cuyahoga county at OH query the POIs in the same county (i.e., Summit with 83.53% and Lorain with 82.65% respectively). Fig.10.(c) illustrates the distribution of the top ten types of user query POIs (i.e., a POI can be with multiple types, such as Walmart can be with types of store and grocery), which shows that: people are more interested in or concerned about premise (e.g., particular building) that accounts for 45.35% and street address that is with 20.09%, followed by prominent local entity (e.g., airport), political (e.g., city hall), store (e.g., drug/food store), restaurant, grocery, health (e.g., hospital), transit station and finance (e.g., company).

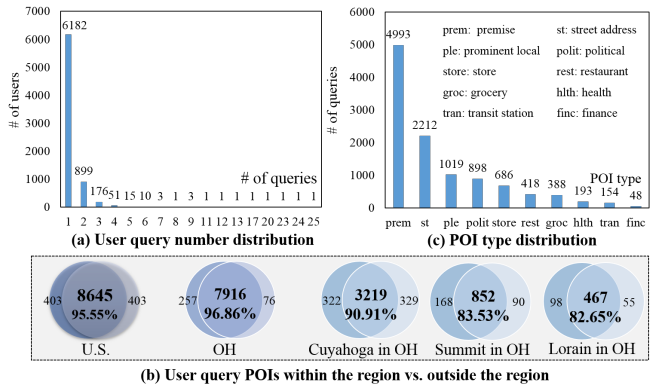


Fig. 10. The distributions of user query POIs.

Analysis 3: user feedback in terms of perceptions towards risks of query POIs. Our launched system enables users to provide their feedback in terms of their perceptions towards COVID-19 risks of their query POIs at five degree levels (i.e., extremely low risk, low risk, medium risk, high risk, and extremely high risk). Using the same dataset in the above analysis, based on the zip codes of user query POIs, we further investigate how users perceive the risk levels of their query POIs. In this study, we categorize the feedback of risks into three groups: low risk (including extremely low and low risks), medium risk, and high risk (including high and extremely high risks). Fig.11.(a) shows that, for all the feedback from 7,348 users related to 9,048 query POIs across the country, 45.59% are ranked as low risk while 41.03% are ranked as high risk; Fig.11.(b) and (c) illustrate similar distributions at the state and county levels respectively. From the analysis, we can see that a large portion of users do not regard COVID-19 as a serious risk while another large portion of users consider it as highly risky. This finding would indicate the difficult situation human society is currently facing - i.e., assuring people's safety and public health while mitigating the negative effects of COVID-19 on society and the economy is truly challenging.

V. Discussion

In the above experiments, we first evaluate our developed system α-*Satellite* for dynamic COVID-19 risk assessment through a set of empirical studies; and then we validate it based on a real-world dataset consisting of 5,060 annotated POIs, which achieves an impressive AUC of 0.9202. After we launched our system for public test, as of June 18, it has had

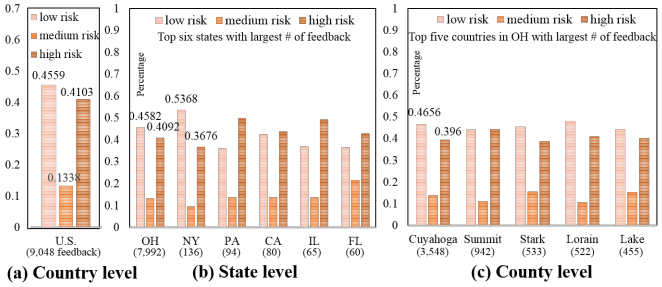


Fig. 11. The distributions of user perceptions about risks of query POIs.

56,980 users with the feedback from 7,348 of them. Based on the analysis of anonymized user feedback, we have three key findings: (1) People from more severe regions (i.e., with larger numbers of COVID-19 cases) have stronger interests using our system to assist with actionable strategies. (2) Users are more concerned about their nearby areas and the top types of POIs users queried are premise and street address followed by prominent local entity, political, store, restaurant, grocery, health, transit station and finance. (3) The user feedback in terms of their perceptions towards COVID-19 risks of their query POIs indicate the challenge of public concerns about safety versus its negative effects on society and the economy; as more and more places start to re-open, the situation could be more challenging. By advancing capabilities of AI and leveraging the large-scale and up-to-date data generated from heterogeneous sources, our proposed and developed system α-*Satellite* provides dynamic COVID-19 risk assessment to the public at the first attempt. After we launched the system for public test, the large number of its users indicate the high demand from the public for effective computational tools to assist people with actionable information.

VI. Conclusion

To track the emerging dynamics of COVID-19 pandemic in the U.S., in this work, we collect and model heterogeneous data from a variety of different sources, devise algorithms to use these data to train and update the models to predict the risks at hierarchical levels, and thus help provide actionable information to users for community mitigation. More specifically, given a POI, our developed system α-*Satellite* will automatically provide risk indices associated with it in a hierarchical manner to enable people to select appropriate actions for protection while minimizing disruptions to daily life. The system and generated benchmark datasets have been made publicly accessible through our website. To comprehensively evaluate α-*Satellite* for dynamic COVID-19 risk assessment, we first conduct a set of empirical studies; and then we validate it based on a real-world dataset consisting of 5,060 annotated POIs, which achieves the AUC of 0.9202. Based on the analysis of its large-scale users (56,980 users by June 18) and their feedback, we have three key findings as discussed above. The discovered knowledge indicates the challenge of assuring people's safety while mitigating the negative effects of COVID-19 on society and the economy. In the future work, we plan to expand the data collections

(e.g., traffic transmission data, Twitter data) and extend our model (e.g., introducing the series model to the original model) to improve its performance for risk estimations. We will continue releasing our datasets and system updates to facilitate researchers and practitioners to combat COVID-19 together.

REFERENCES

[1] WHO, *Coronavirus disease (COVID-19)*, 2020, https://www.who.int/.

[2] CDC, *How COVID-19 Spreads*, 2020, https://www.cdc.gov/coronavirus/2019-ncov/prepare/transmission.html.

[3] ——, *Implementation of Mitigation Strategies for Communities with Local COVID-19 Transmission*, 2020, https://www.cdc.gov/coronavirus/2019-ncov/downloads/community-mitigation-strategy.pdf.

[4] ——, *1918 Pandemic (H1N1 virus)*, 2020, https://www.cdc.gov/flu/pandemic-resources/1918-pandemic-h1n1.html.

[5] Y. Ye, S. Hou, L. Chen, J. Lei, W. Wan, J. Wang, Q. Xiong, and F. Shao, "Out-of-sample node representation learning for heterogeneous graph in real-time android malware detection," in *28th Int. Joint Conf. on Artif. Intell.*, 2019, pp. 4150–4156.

[6] Y. Ye, T. Li, D. Adjeroh, and S. S. Iyengar, "A survey on malware detection using data mining techniques," *ACM Comput. Surv.*, vol. 50, no. 3, p. 41, 2017.

[7] S. Hou, Y. Ye, Y. Song, and M. Abdulhayoglu, "Hindroid: An intelligent android malware detection system based on structured heterogeneous information network," in *Proc. of the 23rd ACM Int. Conf. on Knowl. Discovery and Data Mining*, 2017, pp. 1507–1515.

[8] Y. Ye, D. Wang, T. Li, and D. Ye, "Imds: Intelligent malware detection system," in *Proc. of the 13th ACM Int. Conf. on Knowl. Discovery and Data Mining*, 2007, pp. 1043–1047.

[9] X. Xu, X. Jiang, C. Ma, P. Du, X. Li, S. Lv, L. Yu, Y. Chen, J. Su, G. Lang *et al.*, "Deep learning system to screen coronavirus disease 2019 pneumonia," *arXiv preprint arXiv:2002.09334*, 2020.

[10] J. Chen, L. Wu, J. Zhang, L. Zhang, D. Gong, Y. Zhao, S. Hu, Y. Wang, X. Hu, B. Zheng *et al.*, "Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study," *medRxiv*, 2020.

[11] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng *et al.*, "A deep learning algorithm using ct images to screen for corona virus disease (covid-19)," *medRxiv*, 2020.

[12] Y. Song, S. Zheng, L. Li, X. Zhang, X. Zhang, Z. Huang, J. Chen, H. Zhao, Y. Jie, R. Wang *et al.*, "Deep learning enables accurate diagnosis of novel coronavirus (covid-19) with ct images," *medRxiv*, 2020.

[13] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song *et al.*, "Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct," *Radiol.*, p. 200905, 2020.

[14] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks," *arXiv preprint arXiv:2003.10849*, 2020.

[15] L. Wang and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest radiography images," *arXiv preprint arXiv:2003.09871*, 2020.

[16] O. Gozes, M. Frid-Adar, H. Greenspan, P. D. Browning, H. Zhang, W. Ji, A. Bernheim, and E. Siegel, "Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis," *arXiv preprint arXiv:2003.05037*, 2020.

[17] B. Ghoshal and A. Tucker, "Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection," *arXiv preprint arXiv:2003.10769*, 2020.

[18] H. S. Maghdid, A. T. Asaad, K. Z. Ghafoor, A. S. Sadiq, and M. K. Khan, "Diagnosing covid-19 pneumonia from x-ray and ct images using deep learning and transfer learning algorithms," *arXiv preprint arXiv:2004.00038*, 2020.

[19] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, A. Bridgland *et al.*, "Improved protein structure prediction using potentials from deep learning," *Nature*, pp. 1–5, 2020.

[20] N. Bung, S. R. Krishnan, G. Bulusu, and A. Roy, "De novo design of new chemical entities (nces) for sars-cov-2 using artificial intelligence," *ChemRxiv*, 2020.

[21] H. C. Metsky, C. A. Freije, T.-S. F. Kosoko-Thoroddsen, P. C. Sabeti, and C. Myhrvold, "Crispr-based surveillance for covid-19 using genomically-comprehensive machine learning design," *bioRxiv*, 2020.

[22] F. Hu, J. Jiang, and P. Yin, "Prediction of potential commercially inhibitors against sars-cov-2 by multi-task deep model," *arXiv preprint arXiv:2003.00728*, 2020.

[23] B. Tang, F. He, D. Liu, M. Fang, Z. Wu, and D. Xu, "Ai-aided design of novel targeted covalent inhibitors against sars-cov-2," *bioRxiv*, 2020.

[24] H. Zhang, K. M. Saravanan, Y. Yang, M. T. Hossain, J. Li, X. Ren, Y. Pan, and Y. Wei, "Deep learning based drug screening for novel coronavirus 2019-ncov," *Interdiscip. Sci.: Comput. Life Sci.*, p. 1, 2020.

[25] C. E. Lopez, M. Vasu, and C. Gallemore, "Understanding the perception of covid-19 policies by mining a multilanguage twitter dataset," *arXiv preprint arXiv:2003.10359*, 2020.

[26] L. Li, Q. Zhang, X. Wang, J. Zhang, T. Wang, T.-L. Gao, W. Duan, K. K.-f. Tsoi, and F.-Y. Wang, "Characterizing the propagation of situational information in social media during covid-19 epidemic: A case study on weibo," *IEEE Trans. Computat. Social Syst.*, vol. 7, no. 2, pp. 556–562, 2020.

[27] S. Du, J. Wang, H. Zhang, W. Cui, Z. Kang, T. Yang, B. Lou, Y. Chi, H. Long, M. Ma *et al.*, "Predicting covid-19 using hybrid ai model," *Available at SSRN: https://ssrn.com/abstract=3555202*, 2020.

[28] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala, "The covid-19 social media infodemic," *arXiv preprint arXiv:2003.05004*, 2020.

[29] X. Li, J. Yu, Z. Zhang, J. Ren, A. E. Peluffo, W. Zhang, Y. Zhao, K. Yan, D. Cohen, and W. Wang, "Network bioinformatics analysis provides insight into drug repurposing for covid-2019," *Preprints 202003.0286.v1*, 2020.

[30] M. M. Hossain, "Current status of global research on novel coronavirus disease (covid-19): A bibliometric analysis and knowledge mapping," *Available at SSRN: https://ssrn.com/abstract=3547824*, 2020.

[31] 1point3acres, *COVID-19 in US and Canada*, 2020, https://coronavirus.1point3acres.com/en.

[32] CDC, *Are You at Higher Risk for Severe Illness?*, 2020, https://www.cdc.gov/coronavirus/2019-ncov/specific-groups/high-risk-complications.html.

[33] V. Surveillances, "the epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (covid-19)-china, 2020," *China CDC Weekly*, vol. 2, no. 8, pp. 113–122, 2020.

[34] J. R. Finkel, T. Grenager, and C. D. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proc. of the 43rd Annu. Meet. of the Association for Comput. Linguistics*, 2005, pp. 363–370.

[35] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc", 2009.

[36] StatsIndiana, *City-to-County Finder*, 2020, http://www.stats.indiana.edu/uspr/a/place_frame.html.

[37] M. J, *Table of United States counties*, 2020, https://en.wikipedia.org/wiki/User:Michael_J/County_table.

[38] X. Li, Y. Wu, M. Ester, B. Kao, X. Wang, and Y. Zheng, "Semi-supervised clustering in attributed heterogeneous information networks," in *Proc. of the 26th Int. Conf. on World Wide Web*, 2017, pp. 1621–1629.

[39] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[40] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Adv. in Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[41] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. of the Int. Conf. on Mach. Learning*, 2014, pp. 1188–1196.

[42] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "Pathsim: Meta path-based top-k similarity search in heterogeneous information networks," *Proc. of the VLDB Endowment*, vol. 4, no. 11, pp. 992–1003, 2011.

[43] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[44] S. Fan, J. Zhu, X. Han, C. Shi, L. Hu, B. Ma, and Y. Li, "Metapath-guided heterogeneous graph neural network for intent recommendation," in *Proc. of the 25th ACM Int. Conf. on Knowl. Discovery & Data Mining*, 2019, pp. 2478–2486.

[45] X. Wang, X. He, Y. Cao, M. Liu, and T.-S. Chua, "Kgat: Knowledge graph attention network for recommendation," in *Proc. of the 25th ACM Int. Conf. on Knowl. Discovery & Data Mining*, 2019, pp. 950–958.

[46] T. N. Kipf and M. Welling, "Variational graph auto-encoders," *arXiv preprint arXiv:1611.07308*, 2016.