# Community Mitigation: A Data-driven System for COVID-19 Risk Assessment in a Hierarchical Manner

Yanfang Ye[1*], Yujie Fan[1], Shifu Hou[1], Yiming Zhang[1], Yiyue Qian[1]
Shiyu Sun[1], Qian Peng[1], Mingxuan Ju[1], Wei Song[1], Kenneth Loparo[2]
[1] Department of Computer and Data Sciences, Case Western Reserve University, OH, USA
[2] Department of Electrical, Computer, and Systems Engineering, Case Western Reserve University, OH, USA
{yanfang.ye,yxf370,sxh1055,yxz2092,yxq250,sxs2293,qxp36,mxj255,wxs338,kal4}@case.edu

## 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 with actionable strategies for community mitigation, leveraging the large-scale and real-time pandemic related data generated from heterogeneous sources (e.g., disease related data, demographic data, mobility data, and social media data), in this work, we propose and develop a data-driven system (named *α-Satellite*), *as an initial offering*, to provide real-time COVID-19 risk assessment in a hierarchical manner in the United States. More specifically, given a location (either user input or automatic positioning), the system will automatically provide risk indices associated with the specific location, the county that location is in and the state as a whole to enable people to select appropriate actions for protection while minimizing disruptions to daily life to the extent possible. In *α-Satellite*, we first construct an attributed heterogeneous information network (AHIN) to model the collected multi-source data in a comprehensive way; and then we utilize meta-path based schemes to model both vertical and horizontal information associated with a given location (i.e., point of interest, POI); finally we devise a novel heterogeneous graph neural network to aggregate its neighborhood information to estimate the risk of the given POI in a hierarchical manner. To comprehensively evaluate the performance of *α-Satellite* in real-time COVID-19 risk assessment, a set of studies are first performed to validate its utility; based on a real-world dataset consisting of 6,538 annotated POIs, the experimental results show that *α-Satellite* achieves the area of under curve (AUC) of 0.9378, which outperforms the state-of-the-art baselines. After we launched the system for public tests, it had attracted 51,190 users as of May 30. Based on the analysis of its large-scale users, we have a key finding that people from more severe regions (i.e., with larger numbers of COVID-19 cases) have stronger interests using the system for actionable information. Our system and generated benchmark datasets have been made publicly accessible through our website[1].

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


*CIKM '20, October 19–23, 2020, Virtual Event, Ireland*

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3412753

## CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Information systems** → *Information systems applications.*

## KEYWORDS

Data-driven System, Heterogeneous Data, Community-level COVID-19 Risk Assessment, Community Mitigation.

**ACM Reference Format:**
Yanfang Ye, Yujie Fan, Shifu Hou, Yiming Zhang, Yiyue Qian, Shiyu Sun, Qian Peng, Mingxuan Ju, Wei Song, Kenneth Loparo. 2020. Community Mitigation: A Data-driven System for COVID-19 Risk Assessment in a Hierarchical Manner. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3340531.3412753

## 1 INTRODUCTION

Coronavirus disease (COVID-19) [29] 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 in the United States and globally. The novel coronavirus and its deadly outbreak have posed grand challenges to human society. As of May 30, 2020, there have been 1,810,300 confirmed cases and 105,283 reported deaths in the United States; and the World Health Organization (WHO) characterized COVID-19 - infected more than 6,166,000 people with more than 372,000 deaths in at least 188 countries - a global pandemic. It is believed that the novel coronavirus 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) [5], 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.

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 virus in communities [6]. A growing number of areas reporting community transmission would represent a significant turn for the worse in the battle against the

novel virus; 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 [3] 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 (rather than reactive) interventions that can halt the exponential growth in the disease that is currently being observed. 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 objectively with caution and rigor - i.e., enabling people to select appropriate actions to protect themselves while minimize disruptions to daily life to the extent possible - to mitigate the negative effects of COVID-19 on public health, society, and the economy.

To address the above challenge, leveraging our long-term experiences in combating widespread malware attacks using data-driven techniques [11, 14, 15, 31–34], in this work, we propose to utilize the large-scale and real-time pandemic related data generated from heterogeneous sources to develop a data-driven system to provide real-time COVID-19 risk assessment in a hierarchical manner in the United States for community mitigation *at the first attempt*. More specifically, given a location (either user input or automatic positioning), the system will automatically provide risk indices associated with the specific location, the county that location is in and the state as a whole to enable people to select appropriate actions for protection while minimizing disruptions to daily life. The framework of our proposed and developed system (named *α-Satellite*) is shown in Figure 1. In *α-Satellite*, (1) we first develop a set of tools to collect and preprocess the large-scale and real-time data related to COVID-19 from multiple sources (including disease related data, demographic data, mobility data, and social media data); (2) we then construct an attributed heterogeneous information network (AHIN) to model the collected multi-source data in a comprehensive way; (3) based on the constructed AHIN, we utilize meta-path based schemes to model both vertical and horizontal information associated with a given location (i.e., point of interest, POI); and finally (4) we devise a novel heterogeneous graph neural network (GNN) to aggregate its neighborhood information to estimate the risk of the given POI in a hierarchical manner. The major contributions of our work can be summarized below:

- *Novel heterogeneous graph architecture for abstract representation.* To provide real-time COVID-19 risk assessment for any given location (i.e., POI), in this work, we collect the large-scale and real-time data from multiple sources: i) disease related data (i.e., up-to-date county-based coronavirus related data) from official public health organizations (e.g., WHO, CDC, state and county government websites) and digital media; 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, we present a novel heterogeneous graph architecture, i.e., AHIN which is capable of consisting multiple types of entities and relations, for abstract representation.
- *Heterogeneous GNN for real-time COVID-19 risk assessment.* Based on the constructed AHIN, for any given POI, we propose an innovative heterogeneous GNN to integrate both vertical information (i.e., the information associated with its related county and state) and horizontal information (i.e., the traffic transmissions from its neighborhood areas) for real-time COVID-19 risk assessment.
- *The developed system and generated benchmark datasets have been made publicly accessible to help combat COVID-19.* Based on a real-world dataset consisting of 6,538 annotated POIs, *α-Satellite* achieves the area of under curve (AUC) of 0.9378, which outperforms state of the arts in real-time COVID-19 risk assessment; we also perform a set of case studies to comprehensively validate its utility of COVID-19 risk estimations. After we launched the system for public tests, as of May 30, it had attracted 51,190 users. Based on the analysis of POI queries from the anonymized users, we have a key finding that people from more severe regions (i.e., with larger numbers of COVID-19 cases) have stronger interests using the system for actionable information.

## 2 RELATED WORK

There have been many works on using data-driven and machine learning techniques to help combat COVID-19. For example, in the biomedical domain, [7, 21, 28] use deep learning methods for COVID-19 pneumonia diagnosis and genome study; while [24, 30] develop learning-based models to predict severity and survival for patients. Another research direction is to utilize public accessible data to help the estimation of infection cases or forecast the COVID-19 outbreak [13, 16, 22]. However, many of the existing works are with focus on Wuhan China. The deadly outbreak of COVID-19 in the United States calls for novel computational models to help combat the pandemic; there has no work on real-time COVID-19 risk assessment 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 artificial intelligence (AI) and leveraging the large-scale and real-time pandemic related data generated from heterogeneous sources, we propose and develop a data-driven system to provide real-time COVID-19 risk assessment in a hierarchical manner in the United States *at the first attempt* to help combat the fast evolving COVID-19 pandemic.

## 3 PROPOSED METHOD

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

### 3.1 Data 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 real-time 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,

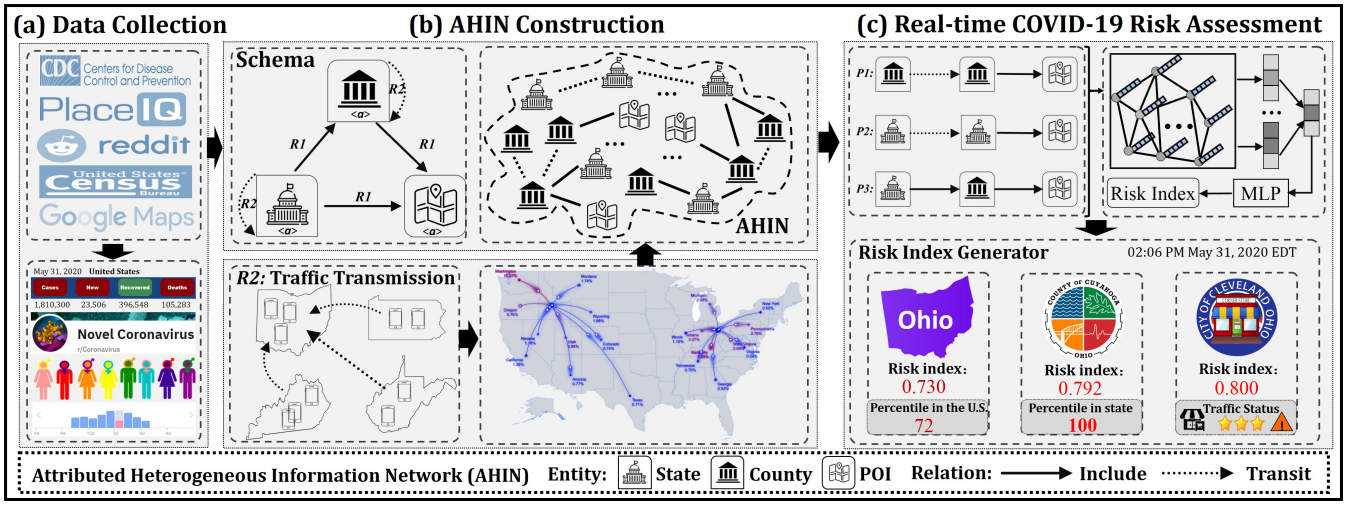



**Figure 1: System architecture of $\alpha$-*Satellite* for real-time COVID-19 risk assessment. In $\alpha$-*Satellite, (a)* we first collect the large-scale and real-time data related to COVID-19 from heterogeneous sources; and then *(b)* we construct an AHIN to model the collected multi-source data in a comprehensive way; finally *(c)* we devise heterogeneous GNN to aggregate both vertical and horizontal information from its neighborhood areas to estimate the risk of the given POI in a hierarchical manner.**

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 real-time updates of COVID-19 (e.g., 1point3acres [1]). For a given area, its related COVID-19 pandemic data will be represented by a numeric feature vector $\mathbf{a}_1$. For example, as of May 30, 2020, Cuyahoga county at Ohio (OH) state has had 4,369 confirmed cases, 51 new cases, 226 deaths and 5.2% fatality rate, which can be represented as $\mathbf{a_1} =< 4369, 51, 226, 0.052 >$. We denote this collected dataset as $\mathbf{DB}_1$, which includes the data from 50 states, Washington, D.C., Puerto Rico and 3,209 counties on a daily basis from Feb. 28, 2020 to date.

***A2: demographic data***. The United States Census Bureau provides the demographic data including basic population, business, and geography statistics for all states and counties. 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 [4], the age distribution of a given area can be considered as an important input. In this work, for a given area, we mainly consider its associated county'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) and gender distribution (i.e., percentage of females). For example, given an area associated with Cuyahoga county at OH, its obtained demographic data are: Cuyahoga county at OH with population of 1,235,072, population density of 1,389, 18.2% people over 65 year-old, and 52.3% females, which will be represented as $\mathbf{a}_2 =< 1235072, 1389, 0.182, 0.523 >$. We have made the dataset (denoted as $\mathbf{DB}_2$) publicly available including information of estimated population and population density for 3,209 counties, 50 states, Washington, D.C. and Puerto Rico.

***A3: mobility data***. Given a specific location (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). The data (denoted as $\mathbf{DB}_3$) including the Global Positioning System (GPS) coordinates for 3,209 counties, 50 states as well as Washington, D.C. and Puerto Rico have been made publicly accessible.

***A4: social media data***. As users in social media are likely to discuss and share their experiences of COVID-19, the data from social media may contribute complementary knowledge such as public perceptions towards COVID-19 in the area they associate with. 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 [12] to extract the location-based information (e.g., county, state), and then utilize NLTK tool [2] to conduct sentiment analysis (i.e., negative, neutral or positive). More specifically, negative indicates less aware or pessimistic of COVID-19, while positive denotes well aware or optimistic of COVID-19. For example, with the analysis of the post by a user (with hash value of "CF***6") 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 automatic sentiment analysis based on the collected posts associated with a given area from Reddit, 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). Such automatically extracted knowledge will be incorporated into the risk assessment of the related area, which may also help inform and educate about the science of coronavirus transmission and prevention. We have crawled and analyzed 54,881 posts by 16,689 users in Reddit associated with 536,996 comments by 60,962 users on the discussion of COVID-19 from Feb. 17, 2020 to date (denoted as $\mathbf{DB}_4$).

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., $a = a_1 \oplus a_2 \oplus a_3 \oplus a_4$. We zero-pad the elements if the data are not available.

## 3.2 AHIN Construction

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

***R1: vertical relation***. Based on the severity of COVID-19 and the available resources as well as the impacts to their 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, including the *state-include-county* and *county-include-city* relations [25].

***R2: horizontal relation***. For a given area, we also consider the estimated traffic transmissions from other states/counties to its associated state/county for risk estimations (i.e., in our application, we consider the top ten traffic transmissions from outside states/counties). The up-to-date traffic transmission data are obtained from PlaceIQ [8] in the way that among smartphones that pinged in a given state/county today, the share of those devices pinged in each state/county at least once during the previous 14 days. Figure 2.(a) shows examples of top ten traffic transmissions to Idaho (ID) and OH states on May 30, 2020 respectively.

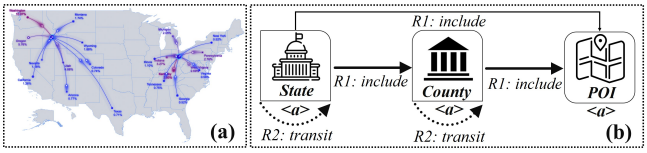



**Figure 2: (a) R2: Traffic transmission. (b) Network schema.**

Given the rich semantics and complex relations extracted above, it is important to model them in a proper way so that different relations among different types of entities 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. We first present the concepts related to *Attributed Heterogeneous Information Network (AHIN)* [18]: 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 $\phi: \mathcal{V} \rightarrow \mathcal{T}$ and a relation type mapping $\psi: \mathcal{E} \rightarrow \mathcal{R}$, 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* [18]: 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 three types of entities (i.e., state, county and POI, $|\mathcal{T}| = 3$), two types of relations (i.e., *R1* and *R2*, $|\mathcal{R}| = 2$), and each entity is attached with an attributed feature vector $a$ as described in Section 3.1. Based on the definitions above, the network schema of AHIN in our application is shown in Figure 2.(b).

## 3.3 Heterogeneous GNN for Risk Assessment

Based on the constructed AHIN, to comprehensively integrate both vertical and horizontal information for COVID-19 risk assessment, we first exploit the concept of meta-path [26] to formulate the relatedness among different areas. 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 $T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} \ldots \xrightarrow{R_L} T_{L+1}$, 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 definition, Figure 3.(a) shows our designed meta-paths (i.e., $\mathcal{P}_1$-$\mathcal{P}_3$). For example, $\mathcal{P}_1$ of $county \xrightarrow{transit} county \xrightarrow{include} POI$ denotes that, to estimate the risk of a specific POI, we not only consider the information from itself, but also the information from its related county and nearby counties (i.e., top ten counties with highest traffic transmissions to its related county).

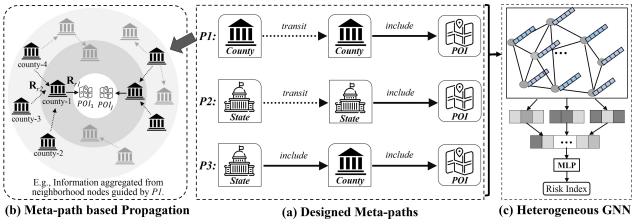



**Figure 3: Heterogeneous GNN for risk assessment.**

Given a node (i.e., POI) in the constructed AHIN, guided by the above designed meta-paths, we propose a heterogeneous GNN (shown in Figure 3.(c)) to aggregate its neighborhood information for real-time COVID-19 risk assessment, which is a three-step learning model: i) meta-path guided neighbor search, ii) information propagation and aggression, and iii) multi-view fusion.

***Meta-path guided neighbor search.*** To find neighbors of a node (i.e., POI) in the constructed AHIN, we first define *k-order neighbors in AHIN*: Given an AHIN $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, let 1-order neighbors of a node $v_i \in \mathcal{V}$ be $\mathcal{N}^1(v_i)$ so that $\mathcal{N}^1(v_i) = \{v_j | (v_i, v_j) \in \mathcal{E}\}$; then, $k$-order neighbors $\mathcal{N}^k(v_i)$ of a node $v_i$ ($k > 1$) can be denoted as $\mathcal{N}^k(v_i) = \{\mathcal{N}^1(v_z) \setminus \mathcal{N}^{(k-2)}(v_i), v_z \in \mathcal{N}^{(k-1)}(v_i)\}$. In our application, given a meta-path $\mathcal{P}$, for each node $v_i$ with type of POI, we will retrieve its 1-order and 2-order neighbors, denoted as $\mathcal{N}^1_\mathcal{P}(v_i)$ and $\mathcal{N}^2_\mathcal{P}(v_i)$ respectively. For example, given the meta-path of *P1*, as shown in Figure 3.(b), the 1-order neighbor of $POI_1$ is $\mathcal{N}^1_{\mathcal{P}_1}(POI_1)$={*county-1*} and its 2-order neighbors are $\mathcal{N}^2_{\mathcal{P}_1}(POI_1)$={*county-2,county-3,county-4*}.

***Information propagation and aggregation.*** For each given node $v_i$ with type of POI, after obtaining its 1-order neighbors $\mathcal{N}^1_\mathcal{P}(v_i)$

and 2-order neighbors $\mathcal{N}^2_{\mathcal{P}}(v_i)$ guided by a specific meta-path, we further consider the heterogeneity of AHIN to aggregate its neighborhood information for real-time COVID-19 risk assessment. More specifically, the information propagated from one node to another would associate with the type of relation between these two nodes. Therefore, we first introduce a relation-specific transformation $\mathbf{R} \in \mathbb{R}^{d \times d}$ for each relation type in the AHIN, where $d$ denotes the dimension of attributed feature vectors attached to the nodes. For each node $v_j$ in $\mathcal{N}^1_{\mathcal{P}}(v_i)$, we then propose the following mechanism for information propagation and aggregation:

$$\mathbf{E}_{v_j} = \sigma\big(\mathbf{W}(\mathbf{E}_{\mathcal{N}(v_j)} + \mathbf{E}_{v_j}) + \mathbf{b}\big), \tag{1}$$

$$\mathbf{E}_{\mathcal{N}(v_j)} = \sum_{v_z \in \mathcal{N}^2_{\mathcal{P}}(v_i)} w_{v_z v_j} \mathbf{R}_{\psi(v_z, v_j)} \mathbf{E}_{v_z}, \tag{2}$$

where $\mathbf{E}_{v_j}$ and $\mathbf{E}_{v_z}$ are embeddings of node $v_j$ and $v_z$ respectively which can be initialized by their attached attributed feature vectors, $\sigma$ is the activation function (i.e., LeakyReLU [19] in this work), $\psi(v_z, v_j)$ and $w_{v_z v_j}$ denote the type and weight of relation between node $v_z$ and $v_j$ respectively. In this way, Eq. (2) propagates the information from $v_i$'s 2-order neighbors in $\mathcal{N}^2_{\mathcal{P}}(v_i)$ to each of its 1-order neighbor $v_j$ in $\mathcal{N}^1_{\mathcal{P}}(v_i)$ in terms of relation type $\psi(v_z, v_j)$; and Eq. (1) aggregates $v_j$'s embedding $\mathbf{E}_{v_j}$ with the information $\mathbf{E}_{\mathcal{N}(v_j)}$ propagated from its neighbors. Similarly, the embedding of node $v_i$ with type of POI can be learned by:

$$\mathbf{E}_{v_i} = \sigma\big(\mathbf{W}'(\mathbf{E}_{\mathcal{N}(v_i)} + \mathbf{E}_{v_i}) + \mathbf{b}'\big), \tag{3}$$

$$\mathbf{E}_{\mathcal{N}(v_i)} = \sum_{v_j \in \mathcal{N}^1_{\mathcal{P}}(v_j)} w_{v_j v_i} \mathbf{R}_{\psi(v_j, v_i)} \mathbf{E}_{v_j}. \tag{4}$$

***Multi-view fusion.*** By applying the above proposed information propagation and aggregation method, given a specific meta-path, we are able to generate embedding of each node $v_i$ with type of POI. As different meta-paths depict the relatedness over nodes in the AHIN in different views. To this end, based on the three designed meta-paths (i.e., $\mathcal{P}_1$-$\mathcal{P}_3$), we propose to concatenate the corresponding embeddings to obtain the fused embedding $\mathbb{E}_{v_i}$:

$$\mathbb{E}_{v_i} = \bigoplus_{k=1}^{3} \mathbf{E}^{\mathcal{P}_k}_{v_i}. \tag{5}$$

After applying the proposed heterogeneous GNN model, we then feed each node (i.e., POI) embedding $\mathbb{E}_{v_i}$ to a classifier consisting of three-layer Multilayer Perceptron (MLP) to train the model, where the loss function is designed as:

$$\mathcal{L} = \sum_{v_i \in \mathcal{Y}} J(y_{v_i}, \hat{y}_{v_i}) + \gamma ||\Theta||_2^2, \tag{6}$$

where $J$ measures the cross-entropy loss between the annotated POI with label of $y_{v_i}$ and prediction score of $\hat{y}_{v_i}$, and $||\Theta||_2^2$ is the L2-regularizer for preventing over-fitting. The estimated risk index of $\hat{y}_{v_i}$ is in the range of [0,1] (i.e., the larger value the higher risk).

# 4 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 tests (https://COVID-19.yes-lab.org). Given a POI (either user input or automatic positioning), the developed system will automatically provide real-time COVID-19 risk indices associated with the POI, the county that POI is in and the state as a whole to enable people to select appropriate actions for protection while minimizing disruptions to daily life. After we launched our system for public tests on April 20, ***α-Satellite had attracted 51,190 users*** as of May 30. We describe our publicized benchmark datasets and the experimental results and analysis below.

## 4.1 Experimental Setup

**Generated datasets and deployed system for public use.** As described in Section 3.1, we have developed a set of tools to collect and parse the large-scale and real-time data related to COVID-19 from multiple sources, including disease related data, demographic data, mobility data, and social media (i.e., Reddit) data. We have made our collected and proprocessed datasets (i.e., $\mathbf{DB}_1$-$\mathbf{DB}_4$) publicly available through our website (https://COVID-19.yes-lab.org). Based on $\mathbf{DB}_1$-$\mathbf{DB}_4$, we construct an AHIN for COVID-19 risk assessment, which consists of 9,799 nodes (i.e., 52 nodes with type of state, 3,209 nodes with type of county, 6,538 nodes with type of POI) and 42,357 edges (i.e., 9,747 edges with relation type of *R1*, 32,610 edges with relation type of *R2*).

**Environmental and Parameter Settings.** The experiments are conducted in Ubuntu 19.10 operating system, plus two Intel i9-9900k, 4-way SLI GeForce RTX 2080 Ti Graphics Cards and 64 GB of RAM. We use Adaptive Moment Estimation (Adam) to optimize our model with learning rate of 0.005, and set epochs to 2000.

**Evaluation Metrics.** To quantitatively assess performances of different methods in COVID-19 risk estimations, we perform ten-fold cross validations and use the measures of precision, recall, accuracy (ACC), F1 and AUC (i.e., area under receiver operating characteristic (ROC) curve) for evaluations.

## 4.2 Utility of *α-Satellite* for Risk Assessment

We first evaluate the utility of our system *α-Satellite* for real-time COVID-19 risk assessment through a set of studies.

**Study 1: real-time risk index of a given POI.** Given a specific 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) associated with the public perceptions towards COVID-19 in this area (i.e., ranging from [0,1], the larger number 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 number means heavier traffic and vice versa). Figure 4.(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 0.720 indicating relatively high risk (i.e., demographic density of 1,389, and traffic status of 2) at 2:06pm EDT on May 31, 2020. Meanwhile, the risk indices of corresponding county and state are also shown in a hierarchical manner: Cuyahoga county with risk index of 0.792, risk percentile of 100 in the state denoting highest risk among all the counties in OH, and public perception of 0.514; OH state with risk index of 0.730, risk percentile of 72 in the country denoting relatively high risk among all the states in the U.S., and public perception of 0.506. If users input POIs in the search bar such as "*grocery stores near me*", then the system will display the nearby grocery stores using Google Maps application programming interface (API) and automatically provide related indices which may vary

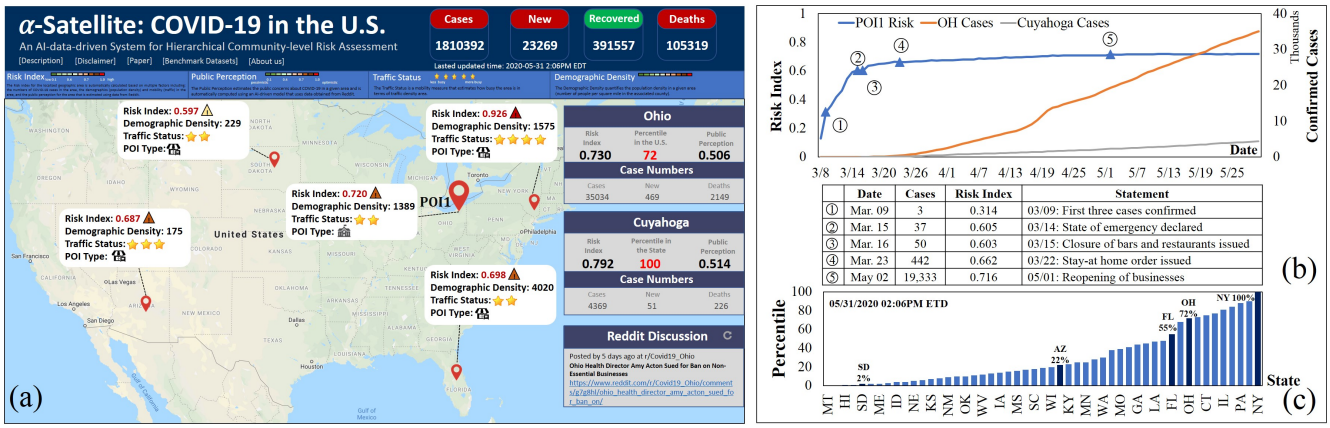



**Figure 4: Comparisons of risk estimations on different dates (given the same POI) and in different states (given the same time).**

due to multiple factors such as traffic statuses, POI types, etc. Given any POI, the provided risk indices in a hierarchical manner could assist people with community mitigation (i.e., selecting appropriate actions for protection while minimizing disruptions to daily life.

**Study 2: comparisons of risk indices on different dates.** In this study, given the same location of $POI_1$, we examine how the generated risk indices change over time. Figure 4.(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 March 8, 2020 (i.e., 0.131) to May 31, 2020 (i.e., 0.720), 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 March 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 March 14, ordered Ohio bars and restaurants to close on March 15 and issued a stay-at-home order on March 22; (4) there has not yet dramatic growth of risks after the reopening of businesses since May 1 till May 31.

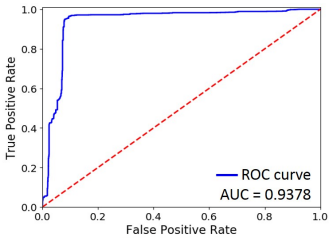
**Study 3: comparisons of risk indices in different areas.** In this study, given the same time, we compare the risk indices of different POIs in different states. Figure 4.(c) shows the risk percentiles of all states at 2:06pm EDT on May 31. For examples, New York (NY) is with 100 percentile, OH - 72 percentile, Florida (FL) - 55 percentile, Arizona (AZ) - 22 percentile, and South Dakoda (SD) - 2 percentile; Figure 4.(a) gives examples of risk indices of specific POIs in these states for illustration. The comparisons of POIs in different states indicate that the risk indices are positively correlated to the numbers of confirmed cases in general but also associated with other complicated factors such as fatality rate, demographics, traffic transmissions, public perceptions, etc.

## 4.3 Systematic Evaluation of *α-Satellite*

In this study, we systematically evaluate the performance of *α-Satellite* for real-time risk assessment. We launched our system for beta test on April 20 and asked a group of users (e.g., professors, students and staff in the university, editors, clinicians and company employees) to use our system and annotate their query POIs (i.e., either relatively low risk (denoted as RL-risk) or relatively high risk (denoted as RH-risk)). As of May 30, we got 6,992 annotated POIs; by excluding the ones with conflicted annotations (i.e., POIs with different labels), we finally obtained 6,538 annotated POIs as the ground-truth (i.e., 2,201 POIs labeled as RL-risk and 4,337 labeled as RH-risk). Figure 5.(a) shows the results with different settings, from which we can see that: (1) for different meta-paths, $\mathcal{P}_1$ and $\mathcal{P}_2$ encoding traffic transmission information perform better than $\mathcal{P}_3$; (2) *α-Satellite* with the combination of three meta-paths ($\mathcal{P}_1$-$\mathcal{P}_3$) outperforms individual one in COVID-19 risk assessment; (3) *α-Satellite* utilizing AHIN representations performs better than merely using augmented features (denoted as *Augment*, that is, given a POI, we concatenate its attributed features with the features of its associated county and state); and finally (4) α-Satellite using relation-specific transformation in the proposed heterogeneous GNN - which differentiates different types of relations in the AHIN - obtains better results than the method (denoted as *Variant*) that simply treats each type of relation equally in the information propagation and aggregation process. Figure 5.(b) illustrates that *α-Satellite* achieves an impressive AUC of 0.9378.

| Setting | Accuracy | F1 |
|---|---|---|
| $P_1$ | 0.8731 | 0.8279 |
| $P_2$ | 0.8702 | 0.8241 |
| $P_3$ | 0.8559 | 0.8067 |
| *Augment* | 0.8483 | 0.7960 |
| *Variant* | 0.8954 | 0.8569 |
| ***α-Satellite*** | **0.9239** | **0.8947** |

**(a) Accuracy and F1 measure**




**(b) ROC curve**

**Figure 5: Systematic evaluation of real-time risk assessment.**

## 4.4 Comparisons with Baselines

In this section, we evaluate the performance of *α-Satellite* in COVID-19 risk assessment by comparisons with the state-of-the-art baselines, including network embedding methods (i.e., DeepWalk, metapath2vec), and GNN-based models (i.e., GCN, GAT, RGCN, MEIRec).

- **DeepWalk** [20] performs truncated random walk and skip-gram model for node embeddings in homogeneous network.
- **metapath2vec** [9] learns latent HIN representations by performing meta-path guided random walk and skip-gram model.
- **GCN** [17] is a semi-supervised graph convolutional network that averages the neighbors' embeddings with linear projection.
- **GAT** [27] is a graph attention network model that aggregates information of neighbors via self-attention mechanism.
- **RGCN** [23] is designed for heterogeneous graph and considers different relations between nodes for information aggregation.
- **MEIRec** [10] is a heterogeneous GNN that merely propagrates and aggregates information of meta-path guided neighbors.

For the methods of DeepWalk and metapath2vec, since they are incapable of dealing with the attributes attached to the nodes, we concatenate the attributed feature vector with the learned node embedding for each node, which is fed to a classifier with three-layer MLP for training and prediction. For GNN-based models that are designed for homogeneous network (i.e., GCN, GAT), we first transform the AHIN to the corresponding homogeneous graph based on each meta-path, and then apply GCN and GAT on each homogeneous graph. Here we test all the meta-paths for GCN and GAT, and report the best performances. The comparison results are shown in Table 1, from which we can see that: (1) generally, GNN-based models (i.e., GCN, GAT, RGCN and MEIRec) which combine the node attributes and structural information in a more comprehensive manner yield better performances than network embedding methods (i.e., DeepWalk, metapath2vec); (2) among GNN-based models, RGCN and MEIRec are designed for heterogeneous graph, which could preserve richer semantic information and thus obtain better results than GCN and GAT; (3) our proposed $\alpha$-Satellite consistently outperforms all baselines in terms of precision, recall, ACC, F1 and AUC. The reason behind this is that, compared with RGCN and MEIRec, *$\alpha$-Satellite* leverages both of their advantages: it first explores meta-path schemes to retrieve neighbors of the nodes, and then considers different semantics of different types of relations for neighborhood information propagation and aggregation.

**Table 1: Comparisons with baselines.**

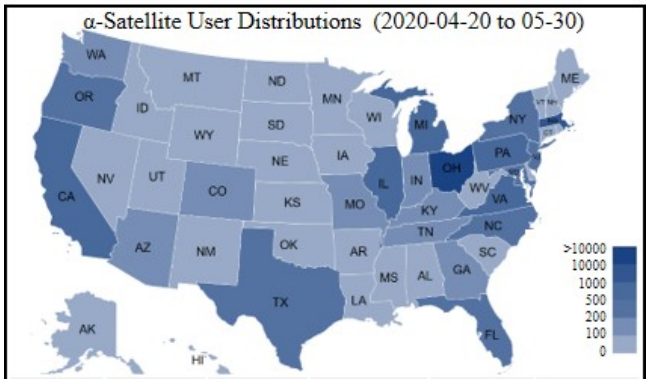
| Method | Precision | Recall | ACC | F1 | AUC |
|---|---|---|---|---|---|
| DeepWalk | 0.7120 | 0.9328 | 0.8572 | 0.8076 | 0.8840 |
| metapath2vec | 0.7364 | 0.9399 | 0.8714 | 0.8258 | 0.8952 |
| GCN | 0.7529 | 0.9441 | 0.8807 | 0.8377 | 0.9036 |
| GAT | 0.7526 | 0.9449 | 0.8808 | 0.8379 | 0.9044 |
| RGCN | 0.7830 | 0.9526 | 0.8974 | 0.8595 | 0.9146 |
| MEIRec | 0.7816 | 0.9530 | 0.8969 | 0.8588 | 0.9175 |
| ***$\alpha$-Satellite*** | **0.8380** | **0.9595** | **0.9239** | **0.8947** | **0.9378** |

## 4.5 Analysis of Large-scale Users

After we launched *$\alpha$-Satellite* to the public for beta test on April 20, ***it had attracted 51,190 users*** as of May 30. We have also received a lot of good feedback from users in terms of the ease of use and its utility for COVID-19 risk estimations, for examples:

- "*I am on the Executive Leadership team of a group of 225 dental practices across the United States. I live in Cleveland and saw your tool profiled in Crain's Cleveland Business. I would like to get access to your tool, as this could be a valuable tool for our clinicians.*"
- "*We'd love to test out the site and give some feedback. Thanks for putting together this tool. It's much needed and I hope will help curb transmission here in NEO.*"
- "*I read with great interest the fact that you released the risk calculator and I looked into it. I think it can be very beneficial for us, especially for clinical decisions, such as which procedures we should do, when to open the Dental School, etc. Thank you so much for doing this, it is great work.*"

The experimental results and user feedback both demonstrate the effectiveness of our system. In this study, based on Google Analytics platform and zip codes of user query POIs (i.e., all the data are anonymized and there are not privacy concerns or issues), we perform further analysis of the distribution of 47,946 users (93.66%) from United States who visited our system during April 20-May 31. Figure 6 illustrates the geo-distributions of the users, from which we have following observations: (1) The system has attracted the users across all the states in the country. (2) The state of OH has largest number of users (i.e., 38,636 users accounting for 80.58%), 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. Figure 7 shows the more severe regions with larger numbers of COVID-19 cases (both at state and county levels) the more *$\alpha$-Satellite* users. This 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 strategies for community mitigation.




| | | | | | |
|---|---|---|---|---|---|
| Users | 38,636 | 1,152 | 712 | 660 | 625 |
| State | OH | MA | MI | PA | IL |
| Cases | 35,514 | 96,301 | 56,884 | 75,697 | 118,909 |
| Users | 559 | 547 | 425 | 380 | 358 |
| State | CA | VA | TX | FL | NY |
| Cases | 109,983 | 43,611 | 62,880 | 55,421 | 374,706 |

**Figure 6: The geo-distributions of *$\alpha$-Satellite* users.**

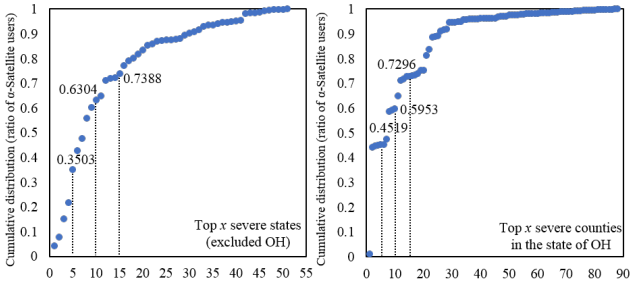



**Figure 7: $\alpha$-*Satellite* user vs. COVID-19 case distributions.**

## 5 CONCLUSION

To track the emerging dynamics of COVID-19 pandemic in the United States, leveraging the large-scale and real-time data generated from heterogeneous sources, we have developed a data-driven system (named $\alpha$-*Satellite*) to provide real-time COVID-19 risk assessment in a hierarchical manner to assist people with actionable information for community mitigation. To comprehensively evaluate the performance of $\alpha$-*Satellite* in real-time COVID-19 risk assessment, a set of studies are first performed to validate its utility; based on a real-world dataset consisting of 6,538 annotated POIs, the experimental results show that $\alpha$-*Satellite* achieves the area of under curve (AUC) of 0.9378, which outperforms state of the arts. After we launched the system for public tests on April 20, it had attracted 51,190 users as of May 30. Based on the analysis of its large-scale users, we have a key finding that people from more severe regions (i.e., with larger numbers of COVID-19 cases) have stronger interests using the system for actionable information. In the further work, we will continue our efforts to expand the data collection and enhance the system to help combat the pandemic. Our developed system and generated benchmark datasets have been made publicly accessible through our website.

## 6 ACKNOWLEDGMENT

This work is partially supported by the NSF under grants IIS-2027127, IIS-2040144, 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.

## REFERENCES

[1] 1point3acres. 2020. *COVID-19 in US and Canada.* https://coronavirus.1point3acres.com/en.
[2] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media".
[3] CDC. 2020. *1918 Pandemic (H1N1 virus).* https://www.cdc.gov/flu/pandemic-resources/1918-pandemic-h1n1.html.
[4] CDC. 2020. *Are You at Higher Risk for Severe Illness?* https://www.cdc.gov/coronavirus/2019-ncov/specific-groups/high-risk-complications.html.
[5] CDC. 2020. *How COVID-19 Spreads.* https://www.cdc.gov/coronavirus/2019-ncov/prepare/transmission.html.
[6] CDC. 2020. *Implementation of Mitigation Strategies for Communities with Local COVID-19 Transmission.* https://www.cdc.gov/coronavirus/2019-ncov/downloads/community-mitigation-strategy.pdf.
[7] Jun Chen, Lianlian Wu, Jun Zhang, Liang Zhang, Dexin Gong, Yilin Zhao, Shan Hu, Yonggui Wang, Xiao Hu, Biqing Zheng, et al. 2020. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study. *medRxiv* (2020).
[8] Victor Couture, Jonathan Dingel, Allison Green, Jessie Handbury, and Kevin Williams. 2020. *Exposure indices derived from PlaceIQ movement data.* https://github.com/COVIDExposureIndices/COVIDExposureIndices.
[9] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *KDD*. 135–144.
[10] Shaohua Fan, Junxiong Zhu, Xiaotian Han, Chuan Shi, Linmei Hu, Biyu Ma, and Yongliang Li. 2019. Metapath-guided Heterogeneous Graph Neural Network for Intent Recommendation. In *KDD*. 2478–2486.
[11] Yujie Fan, Shifu Hou, Yiming Zhang, Yanfang Ye, and Melih Abdulhayoglu. 2018. Gotcha-Sly Malware! Scorpion A Metagraph2vec Based Malware Detection System. In *KDD*. 253–262.
[12] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*. 363–370.
[13] Slav W Hermanowicz. 2020. Forecasting the Wuhan coronavirus (2019-nCoV) epidemics using a simple (simplistic) model. *medRxiv* (2020).
[14] Shifu Hou, Yujie Fan, Yiming Zhang, Yanfang Ye, Jingwei Lei, Wenqiang Wan, Jiabin Wang, Qi Xiong, and Fudong Shao. 2019. $\alpha$Cyber: Enhancing Robustness of Android Malware Detection System against Adversarial Attacks on Heterogeneous Graph based Model. In *CIKM*. 609–618.
[15] Shifu Hou, Yanfang Ye, Yangqiu Song, and Melih Abdulhayoglu. 2017. Hindroid: An intelligent android malware detection system based on structured heterogeneous information network. In *KDD*. 1507–1515.
[16] Kia Jahanbin and Vahid Rahmanian. 2020. Using twitter and web news mining to predict COVID-19 outbreak. *Medknow Publications* (2020).
[17] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
[18] Xiang Li, Yao Wu, Martin Ester, Ben Kao, Xin Wang, and Yudian Zheng. 2017. Semi-supervised clustering in attributed heterogeneous information networks. In *WWW*. 1621–1629.
[19] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, Vol. 30. 3.
[20] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *KDD*. 701–710.
[21] Gurjit S Randhawa, Maximillian PM Soltysiak, Hadi El Roz, Camila PE de Souza, Kathleen A Hill, and Lila Kari. 2020. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. *bioRxiv* (2020).
[22] Arni SR Srinivasa Rao and Jose A Vazquez. 2020. Identification of COVID-19 Can be Quicker through Artificial Intelligence framework using a Mobile Phone-Based Survey in the Populations when Cities/Towns Are Under Quarantine. *Infection Control & Hospital Epidemiology* (2020), 1–18.
[23] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *ESWC*. 593–607.
[24] Weiya Shi, Xueqing Peng, Tiefu Liu, Zenghui Cheng, Hongzhou Lu, Shuyi Yang, Jiulong Zhang, Feng Li, Mei Wang, Xinlei Zhang, et al. 2020. Deep Learning-Based Quantitative Computed Tomography Model in Predicting the Severity of COVID-19: A Retrospective Study in 196 Patients. (2020).
[25] StatsIndiana. 2020. *City-to-County Finder.* http://www.stats.indiana.edu/uspr/a/place_frame.html.
[26] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. 2011. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *VLDB Endowment* 4, 11 (2011), 992–1003.
[27] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).
[28] Shuai Wang, Bo Kang, Jinlu Ma, Xianjun Zeng, Mingming Xiao, Jia Guo, Mengjiao Cai, Jingyi Yang, Yaodong Li, Xiangfei Meng, et al. 2020. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). *medRxiv* (2020).
[29] WHO. 2020. *Coronavirus disease (COVID-19).* https://www.who.int/.
[30] Li Yan, Hai-Tao Zhang, Yang Xiao, Maolin Wang, Chuan Sun, Jing Liang, Shusheng Li, Mingyang Zhang, Yuqi Guo, Ying Xiao, et al. 2020. Prediction of survival for severe Covid-19 patients with three clinical features: development of a machine learning-based prognostic model with clinical data in Wuhan. *medRxiv* (2020).
[31] Yanfang Ye, Lingwei Chen, Shifu Hou, William Hardy, and Xin Li. 2017. DeepAM: A Heterogeneous Deep Learning Framework for Intelligent Malware Detection. *Knowledge and Information Systems* (2017), 1–21.
[32] Yanfang Ye, Shifu Hou, Lingwei Chen, Jingwei Lei, Wenqiang Wan, Jiabin Wang, Qi Xiong, and Fudong Shao. 2019. Out-of-sample Node Representation Learning for Heterogeneous Graph in Real-time Android Malware Detection.. In *IJCAI*. 4150–4156.
[33] Yanfang Ye, Shifu Hou, Lingwei Chen, Xin Li, Liang Zhao, Shouhuai Xu, Jiabin Wang, and Qi Xiong. 2018. ICSD: An automatic system for insecure code snippet detection in stack overflow over heterogeneous information network. In *ACSAC*. 542–552.
[34] Yanfang Ye, Tao Li, Donald Adjeroh, and S Sitharama Iyengar. 2017. A Survey on Malware Detection Using Data Mining Techniques. *Comput. Surveys* (2017).