Multiwave COVID-19 Prediction from Social Awareness using Web Search and Mobility Data

Jiawei Xue
xue120@purdue.edu
Purdue University
West Lafayette, IN, USA

Takahiro Yabe
tyabe@mit.edu
Massachusetts Institute of Technology
Cambridge, MA, USA

Kota Tsubouchi
ktsubouc@yahoo-corp.jp
Yahoo Japan Corporation
Tokyo, Japan

Jianzhu Ma*
majianzhu@pku.edu.cn
Peking University
Beijing, China

Satish Ukkusuri*
sukkusur@purdue.edu
Purdue University
West Lafayette, IN, USA

ABSTRACT

Recurring outbreaks of COVID-19 have posed enduring effects on global society, which calls for a predictor of pandemic waves using various data with early availability. Existing prediction models that forecast the first outbreak wave using mobility data may not be applicable to the multiwave prediction, because the evidence in the USA and Japan has shown that mobility patterns across different waves exhibit varying relationships with fluctuations in infection cases. Therefore, to predict the multiwave pandemic, we propose a Social Awareness-Based Graph Neural Network (SAB-GNN) that considers the decay of symptom-related web search frequency to capture the changes in public awareness across multiple waves. Our model combines GNN and LSTM to model the complex relationships among urban districts, inter-district mobility patterns, web search history, and future COVID-19 infections. We train our model to predict future pandemic outbreaks in the Tokyo area using its mobility and web search data from April 2020 to May 2021 across four pandemic waves collected by Yahoo Japan Corporation under strict privacy protection rules. Results demonstrate our model outperforms state-of-the-art baselines such as ST-GNN, MPNN, and GraphLSTM. Though our model is not computationally expensive (only 3 layers and 10 hidden neurons), the proposed model enables public agencies to anticipate and prepare for future pandemic outbreaks.

CCS CONCEPTS
• Information systems → Spatial-temporal systems; • Applied computing → Sociology.

KEYWORDS
COVID-19 Forecasting, Web Search Data, Human Mobility Data, Graph Neural Networks, Social Awareness Decay

ACM Reference Format:
Jiawei Xue, Takahiro Yabe, Kota Tsubouchi, Jianzhu Ma*, and Satish Ukkusuri*. 2022. Multiwave COVID-19 Prediction from Social Awareness using Web Search and Mobility Data. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD ’22), August 14–18, 2022, Washington, DC, USA. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3534678.3539172

1 INTRODUCTION

The spreading mechanism of COVID-19 is complicated due to its dependency on disease features and social factors such as human mobility [35, 51], public awareness [36], and intervention policies. One prominent phenomenon of the complex disease spreading process is the multiple outbreak wave, which implies the periodic rebound to a large number of infection cases [28] and is obvious in many countries such as the USA, UK, France, and Japan (Figure 1). Abrupt and uncertain disease outbreaks disturb individuals' daily life, government’s reopening policies [5], medical resources management [30], and risk assessment [57]. It has an enormous social impact to investigate and construct an accurate model to predict the multiple waves by fully utilizing different types of data [48].

Many prediction models for the first outbreak wave have been proposed to anticipate the infection and death cases [9, 27, 41, 43]. One critical input for these models is the mobility data [6, 19, 22], which describes population movements and is positively related to the disease infections [49]. Nevertheless, continuous tracking of human mobility dynamics shows that the mobility strength did not exhibit consistent relationships with infection cases: (1) the USA: human mobility fluctuated around 95% of the normal period from July 1 to Dec. 1, 2020 [10] which witnessed the second wave in July and the third wave in November (Figure 1); (2) Tokyo: the social contact index resumed to the normal level and decreased slightly after July 2020, but Tokyo experienced the second wave since then [52]. The research community has also noticed and discussed the limitation of mobility data in long-term multiwave infection prediction [2, 29]. The inconsistency in the mobility and infection necessitates other data that is more representative of disease waves.

During COVID-19, many text-based methods have been proposed to aid communities such as to understand human’s emotion states [21] and to answer peoples’ questions [54]. Web search records collected by the web service provider have extensive applications such as customer behavior analysis [14], disease outbreak monitoring [17], and evacuation prediction [53]. For COVID-19, symptom-related web search records (e.g., fever, cough, and
In this study, we propose a multiwave infection prediction approach \(^1\), with the direct application as urban district-level disease outbreak early warning. District-level disease prediction has the following three requirements: (1) comprehensive data sources, such as people movement and social responses, should be included to contain various hints that are closely related to the disease spreading; (2) spatial and temporal disease transmission patterns of COVID-19 should be taken into consideration; (3) it captures complex dependency between infection cases and other factors. To deal with these challenges, we first define the Web search Mobility Network (WMN) whose nodes and edges maintain web search frequency and inter-district human flow information, respectively. Afterward, we propose a Social Awareness-Based Graph Neural Network (SAB-GNN) architecture upon the WMN to capture the spatio-temporal infection case dynamics in different urban districts. We train and test the model using real-world infection, human mobility, and web search data in Tokyo from April 2020 to May 2021, and obtain better prediction performance than state-of-the-art baseline models. Our method has three contributions:

- We propose the SAB-GNN by fusing historical infection, mobility, and web search data that provide sufficient evidence of potential disease outbreaks. The spatial module, temporal module, and social awareness module take separate responsibilities and jointly contribute to the prediction.
- The proposed method is implemented on a mega-city, Tokyo, with a period spanning more than one year across four pandemic waves. We conduct a comprehensive analysis of disease outbreaks and prediction results from different models at different time intervals, which promotes a more nuanced understanding of the disease waves.

\(^1\)The code and data of this study is open to public and can be found at https://github.com/JiaweiXue/MultiwaveCovidPrediction.

2 RELATED WORK

2.1 Time Series Models

Existing COVID-19 time series models cover multiple types such as auto-regressive integrated moving average (ARIMA) [27], and long short-term memory (LSTM) [9, 20]. Moreover, biologists and engineering scientists focus on the relationship between fatality rate and biochemical indicators [60], human mobility [20]. When it comes to urban district-level disease infection prediction, while the inter-district connections provide crucial pathways for both human movement and disease dissemination, these models are insufficient to capture such spatial dependency between different urban districts. In fact, the spatial dependency information enables us to deal with the data scarcity issue which may occur during the pandemic season. For instance, assume that we have sufficient mobility and social media data for district \(A\) and deficient data for district \(B\), and recognize strong mobility connections between districts \(A\) and \(B\). Considering the connections between the two districts helps to fully utilize the infection information and predict the infection cases for both districts \(A\) and \(B\).

2.2 Graph Neural Networks

Graph neural network (GNN) is an innovative neural network that captures the relationship between multi-hop neighborhood nodes via the message passing mechanism [61]. In the past years, various GNN models such as graph convolutional network (GCN) [26], GraphSage [16], graph attention network (GAT) [45] were developed and applied to many fields such as neural machine translation [3], visual question answering [31], traffic prediction [7, 34, 55, 59], and network metric generation [50].

Researchers have harnessed the superiority of the GNN in modeling spatial dependency to perform the disease infection case prediction. Most published GNN approaches [13, 23, 33] focused on the infection prediction before July 2020 when the first global outbreak occurred using the historical infection and mobility data. The mobility data was sufficient to reflect the disease spreading patterns during the first wave thanks to its simple relationship with the infection. First, areas attracting more passengers had higher risks of experiencing rapidly increasing cases than some lonely areas at the beginning of the first wave. Second, the travel restriction policies during the first wave suppressed the mobility strength and thus decelerated the disease outbreak [49]. Nevertheless, the interaction between infection and mobility evolved into a much more complicated status during the later waves because the infection cases were affected by a large variety of causes.
such as mask policy, the vaccination, which hindered the ability of raw mobility data to reflect the infection tendency. Besides, many studies have recognized that the available mobility data for COVID-19 infection cases prediction was limited by the period length [33, 38], which resulted in unstable learned models. In summary, the long-term multiwave infection prediction requests alternative data sources that provide sufficient interconnections with the infection case under the dynamic environment. In this study, we turn to the novel web search data, which directly reveals human’s awareness to the disease and potential symptoms [58], to perform the multiwave disease prediction.

3 PRELIMINARIES

We first describe the mobility and web search data used as features in the prediction, and then formally define our prediction task. Assume an urban area is divided into $n$ pre-defined urban districts.

Mobility feature: for mobile phone users within the urban area, we collect location records (longitude, latitude, and time) with the temporal resolution as around 30 minutes and spatial error as at most 100 meters. We extract each individual’s trajectory points as a sequence. Next, we project the sequence points of each mobile phone user into the urban districts to obtain the number of daily inter-district trips. Note that inter-district trips build natural pathways to transmit disease viruses in the city so that these districts have correlated exposure to disease outbreak risks.

Web search feature: we scope sources that provide sufficient interconnections with the infection cases prediction was limited by the period length [33, 38], which resulted in unstable learned models. Lastly, the existing study finds that the public’s symptom-related web search frequency decreases from the first wave to the second wave, which reveals a prevalent social phenomenon that people’s awareness of a hot topic gradually declines (which is referred to as social awareness decay in this study) [52]. Based on the fact that people adapt themselves to the mask policies, travel restrictions, and routine testings and pay less attention to the COVID-19, we propose a social awareness recovery module to estimate the actual occurrence of COVID-19-related symptoms. In summary, we build an integrated future infection case prediction model with three modules (i.e., spatial module, awareness recovery module, and temporal module) by fusing the historical infection, mobility, and web search data.

4 SOCIAL AWARENESS-BASED GRAPH NEURAL NETWORKS

In this section, we first present the intuition of the SAB-GNN model, and then sequentially introduce its three modules: the spatial information propagation module, the social awareness recovery module, and the temporal information passing module, and finally declare the loss function.

We deduce that the future reported district-level infection cases are jointly influenced by the existing infection cases, the number of susceptible individuals that may have been infected (which can be mined from public’s symptom-related web search data), and in-person contact patterns across the city (which is reflected by the inter-district trip numbers) and use them as features. Given that the pandemic spreading is indeed a complicated temporal process embedding on the space, we propose to build temporal and spatial modules [56] to track the infection dynamics. Lastly, the existing study finds that the public’s symptom-related web search frequency decreases from the first wave to the second wave, which reveals a prevalent social phenomenon that people’s awareness of a hot topic gradually declines (which is referred to as social awareness decay in this study) [52]. Based on the fact that people adapt themselves to the mask policies, travel restrictions, and routine testings and pay less attention to the COVID-19, we propose a social awareness recovery module in the SAB-GNN to estimate the actual occurrence of COVID-19-related symptoms. In summary, we build an integrated future infection case prediction model with three modules (i.e., spatial module, awareness recovery module, and temporal module) by fusing the historical infection, mobility, and web search data.

4.1 Spatial Module: Graph Neural Networks

Recall that the web search frequency vector $h_i^0$ reflects the number of potential infected individuals in the urban district $d_i$ and $E_t$ records the daily inter-district trips. We therefore perform the convolution operation on $h_i^0$ using the $E_t$ information under the graph neural network framework to capture the disease risk propagation properties. As shown in Figure 2, using the symptom-related web search frequency in each urban district, we employ the one-hot encoding to initialize the representation for each urban district (i.e., each node in $G_t$) as the input matrix: $X_t^{(0)} = H_t$. Following the GCN model [26], we define the node representation propagation rule between the layers $k$ and $(k + 1)$ as:

$$X_t^{(k+1)} = \sigma(\tilde{D}_t^{-1/2} \tilde{E}_t \tilde{D}_t^{-1/2} X_t^{(k)} W^{(k)}),$$

(1)

where

$$\tilde{E}_t = E_t + I_{n \times n}, \tilde{D}_{ii} = \sum_{j=1}^{n} \tilde{E}_{ij}.$$ 

(2)

and $W^{(k)}$ is a learnable weight matrix, $I_{n \times n}$ is the $n$ by $n$ identity matrix, $\sigma(\cdot)$ is the activation function ReLU.

Note that we normalize the matrix $E_t$ such that the sum of each column is equal to 1 (i.e., the sum of incoming edges on one node is 1), which is used in the existing study [33]. In practice, it is possible to replace the spectral convolution with other GNN variants such as GAT [45], GraphSage [16]. We also implement them and find quite approximate prediction performance as the GCN. The outcome of the spatial module is a matrix $H_t^S = X_t^{(K)} = [x_t^{v_1}, x_t^{v_2}, ..., x_t^{v_n}]^T$, with $n$ rows that encode the web search frequency and mobility where $K$ is the number of layers.

4.2 Social Awareness Recovery Module

The symptom-related web search frequency is positively related to the number of actual symptom occurrences among the population [52]. As mentioned earlier, the social awareness decay effect
where \( \lambda \) varying social awareness rates, we specify district-dependent \( v \) in to recover the social awareness. Note that we introduce the square social awareness recovery module transforms \( H \) to \( \lambda \) in to encode its unique awareness decay behavior. A large value of type, demographic characteristic discrepancy may lead to spatially served web search representation by a monotonically increasing recover the social awareness: \( r \) define an increasing function with respect to the time. The actual symptom occurrences, we propose to multiply the ob-
ally decreases with time after the first COVID-19 wave. To estimate in forms that probability of symptom-related word searching gradu-
ally decreases with time after the first COVID-19 wave. To estimate the actual symptom occurrences, we propose to multiply the ob-
served web search representation by a monotonically increasing function with respect to the time.

Specifically, we first linearly normalize each entry of the web search record vector \( h_i^t \) to 0 and 1 by the maximal and minimal web search frequency of each word across all days in the urban district \( v_i \), and feed them into the spatial module, and obtain \( H_i^2 \). Next, we define an increasing function \( r(t|i, t_0) = e^{\lambda_i^2(t-t_0)} \) regarding \( t \) to recover the social awareness:

\[
\tilde{x}_{i}^{2n} = x_{i}^{2n} r(t|i, t_0) = x_{i}^{2n} e^{\lambda_i^2(t-t_0)},
\]

where \( \lambda_i^2 \) is learnable and measures the social awareness recovery rate in \( v_i \). \( t, t_0 \) represent the current day and the first day of the study period, respectively. Given that the land use, economy type, demographic characteristic discrepancy may lead to spatially varying social awareness rates, we specify district-dependent \( \lambda_i^2 \) to encode its unique awareness decay behavior. A large value of \( \lambda_i^2 \) implies that social awareness of COVID-19 for residents living in \( v_i \) declines rapidly and we therefore adopt this large value to recover the social awareness. Note that we introduce the square in \( \lambda_i^2 \) to ensure that it is non-negative. Collectively speaking, the social awareness recovery module transforms \( H_i^2 \) to \( H_i^3 \) by

\[
H_i^3 = H_i^2 \circ M_{t,t_0},
\]

where \( M_{t,t_0} \) is the awareness recovery matrix (ARM):

\[
M_{t,t_0} = \begin{bmatrix}
    e^{\lambda_{11}(t-t_0)} & e^{\lambda_{12}(t-t_0)} & \cdots & e^{\lambda_{1n}(t-t_0)} \\
    e^{\lambda_{21}(t-t_0)} & e^{\lambda_{22}(t-t_0)} & \cdots & e^{\lambda_{2n}(t-t_0)} \\
    \vdots & \vdots & \ddots & \vdots \\
    e^{\lambda_{n1}(t-t_0)} & e^{\lambda_{n2}(t-t_0)} & \cdots & e^{\lambda_{nn}(t-t_0)}
\end{bmatrix},
\]

and \( \circ \) is the Hadamard product. This transform implies that the entries in \( H_i^3 \) are amplified to capture the actual disease risk which is underestimated due to the social awareness decay effect when \( t \) is large. In the end, we perform 0-1 normalization on the infection matrix \( I_t \) to obtain \( \tilde{I}_t \), and arrive at the output of the social awareness recovery module by concatenating the web search and infection representations, which is

\[
H_i^4 = [H_i^3, \tilde{I}_t].
\]

### 4.3 Temporal Module: LSTM

To capture the temporal dependency of district-level features and infection cases, we adopt the existing LSTM model [18]. For \( i \in \{1, 2, \ldots, n\} \), we extract the \( i \)-th row of matrices \( H_i^4, t \in [T-D_1+1, T] \) and feed them into an LSTM sequence (Figure 2). Since we have already modelled the spatial dependency in the spatial module, in the temporal module we pass the node representations from \( H_i^4 \) separately for different nodes, and these LSTM sequences share the identical structures and parameters.

---

**Figure 2:** The framework of SAB-GNN. We first model urban districts as nodes, and propagate the web search frequency embedding using the mobility information for the last \( D_1 \) days (i.e., Spatial Module). Next, we use a learnable social awareness matrix to recover node representations (i.e., Social Awareness Recovery Module), and feed them into an LSTM sequence to predict the next \( D_2 \) days’ infection cases (i.e., Temporal Module).
We utilize four datasets: infection cases, mobile phone location data, web search data, and symptom data in Tokyo from Jan. 6, 2020, to May 15, 2021. The mobility and web search data were owned by Yahoo Japan Corporation with strict privacy protection regulations. The number of inter-district trips on three days.

5.1 Mobility, Web Search, and Symptom Data

Recall that our objective is to perform the infection case prediction for the next $D_2$ days, we define the loss function as the mean squared error, which is:

$$L_T = \frac{1}{D_2 n} \sum_{t=T}^{T+D_2} \sum_{i=1}^{n} (l_{t,i} - \hat{l}_{t,i})^2,$$

where $l_{t,i}$ and $\hat{l}_{t,i}$ denote the actual and predicted infection cases for the district $i$ on the day $t$. We present the training process as Algorithm in Supplement.

5.2 Data Preprocessing

Note that the raw mobility and web search data are at the individual level. To prepare the features $E_t$, $H_t$ used in the machine-learning framework, we conduct the data preprocessing to aggregate the raw mobility and web search data to the urban district level. Please find the complete description of data prepossessing in the Appendix.

The statistics of aggregate daily new infection cases, web search frequency, and inter-district trip number are shown in Figure 3. The vertical dash lines in Figures 3ac mark March 1, 2020, May 1, 2020, and January 1, 2021, whose web search distribution and mobility flow are visualized in Figures 3bd, respectively. We find from Figure 3a that the dynamic of symptom-related web search exhibits consistent peaks with daily infection cases, especially near April 2020, January 2021, and May 2021, which provides the direct evidence to inspire us to utilize the web search data in the multiwave pandemic prediction.

6 RESULTS

6.1 Setting, Evaluation Metrics, Baselines

We conduct experiments to predict two outbreak waves, i.e., the third wave (from Dec. 10, 2020 to Feb. 7, 2021) and the fourth wave (from March 17 to May 15, 2021). The period for each experiment covers 10 months of observations where the train/validation/test ratio is 70%/10%/20%. We use the first 8 months (months 1 to 8) as training and validation data, and the last 2 months (months 9 and 10) as testing data. Note that the validation data has the size of 1 month and is evenly distributed from months 7 to 8. Since most existing disease prediction studies [33, 38, 47] predict the infection at the week-level, we design three scenarios:...
We present the prediction performance of the SAB-GNN and other models (9) GraphLSTM: a prediction framework integrating GraphSage and LSTM [42]; (8) MPNN+LSTM: a message passing neural network using two separate LSTMs [11]; (7) ST-GNN: an Spatio-Temporal Seq2seq: encode the input infection case and decode the sequence to connect with the future infection case. Finally, we observe from the LSTM and Seq2seq that simply concatenating the web search frequency embedding with the past infection cases is unable to consistently boost the prediction, which affirms the necessity of designing a suitable model architecture to utilize the web search data. For each model in Table 1, the running time for each scenario is within 20 minutes on the Ubuntu system with 32 GB RAM and 3.3 GHz Xeon w-2155 CPU.

### 6.3 Prediction Results for Urban Districts

We visualize the prediction errors and compare the predicted cases with the actual cases for each urban district in Figures 4, 5. As shown in Figures 4ab, the relative prediction errors for most districts are homogeneously lower than 0.50 except for the central district (i.e., Chiyoda). This reveals that SAB-GNN mines the relationships between features and future infection cases in a global manner thanks to the message passing mechanism among neighborhood nodes. The reason for the relatively large prediction error in Chiyoda (i.e., District 1) is because Chiyoda is where Imperial Palace locates and has quite a few real infection cases (Figure 5, the top-left panel). Figure 4c displays that the predicted daily case curve is able to capture the increasing tendency of actual cases from day 240 to day 280 and also maintains a small prediction error, which further demonstrates the power of our proposed SAB-GNN.

Figure 5 exhibits that for most urban districts, the model is able to anticipate the general increasing tendency of infection cases during the fourth wave. This information is especially valuable for local public agencies to make timely preparations at the beginning of a wave. Note that the prediction of District 4 (i.e., Shinjuku) is not as accurate as other districts. One potential interpretation is that Shinjuku is a commercial area with many entertainment industries and thus does not have similar infection patterns as other districts.

### 6.4 Parameter Sensitivity

The influences of model parameters in the SAB-GNN on the prediction performance are shown in Figure 6. Figure 6a displays the changes of RMSE for the fourth peak prediction with respect to \( n_w \) (i.e., numbers of symptom-related web search features used in the model). Recall that there is a huge frequency discrepancy between different symptoms and we only feed the most frequent \( n_w \) symptoms into our model. The result suggests that \( n_w = 8 \) provides the best prediction performance. The underlying reason is that when \( n_w \) is small, more symptoms contribute stronger connections with future infection cases; when \( n_w \) is large, many weakly-related symptoms bring noises to the training and thus harm the prediction performance.

Next, we tune the numbers of layers \( L_1, L_2 \) in the GNN and LSTM modules (Figures 6bc) and conclude that \( (L_1, L_2) = (1, 2) \) yields the best prediction. It is reasonable that the simple model with a small number of parameters is preferred in this infection case prediction task whose training data is at the hundreds level [37] (each day is associated with one training sample). Finally, we test the model in Figure 6d with varying \( D_1 \) and \( D_2 \) and observe the overall tendency.
Table 1: Performance evaluation using past $D_1$ days’ features to predict next $D_2$ days’ infection ($I$: infection data; $W$: web search data; $M$: inter-district mobility data; SAB-GNN-wsa: the SAB-GNN model without the social awareness recovery).

| Model          | Features | $D_1$ | $D_2$ | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE |
|----------------|----------|-------|-------|-----|------|-----|------|-----|------|-----|------|-----|------|-----|------|
| HA (all)       | I        | 22.63 | 26.45 | 22.53 | 27.22 | 22.16 | 27.65 | 8.76 | 10.04 | 9.14 | 10.59 | 9.41 | 11.08 |
| HA (X days)    | I        | 14.15 | 17.02 | 16.92 | 20.72 | 19.14 | 23.81 | 4.52 | 5.76 | 5.52 | 7.02 | 6.47 | 8.23 |
| LSTM           | I        | 19.67 | 24.07 | 17.23 | 21.93 | 20.61 | 26.61 | 5.94 | 7.67 | 8.20 | 10.64 | 8.32 | 10.95 |
| LSTM           | I/W      | 20.61 | 24.89 | 20.72 | 25.91 | 19.93 | 25.94 | 7.43 | 9.79 | 8.28 | 10.55 | 8.11 | 10.63 |
| Seq2seq        | I        | 17.27 | 22.24 | 23.04 | 28.73 | 16.79 | 22.77 | 6.37 | 8.38 | 8.84 | 11.42 | 8.45 | 11.08 |
| Seq2seq        | I/W      | 15.69 | 19.83 | 21.76 | 27.46 | 19.66 | 25.69 | 5.96 | 7.74 | 7.93 | 9.86 | 10.43 | 12.94 |
| ST-GNN [23]    | I/W/M    | 20.20 | 25.86 | 21.53 | 27.32 | 20.24 | 26.36 | 5.46 | 6.95 | 7.63 | 9.97 | 10.54 | 13.96 |
| MPNN+LSTM [33] | I/W/M    | 12.40 | 17.31 | 16.30 | 21.73 | 19.78 | 25.59 | 3.34 | 4.49 | 5.27 | 7.97 | 4.80 | 6.56 |
| GraphLSTM [42] | I/W/M    | 10.22 | 12.82 | 13.27 | 16.91 | 15.56 | 20.07 | 3.44 | 4.61 | 4.38 | 5.88 | 5.32 | 7.13 |
| SAB-GNN-wsa    | I/W/M    | 10.75 | 13.43 | 12.72 | 16.33 | 15.61 | 20.10 | 3.32 | 4.46 | 4.63 | 6.18 | 5.03 | 6.77 |
| SAB-GNN        | I/W/M    | 8.03  | 10.43 | 11.23 | 14.78 | 13.76 | 18.24 | 3.25 | 4.24 | 4.28 | 5.57 | 5.25 | 6.82 |

6.5 Ablation Study

To confirm the effect of each module in the SAB-GNN on prediction results, we perform the ablation experiments for both the third wave and fourth wave (Figure 7). We remove one of the three modules in the SAB-GNN, and perform the prediction for the same temporal horizons as the full SAB-GNN model. We show quantitatively that the SAB-GNN model obtains lower prediction RMSE than other incomplete models, especially for the third wave, which reveals that both the spatial module, temporal module, and social awareness module contribute to the final prediction results in a positive manner. While existing studies have paid much attention to the spatial and temporal relationships between features and infection cases, we recommend introducing suitable social knowledge into the prediction, given the positive effect of the social awareness recovery mechanism.

7 DISCUSSION

Model performance analysis: Underlying reasons for the superior prediction of SAB-GNN are threefold: (1) SAB-GNN leverages the power of GNN and LSTM to capture the spatio-temporal dynamic of disease spreading; (2) Web search features provide extra information as unconfirmed symptoms; (3) Social memory decay module properly reflects the social awareness decay as people get accustomed to COVID-19 during the pandemic. Note that we do not quantitatively test epidemiological models (e.g., SIR, SEIR) for two reasons: (1) Vanilla epidemiological models provide a good fit to the single wave or the first wave, but are insufficient to describe the multiwave infection outbreaks; (2) Quantitative experiments have
8 CONCLUSION

Motivated by the multiwave outbreak of COVID-19 across the globe, we establish the SAB-GNN to predict future infection cases. Except for the historical infection and mobility data, our approach utilizes the novel symptom-related web search data which provides alternative evidence of future waves. More importantly, we consider the social awareness decay effect and propose the social awareness recovery module to estimate the actual infection risks. Experiments on the third and fourth peaks of Tokyo affirm that the SAB-GNN outperforms other baseline models and captures the increasing trend of pandemic waves. Our method is applicable to many countries given the wide coverage of web search data.

ACKNOWLEDGEMENTS

We thank Yahoo Japan Corporation for the collaboration. Jiawei Xue and Satish Ukkusuri are partly supported by the National Science Foundation (Award Number: 1638311) grant for which the authors are grateful.

REFERENCES

[1] Aniruddha Adiga, Lijing Wang, Benjamin Hurt, Akkil Peddiredy, Przemyslaw Poerbski, Srirangan Venkataramanan, Bryan Leroy Lewis, and Madhav Marathe. 2021. All models are useful: Bayesian ensembling for robust high resolution covid-19 forecasting. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2505–2513.
[2] Hamada S Badr and Lauren M Gardner. 2021. Limitations of using mobile phone data to model COVID-19 transmission in the USA. The Lancet Infectious Diseases 21, 5 (2021), e113.
[3] Joost Bastings, Ivan Títov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima’an. 2017. Graph Convolutional Encoders for Syntax-aware Neural Machine Translation. In EMNLP.
[4] Centers for Disease Control and Prevention. 2021. Symptoms of COVID-19. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html.
[5] Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. 2020. Mobility network models of COVID-19 explain inequities and inform reopening. Nature (2020), 1–6.
[6] Serina Chang, Mandy L Wilson, Bryan Lewis, Zakaria Mehrab, Komal K Dudakiya, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, et al. 2021. Supporting covid-19 policy response with large-scale mobility-based modeling. In SIGKDD. 2632–2642.
[7] Chen Chen, Kenli Li, Sin G Teo. Xiaodong Zou, Kang Wang, Jie Wang, and Zeng Zeng. 2019. Gated residual recurrent graph neural networks for traffic prediction. In Proceedings of the AAAI conference on artificial intelligence. Vol. 33. 485–492.
[8] Yizong Cheng. 1995. Mean shift, mode seeking, and clustering. IEEE transactions on pattern analysis and machine intelligence 17, 8 (1995), 790–799.
[9] Vinay Kumar Reddy Chimmula and Lei Zhang. 2020. Time series forecasting patterns across the United States during the COVID-19 outbreak. Available online at https://covid19.gleamproject.org/mobility.
[10] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014).
[11] Estee Y Cramer et al. 2021. The United States COVID-19 Forecast Hub dataset. medRxiv (2021). https://doi.org/10.1101/2021.11.04.21265886.
[12] Junyi Gao, Rakshit Sharma, Cheng Qian, Lucas M Glass, Jeffrey Spudener, Justin Romberg, Jieping Sun, and Cao Xiao. 2020. STAN: spatio-temporal attention network for pandemic prediction. In Proceedings of the AAAI conference on artificial intelligence. Vol. 34. 485–492.
[13] Sharad Goel, Jake M Hofman, Sébastien Lahaie, David M Pennock, and Duncan J Watts. 2010. Predicting consumer behavior with Web search. Proceedings of the National academy of sciences 107, 41 (2010), 17486–17490.
[14] Google Inc. 2021. Covid-19 infection cases. https://news.google.com/covid19/map?hl=en-US&gl=US&ceid=US%3Aen.
[15] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In Advances in neural information processing systems. 1024–1034.
Multiwave COVID-19 Prediction from Social Awareness using Web Search and Mobility Data

KDD ’22, August 14–18, 2022, Washington, DC, USA

[17] Shohret Risada, Taichi Murayama, Kota Tsoubouchi, Sumio Fujita, Shuntaro Yada, Shoko Wakahami, and Eji Arakami. 2020. Surveillance of early stage COVID-19 clusters using search query logs and mobile device-based location information. Scientific Reports 10, (2020), 1–8.

[18] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural computation 9, 8 (1997), 1735–1780.

[19] Xiao Huang, Zhengdong Li, Yuqin Jiang, Xiaoming Li, and Dwayne Porter. 2020. Twitter reveals human mobility dynamics during the COVID-19 pandemic. PloS one 15, 11 (2020), e0241957.

[20] HyongChon Ju, Juyeon Kim, Tru-Chen Huang, and Yu-Li Ni. 2020. condLSTM-Q: A novel deep learning model for predicting COVID-19 mortality in fine geographical Scale. arXiv preprint arXiv:2011.11907 (2020).

[21] Mingxuan Ju, Wei Song, Shiyu Sun, Yanfei Ye, Yuefei Fan, Shifu Hou, Kenneth Loparo, and Lian Zhao. 2021. Dr. Emotion: Disentangled Representation Learning for Emotion Analysis on Social Media to Improve Community Resilience in the COVID-19 Era and Beyond. In Proceedings of the Web Conference 2021. 518–528.

[22] Yuhao Kang, Song Gao, Yunlei Liang, Mingxiao Li, Jimmeng Rao, and Jake Kruse. 2020. Multiscale dynamic human mobility flow dataset in the US during the COVID-19 epidemic. Scientific data 7, 1 (2020), 1–13.

[23] Amol Kapoor, Xue Ben, Luyang Liu, Bryan Perozzi, Matt Barnes, Martin Blais, and Alexander Rodriguez, Nikhil Muralidhar, Bijaya Adhikari, Anika Tabassum, George Panagopoulos, Giannis Nikolentzos, and Michalis Vazirgiannis. 2021. World Health Organization. accessed June 2022. Coronavirus disease (COVID-19) pandemic. Available online at https://covid19.who.int/table.

[24] Roman Levin, Dennis L Chao, Edward A Wenger, and Joshua L Proctor. 2021. Taming the Box: Reasoning with Graph Convolutional Networks for Factual Visual Question Answering. Advances in Neural Information Processing Systems 34, 7, 1 (2021), 1382–1393.

[25] Xiao Huang, Zhenlong Li, Yuqin Jiang, Xiaoming Li, and Dwayne Porter. 2020. Spatial temporal incidence dynamic graph neural networks for traffic flow forecasting. IEEE Journal of Biomedical and Health Informatics 140 (2020), 110212.

[26] Nathan Sesti, Juan Jose Garau-Luis, Edward Crawley, and Bruce Cameron. 2021. Predicting COVID-19 Spread from Large-Scale Mobility Data. In KDD 2022.

[27] Amray Schwabe, Joel Persson, and Stefan Feuerriegel. 2021. Predicting COVID-19 Spreads and DISSem. In IJCAI 2021.

[28] O. Kara, T. Torlak, Y. Can, and M. F. Yildirim. 2020. Predicting COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. Chaos, Solitons & Fractals 120 (2020), 119212.

[29] Amol Kapoor, Xue Ben, Luyang Liu, Bryan Perozzi, Matt Barnes, Martin Blais, and Alexander Rodriguez, Nikhil Muralidhar, Bijaya Adhikari, Anika Tabassum, George Panagopoulos, Giannis Nikolentzos, and Michalis Vazirgiannis. 2021. Predicting Evacuation Decisions using Representations of Individuals’ Social Neighbors Ahead of COVID-19 Outbreak. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35. 4892–4900.

[30] Chenfeng Xiong, Songhua Hu, Mofeng Yang, Weiyu Luo, and Lei Zhang. 2020. Mobile device data reveals the dynamics in a positive relationship between human mobility and COVID-19 infections. Proceedings of the National Academy of Sciences 117, 44 (2020), 27087–27089.

[31] Jiawei Xue, Nan Jiang, Senwei Liang, Qiyuan Pang, Takahiro Yabe, Satoshi Ukkusuri, and Jianhua Ma. 2020. Quantifying the spatial homogeneity of urban road networks via graph neural networks. Nature Machine Intelligence 4, 3 (2022), 246–257.

[32] Takahiro Yabe, Kota Tsoubouchi, Naoya Fujiwara, Takayuki Wada, Yoshihide Sekimoto, and Satoshi Ukkusuri. 2020. Non-compulsory measures sufficiently reduced human mobility in Tokyo during the COVID-19 epidemic. Scientific reports 10, 1 (2020), 1–9.

[33] Takahiro Yabe, Kota Tsoubouchi, Yoshihide Sekimoto, and Satoshi Ukkusuri. 2021. Early warning of COVID-19 hotspots using human mobility and web search query data. Computers, Environment and Urban Systems (2021), 101747.

[34] Takahiro Yabe, Kota Tsoubouchi, Toru Shimizu, Yoshihide Sekimoto, and Satoshi Ukkusuri. 2019. Predicting Evacuation Decisions using Representations of Individuals’ Pre-Disaster Web Search Behavior. In SIGKDD ’19: Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Vol. 28. 2707–2717.

[35] Hai Xue, Fei Li, Jiaxin Su, Suyao Mu, Zhibiao Kang, Hailei Zhang, Yuekun Wang, and Tiejun Hu. 2022. Deep multi-view spatial-temporal network for taxi demand prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36. 5668–5675.

[36] Haijun Yao, Xiongfang Tang, Huo Wei, Guanjie Zheng, and Zhenhui Li. 2019. Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction. In Proceedings of the AAAI conference on artificial intelligence, Vol. 33. 5648–5657.

[37] Haijun Yao, Fei Wu, Jiantao Ke, Xianfeng Tang, Yitian Jia, Suyao Mu, Zhibiao Kang, Jieping Ye, and Zhenhui Li. 2018. Deep multi-view spatial-temporal network for taxi demand prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32.

[38] Yanfei Ye, Shifu Hou, Yuefei Fan, Yiming Zhang, Yiqian Sun, Shiyu Sun, Qian Peng, Mingxuan Ju, Wei Song, and Kenneth Loparo. 2020. alpha-Satellite: An AI-Driven System and Benchmark Datasets for Dynamic COVID-19 Risk Assessment in the United States. IEEE Journal of Biomedical and Health Informatics 24, 10 (2020), 2755–2764.

[39] Elad Yom-Tov, Vasileios Lampous, Thomas Inns, Ingrid J Cox, and Michael Edelstein. 2022. Providing early indication of regional anomalies in COVID-19 case counts in England using search engine queries. Scientific reports 12, 1 (2020), 1–10.

[40] Xiuye Zhang, Sao Huang, Yong Xu, Lianghao Xia, Peng Dai, Liefeng Bo, Junbo Zhang, and Yu Zheng. 2021. Traffic Flow Forecasting with Spatial-Temporal Graph Diffusion Network. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35. 15008–15015.

[41] Feng Zhou, Tao Chen, and Baisong Lei. 2020. Do not forget interaction: Predicting fatality of COVID-19 patients using logistic regression. arXiv preprint arXiv:2006.19442 (2020).

[42] Jie Zhou, Ganqu Cui, Shengding Hu, Zeyuan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. AI Open 1 (2020), 57–81.
A SUPPLEMENT: REPRODUCIBILITY

A.1 Dataset

| Location data | Web search data             |
|---------------|-----------------------------|
| ID1           | thissifakeid                |
| Latitude      | 35.683                      |
| Longitude     | 139.763                     |
| Unix time     | 1584390800                  |
| Date          | 20200317                    |

Table 2: Samples of mobility data and web search data.

| Symptom         | Symptom          |
|-----------------|------------------|
| Abdominal pain  | Hot flash        |
| Ageusia         | Hyperhidrosis    |
| Anosmia         | Insomnia         |
| Anxiety         | Lethargic        |
| Arthralgia      | Loss of appetite |
| Body ache       | Myalgia          |
| Chest pain      | Nasal dryness    |
| Chest tightness | Nausea           |
| Chills          | Oropharyngeal pain|
| Confusion       | Pain             |
| Cough           | Palpititation    |
| Dehydration     | Pyrexia          |
| Diarrhea        | Rash             |
| Disorientation  | Sinusitis        |
| Dizziness       | Rhinorrhea       |
| Dyspnea         | Sleep disturbance|
| Ear infection   | Sneezing         |
| Ear pain        | Stress           |
| Eye infection   | Sweating         |
| Eye pain        | Upper Respiratory Tract Infection |
| Fatigue         | Vomiting         |
| Headache        | Wheezing         |

Table 3: Specified COVID-19 related symptoms.

This study uses four datasets: [1] mobility data; [2] web search data; [3] infection data; [4] symptom data.

- We aggregate mobility data and web search data for individuals from the mobile phone data owned by Yahoo Japan Corporation\(^2\), and feed aggregated mobility (i.e., \(E_t\)) and web search data (i.e., \(H_t\)) into our machine learning model. \([1]\) \(E_t\) and \([2]\) \(H_t\) are accessible at Yahoo! JAPAN R&D\(^3\). Please go to YJ Covid-19 Prediction Data.
- [3] infection data and [4] symptom data originate from public resources and can be found under the data-collection directory in our Github repository\(^4\).
- Machine learning codes and implementation results are under the SAB-GNN directory in our Github repository.

Using these data and codes, readers can fully reproduce the disease prediction results for Tokyo. Besides, readers may use our codes to conduct the disease infection prediction on other cities if their mobility and web search data are available.

\(^2\)https://about.yahoo.co.jp/en/info/company/
\(^3\)https://randd.yahoo.co.jp/en/softwaredata
\(^4\)https://github.com/JiaweiXue/MultiwaveCovidPrediction

A.2 Data Preprocessing

Recall that our prediction task requires the standard mobility matrix \(E_t\) as well as the web search matrix \(H_t\) on the day \(t\). We design a data preprocessing framework to convert the raw mobility and web search data to \(E_t\) and \(H_t\) (Figure 8). We notice from the raw mobility data that a proportion of user IDs only appear a few times, which indicates that they might be contemporary travelers in Tokyo and may not be strongly related to the inter-city disease spreading in several months, and thus exclude their mobility data in this study. We finally identify 551,745 permanent users in the 23 special wards of Tokyo, which take around 6% of the total population in these wards. Our mobility and web search data preprocessing procedure is shown as Figure 8:

- **Identify Tokyo residents.** We apply the Mean Shift Algorithm [8] to each permanent users’ location points during the night hours (from 6:00 PM to 9:00 AM) for 2 weeks starting from Jan. 6, 2020 to estimate the longitude and latitude of their homes using Java (step (1)).
- **Generate mobility feature.** We extract the daily trajectories of these permanent residents using Java (step (2)) and then project the location points along the trajectories to the urban districts using the spatial-join function in geopandas package in Python and obtain the mobility matrix \(E_t\) (step (3)). Note that we mine the daily location trajectory of each user ID and identify cross-district trips with the duration of at least 10 minutes (these trips are referred to as valid trips).
- **Generate web search feature.** Based on the specified 44 COVID-19 symptoms, we count the number of symptom searches for each permanent resident using Java (step (4)), aggregate them by urban districts (step (5)), and arrive at the web search matrix \(H_t\).

A.3 Training Algorithm

We present the training algorithm as Algorithm 1. We calculate the gradient and update model parameters for each batch of data.
For SAB-GNN-wsa and SAB-GNN, we try following hyperparameter selection:

- Learning rate $\in \{1e^{-1}, 1e^{-2}, 1e^{-3}, 1e^{-4}, 1e^{-5}\}$.
- Batch size $\in \{2, 4, 8, 16\}$.
- Dropout rate $\in \{0.3, 0.5, 0.7\}$.
- $L_1 \in \{1, 2, 3\}$: the number of layers in GNN.
- $L_2 \in \{1, 2, 3\}$: the number of layers in LSTM.
- $w_u \in \{1, 3, 8, 12, 44\}$: $w_u$: the number of used symptom features (recall that $h^u_t \in \mathbb{R}^{n_u}$).

### A.5 Comparison with Existing Studies

We present existing COVID-19 infection studies in Table 4 based on prediction period, used feature data, and model architecture. We find that mobility data owned by different companies such as Baidu, Facebook, Google, and Swisscom served as the primary sources of COVID-19 infection prediction before Jun. 2020 when the first COVID-19 outbreaks occurred in many countries. As the pandemic propagated and evolved worldwide with the total infection cases exceeding 530 million as of June 7, 2022 [32], the infection dynamic involved more factors such as virus mutation, mask policy, vaccination, and travel policy changes.

Our study makes the first attempt to introduce web search data, which is a representative signal of unconfirmed positive cases within the population, into the multiwave infection prediction. Besides, we consider the temporal decay of social awareness of COVID-19 symptoms in our proposed social awareness GNN, to fully leverage the power of web search data. The reason is that people’s cognition changed during the two-year pandemic.

**Our goal is not to beat all other models.** COVID-19 spreading is a highly uncertain process affected by varying transmission discrepancy, virus mutation, human behavior, and vaccination penetrations in different periods and locations worldwide, which makes a universally best model impossible. Centers for Disease Control and Prevention announced that even the ensemble model was unable to provide reliable predictions all the time. Instead, the main takeaway is that we can leverage web search data to complement existing mobility-based infection prediction to confront the challenges of the multiwave infection uncertainty.

### A.6 Application on Social Media Data Mining

One contribution of this study is the social awareness recovery module (Equations 4, 5, 6). Social awareness depicts the general phenomenon that human perception and attention on events / products / news becomes weaker as time proceeds. Beyond disease prediction, it is promising to apply the social awareness recovery module to social media data from other sources (e.g., Facebook, Twitter, Weibo) in other applications such as predicting the visits to a new location, and the shopping demand of a new product.

---

Table 4: Comparison between existing studies and this study.

| Study                | Period          | Used features          | Model                  | Publication         |
|----------------------|-----------------|------------------------|------------------------|---------------------|
| Chimmula et al. [9]  | Jan. 2020 - Mar. 2020 | Infection             | LSTM                   | Chaos, Solitons & Fractals |
| Xiao et al. [48]     | Jan. 2020 - Mar. 2020 | Baidu mobility, etc.  | TL                     | AAAI 2021           |
| Schwabe et al. [41]  | Feb. 2020 - Apr. 2020 | Swisscom mobility, etc. | Hawkes model          | KDD 2021            |
| Panagopoulos et al. [33] | Feb. 2020 - May 2020 | Facebook mobility     | GPNN + TL             | AAAI 2021           |
| Kapoor et al. [23]   | Mar. 2020 - May 2020 | Google mobility       | GCN + Skip-connection  | ArXiv               |
| Kargas et al. [24]   | Mar. 2020 - Jun. 2020 | Medical data          | Epidemiological model  | AAAI 2021           |
| Gao et al. [13]      | Mar. 2020 - Jun. 2020 | Medical data          | GAT + GRU             | JAMIA               |
| Wang et al. [47]     | Mar. 2020 - Aug. 2020 | Medical data          | Ensemble               | IEEE Big Data 2020  |
| Adiga et al. [1]     | Aug. 2020 - Jan. 2021 | Infection             | Ensemble               | KDD 2021            |
| Chang et al. [6]     | Mar. 2020 - Feb. 2021 | SafeGraph mobility, etc. | Epidemiological model | KDD 2021            |
| Sesti et al. [42]    | Jan. 2020 - May 2021 | Infection             | GraphLSTM             | ICML 2021 Workshop  |

This study: Apr. 2020 - May 2021, Yahoo mobility, web search, Social Awareness GNN, KDD 2022.