Supplementary Information for

Temporal Contact Graph Reveals the Evolving Epidemic Situation of COVID-19

Mincheng Wu[1,†], Chao Li[1,†], Zhangchong Shen[1], Shibo He[1,7,*], Lingling Tang[2], Jie Zheng[3], Yi Fang[4], Kehan Li[1], Yanggang Cheng[1], Zhiguo Shi[5,7], Guoping Sheng[2], Yu Liu[4,7], Jinxing Zhu[4], Xinjiang Ye[4], Jinlai Chen[4,7], Wenrong Chen[4], Lanjuan Li[6,*], Youxian Sun[1], Jiming Chen[1,7,*]

[1]*College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China.*
[2]*Shulan (Hangzhou) Hospital Affiliated to Shulan International Medical College, Zhejiang Shuren University, Hangzhou, China.*
[3]*Zhejiang Institute of Medical-care Information Technology, Hangzhou, China.*
[4]*Westlake Institute for Data Intelligence, Hangzhou, China.*
[5]*College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China.*
[6]*State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, Zhejiang University, Hangzhou, China.*
[7]*Data Intelligence Research Center, Institute of Wenzhou, Zhejiang University, Wenzhou, China.*

[*] Corresponding author. E-mail:s18he@zju.edu.cn, ljli@zju.edu.cn, cjm@zju.edu.cn.

[†] These authors contributed equally.

Supplementary Note I: Data description

The original data of location-related information was collected by location-based service (LBS) providers in China, which have a long-term cooperation agreements with Westlake Institute for Data Intelligence (the affiliate of some coauthors of this article). The location-related information was uploaded every time smartphone users are using LBS. Smartphone users authorized such data collection process. According to the Personal Information Security Specification of China (2019), privacy protection mechanisms such as perturbation and pseudonymization are adopted during data collection.

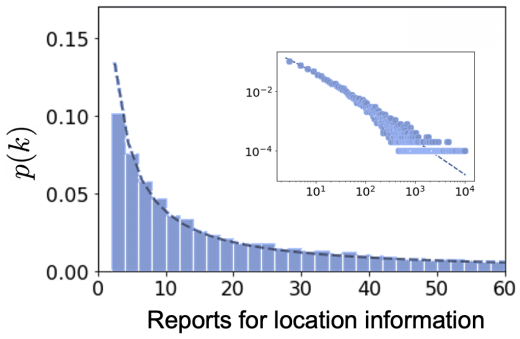
Since the outbreak of COVID-19 in China, Prof. Lanjuan Li, one of the five members in the high-level expert team for COVID-19 convened by the National Health Commission of China, Dr. Lingling Tang and Dr. Guoping Sheng, famous epidemiologists from Shulan (Hangzhou) Hospital, have been working on the front line for prevention and control of COVID-19. Westlake Institute for Data Intelligence led by Yi Fang also worked closely with Center for Disease Control and Prevention (CDC) on tracking contacts with infectious cases. Based on Public Health Emergencies Regulations of China, the local authority provided the device identifications of confirmed cases. Such information can be incorporated to link the health status of smartphone users since the app accounts in China are registered by these identifications.

Many recent works used very similar data collected from different sources. For example, Kraemer et al. evaluated how the individual mobility affects the disease transmission, and, as control measures can be reflected by the changes in individual mobility, further showed the positive effect of control measures in Wuhan on mitigating the spread of COVID-19[1]. Nicholas et al. used country-wide aggregated mobile phone data to show that the outflow from Wuhan accurately predicts the distribution of infections across all of China, and developed a "risk source" model to forecast confirmed cases and identify high risk locales at early stage. This work also derived the geographic spread and growth pattern of COVID-19[2]. Vespignani et al. used the data collected by Cuebiq Inc[3], which is similar with the location data we have. However, they simulate the infectious cases using a stochastic discrete-time compartmental model, which is not calibrated to account for the specific evolution in Boston. From the above summary, we can see that we utilized the real health status of users to construct the temporal contact graph, and provides an empirical evidence to reveal evolving epidemic situation of COVID-19. These distinguishes our work from existing ones. In other words, we took the first attempt to construct the contact graph between susceptible and infectious individuals to represent the process of digital contact tracing, while previous studies

did not utilized the information of confirmed cases.

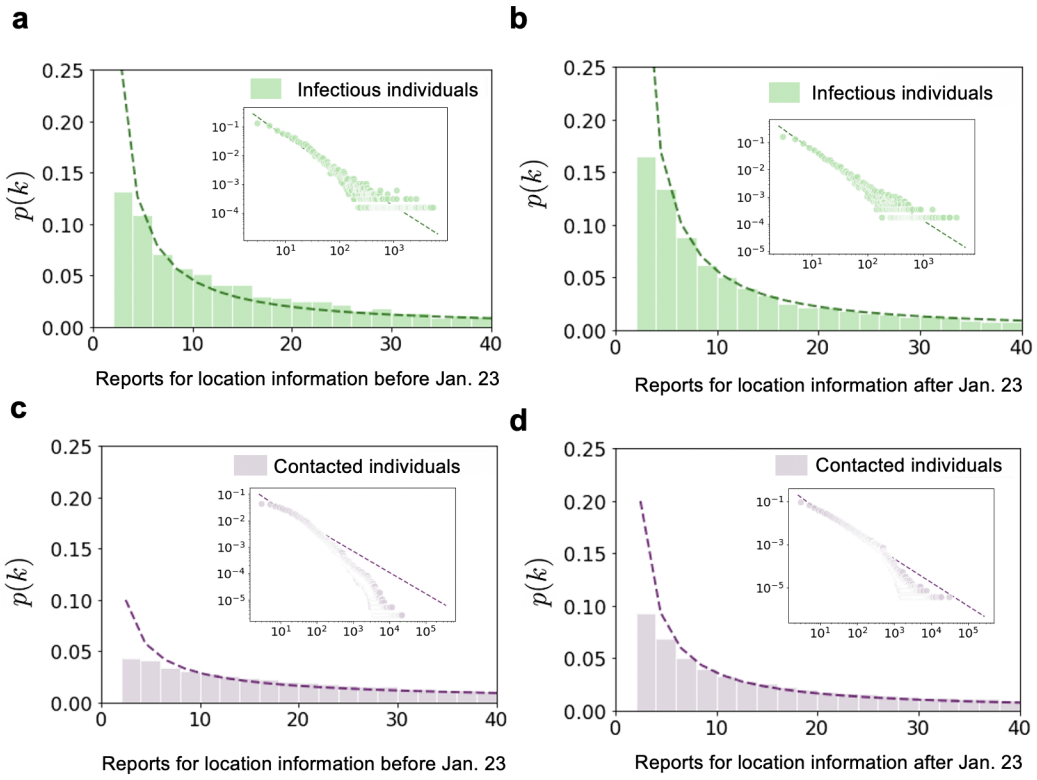
We note that all individual location-related data and health status information were collected, stored, and used by following the Personal Information Security Specification (2019) and Public Health Emergencies Regulations of China. All raw data was stored in specialized data servers with limited access by LBS providers. For the research purpose, the Westlake Institute for Data Intelligence constructed the temporal contact graph according to the contact model we built and provided it for further analysis. All the results in the article can be obtained and validated by the temporal contact graph, which does not contain private geographic and identity information about the contacted individuals and the confirmed cases. This project was reviewed and approved by the Medical Ethics Committee of School of Medicine, Zhejiang University.

Supplementary Fig. **1** analyzes the distribution of the number of reports for all the confirmed cases and the contacted individuals under the proposed contact model. Every report is counted no matter whether it contributes to a contact or not. The number of reports related to location-based information follow well-known power-law distribution, i.e., $p(k) \propto k^{-\gamma}$, where the average piece of information $< k >$ equals to 311.63 and the power exponent γ equals to 1.15.



Supplementary Figure 1: **Distribution of reports history in 20 days.** Distribution of reports for location-based information, and 10.11% of users contribute less than three time in 20 days.

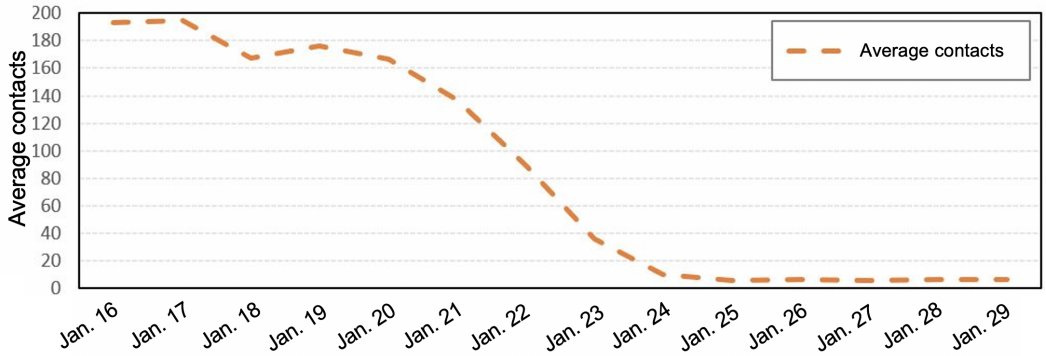
As shown in Supplementary Fig. **2**, we studied the crowdsourced data before and after the date of the travel restriction, i.e., 23 January, 2020. Among the data, the location information, which is used to calculate contacts, is preprocessed firstly. Then, we count the reporting data of each user in seven days and study the distribution of the crowdsourced data. In the first instance, we can find that the contributed data are sparse and the majority of users contribute less than three time



Supplementary Figure 2: **Distribution of reporting crowdsourced data before and after 23 January, 2020.** **a** Distribution of confirmed cases' reports for location information before 23 January, 2020. **b** Distribution of confirmed cases' reports for location information after 23 January, 2020. **c** Distribution of contacted individuals' reports for location information before 23 January, 2020. **d** Distribution of contacted individuals' reports for location information after 23 January, 2020.

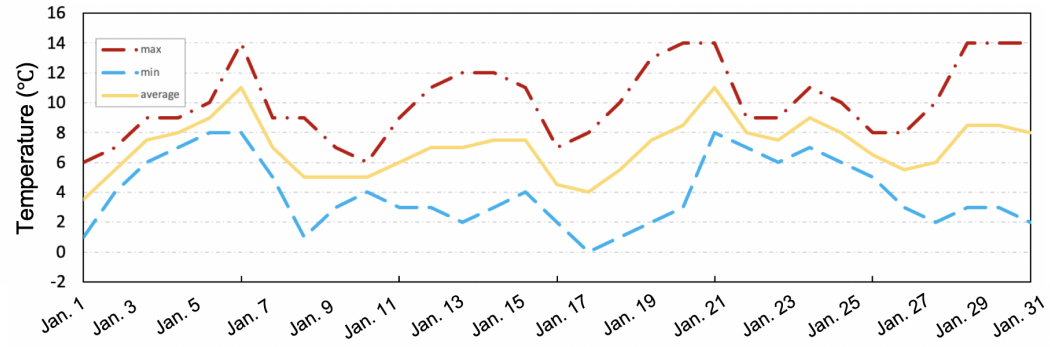
in seven days because smartphone users report data in a very low and irregular frequency. Besides, the distributions of the reporting data from both confirmed cases and contacted individuals in the whole period follow power-law distributions, following $p(k) \propto k^{-\gamma}$, where $< k >= 93.04, \gamma = 1.2$ for location information reported by confirmed cases before 23 January, 2020 (Supplementary Fig. **2**a) and $< k >= 58.68, \gamma = 1.3$ for location information reported by confirmed cases after 23 January, 2020 (Supplementary Fig. **2**b) and $< k >= 310.09, \gamma = 0.8$ for location information reported by contacted individuals before 23 January, 2020 (Supplementary Fig. **2**c) and $< k >= 142.46, \gamma = 1.1$ for location information reported by contacted individuals after 23 January, 2020 (Supplementary Fig. **2**d). We can find that for both confirmed cases and contacted individuals, the mean value of crowdsourced reporting frequency decreased after 23 January, 2020, while the power

exponent increased after 23 January, 2020, which means people report less location information. The reason is that the travel restriction was implemented in Wuhan on 23 January, 2020 and people tended to stay at home.



Supplementary Figure 3: **Average number of contacts of the entire population.** Before and after the travel restriction of the general population daily contact number changes.

Supplementary Fig. **3** shows the average contact trends of the whole population (10,527,737 smartphone users) in Wuhan during the period from 16 to 29 January, 2020. After the Chinese authority confirmed the COVID-19 coronavirus can be transmitted among human on 20 January, 2020, the average daily contacts of the whole population dropped sharply. After the ban of non-essential vehicles in Wuhan downtown area on 26 January, 2020, the average contacts reached the lowest, at about 6.3, respectively. And then the average contacts stayed stable at such a level for a long period.

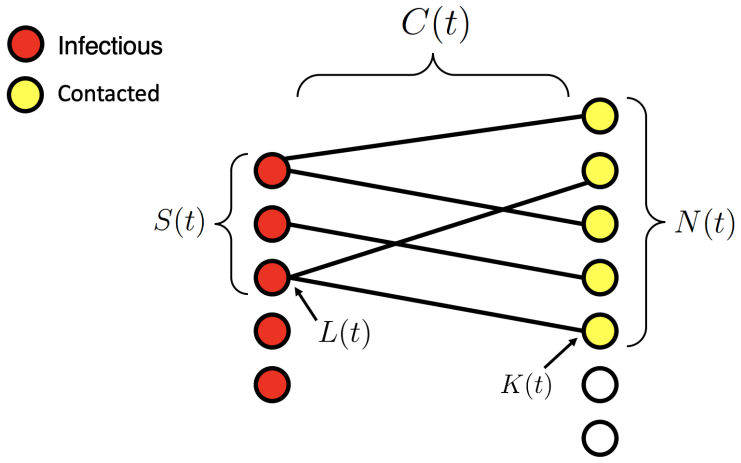


Supplementary Figure 4: **Daily temperature in Wuhan from 1 January, 2020 to 31 January, 2020.** The weather data comes from meteoblue.com.

Supplementary Note II: Statistical analysis

Constructed graph structure

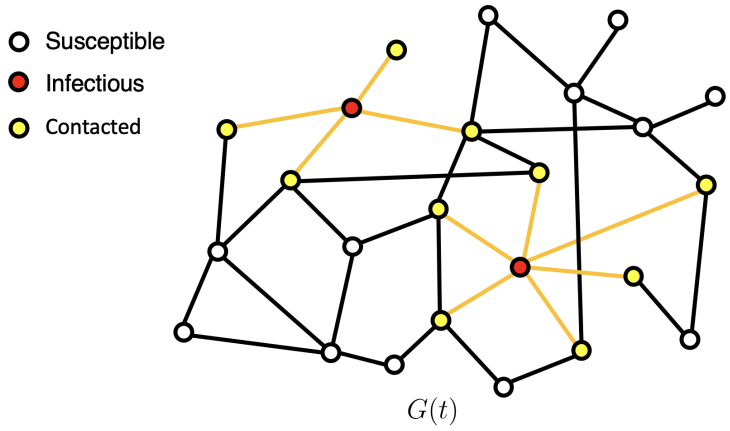
In the article we have introduced five indicators by leveraging the temporal contact graph: the total number of daily contacts $C(t)$ by calculating the number of edges of the temporal contact graph in day t, the total number of active infectious individuals $S(t)$, the number of susceptible individuals $N(t)$ who had at least one contact with infectious individuals, the average number of susceptible individuals $L(t)$ that each infectious individual contacted in day t (by calculating the average degree of the nodes representing infectious individuals in the graph), the number of infectious individuals $K(t)$ that each susceptible individual encountered in day t (by calculating the average degree of nodes representing the susceptible). A toy example is shown for specific calculation (Figure **5**).



Supplementary Figure 5: **A toy example for a snapshot (one day) of the temporal contact graph.** Nodes on the left are infectious individuals and the nodes on the right are susceptible individuals. An edge indicates a contact between the two individuals in this day. The total contacts $C(t)$ is the number of edges in the bipartite graph, i.e., $C(t) = 5$. The number of active infectious individuals equals three, i.e., $S(t) = 3$, while the number of contacted individuals equals five, i.e., $N(t) = 5$. Also, we can calculate that $L(t) = 5/3$ and $K(t) = 1$.

Since the fraction of the infectious individuals in a population is very small, $L(t)$ is much larger than $K(t)$ generally. For example, in the contact graph $G(t)$, $K(t) = 1$ while $L(t) = 5$ (Supplementary Figure **6**). As a matter of fact, the two metrics are on very different orders

of magnitude, since the contact graph is much larger than the toy example. The units for $K(t)$ and $L(t)$ are people/day. Thus, once the temporal contact graph is constructed by the contact model, the five informative indicators can be determined. However, different parameters do impact the measurements for the contact such as the infectious period and time interval in the contact model. We have done more analysis for the sensitivity of these parameters to support the results statistically. In the manuscript, we claim that $N(t)$ and $K(t)$ are similar rather than $N(t)$ and $S(t)$. Obviously, $N(t)$ and $K(t)$ increase from 1 January, 2020 to 20 January, 2020, and decrease from 20 January, 2020 to 25 January, 2020.

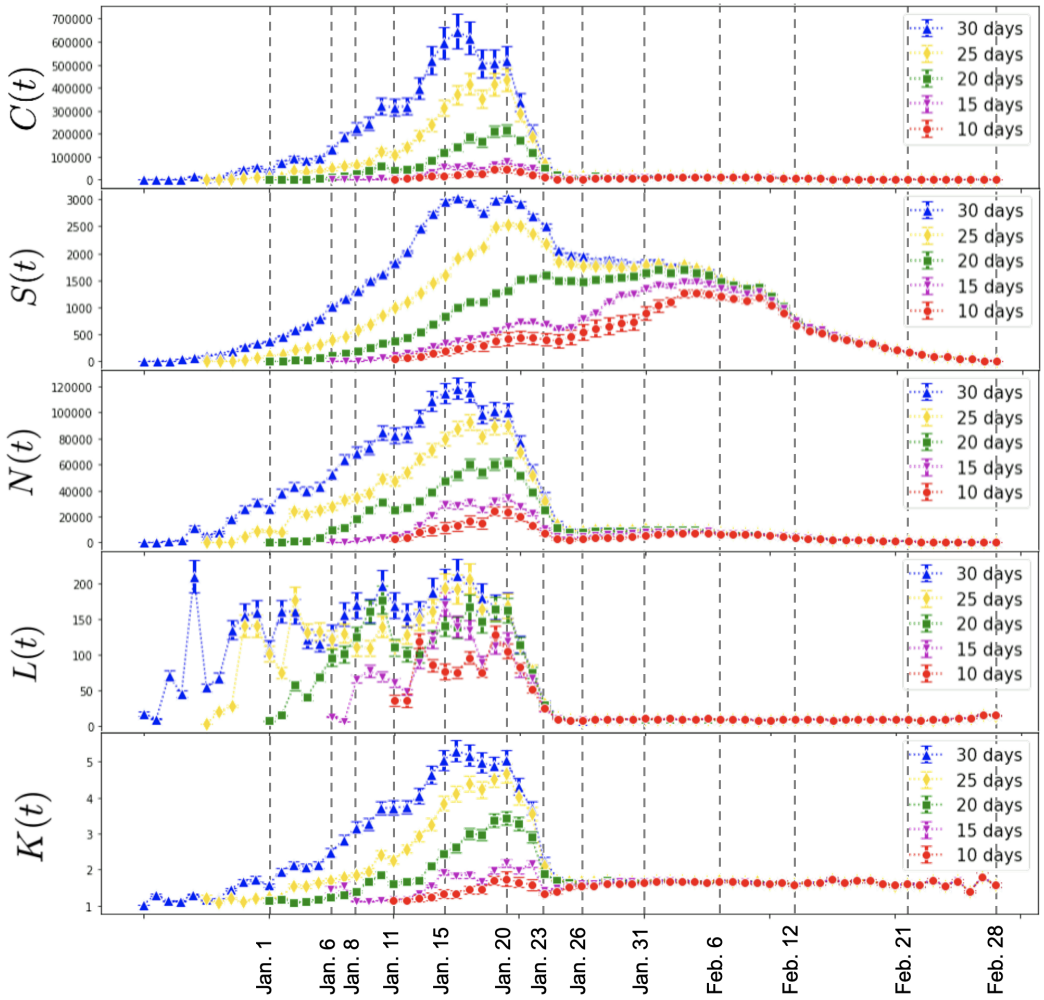


Supplementary Figure 6: **A toy example for a snapshot of the spreading.** The spreading in day t is represented by the graph $G(t)$, where nodes are individuals and edges are contacts between two individuals. We denote different health statuses by distinct colors, and the contacts between infectious and contacted individuals are colored by orange.

Since the infectious period is another parameter to determine the temporal contact graph, we perform sensitivity analyses for this period from 10 days to 30 days (Supplementary Figure **7**). Although the absolute values of the five indicators change, we can see that the trends remain stable.

Infectious period

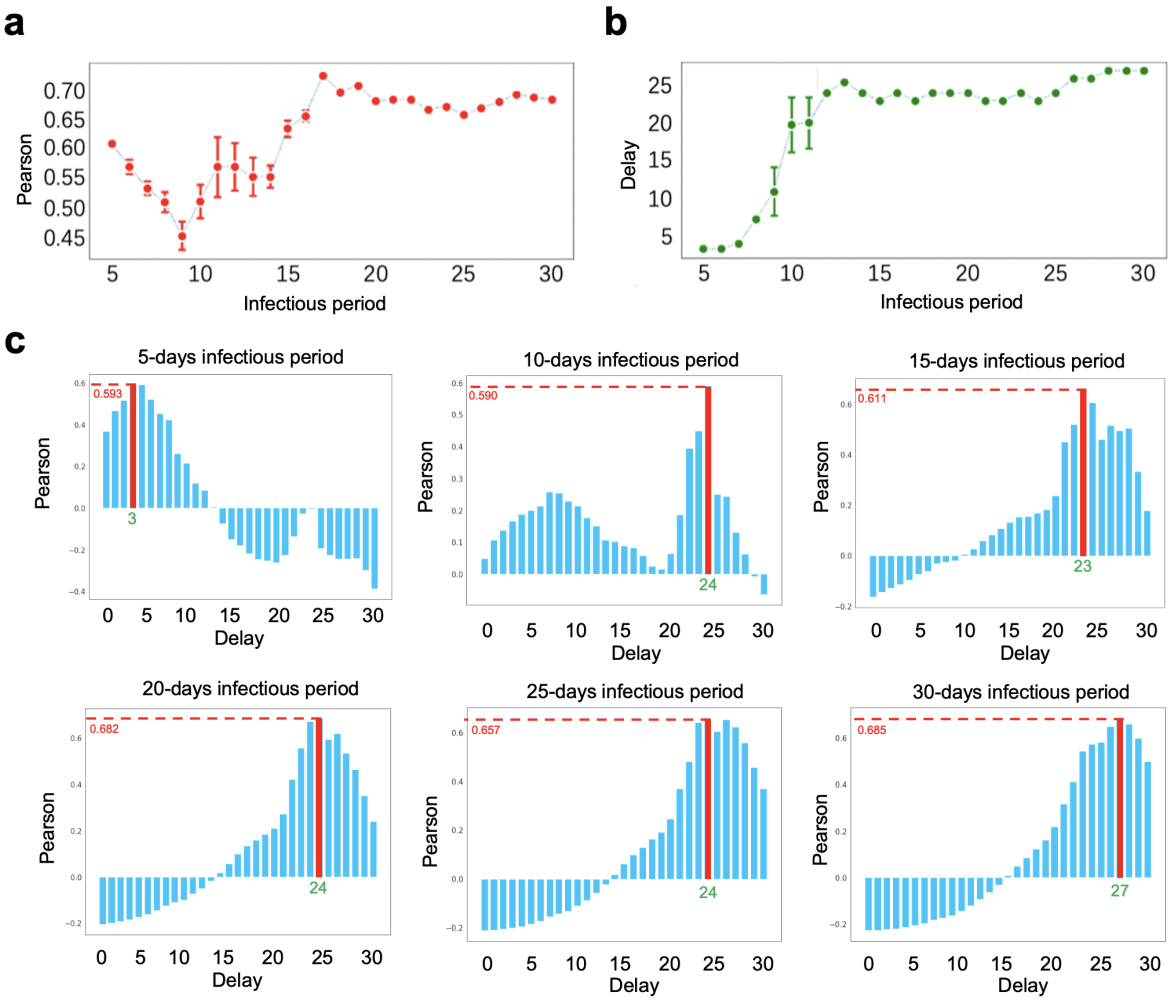
As mentioned in the Data description, the collected data include location-based information. Based on the collected data, we introduce a contact model, quantifying a contact between two individuals when they are geographically close to each other. Specifically, since locations of individuals is perturbed, we define a contact occurring when two individuals are in the same perturbed area within a given time interval. Based on the proposed contact model, we constructed the



Supplementary Figure 7: **Sensitivity analysis for the infectious period a.** $C(t)$, the daily total times of contacts between infectious individuals and contacted individuals. **b.** $S(t)$, the daily total number of infectious individuals that have contacts with contacted individuals at least once. **c.** $N(t)$, the daily total number of contacted individuals that have contacts with infectious individuals at least once. **d.** $L(t)$, the daily average times of contacts with contacted individuals for infectious individuals before confirmation. **e.** $K(t)$, the daily average times of contacts with infectious individuals for contacted individuals.

temporal contact graph, which describes the daily contacts between the susceptible and infectious individuals: a node in the graph means a susceptible or infectious individual, and an edge in day t indicates a contact occurring in day t between a susceptible individual and infectious individual.

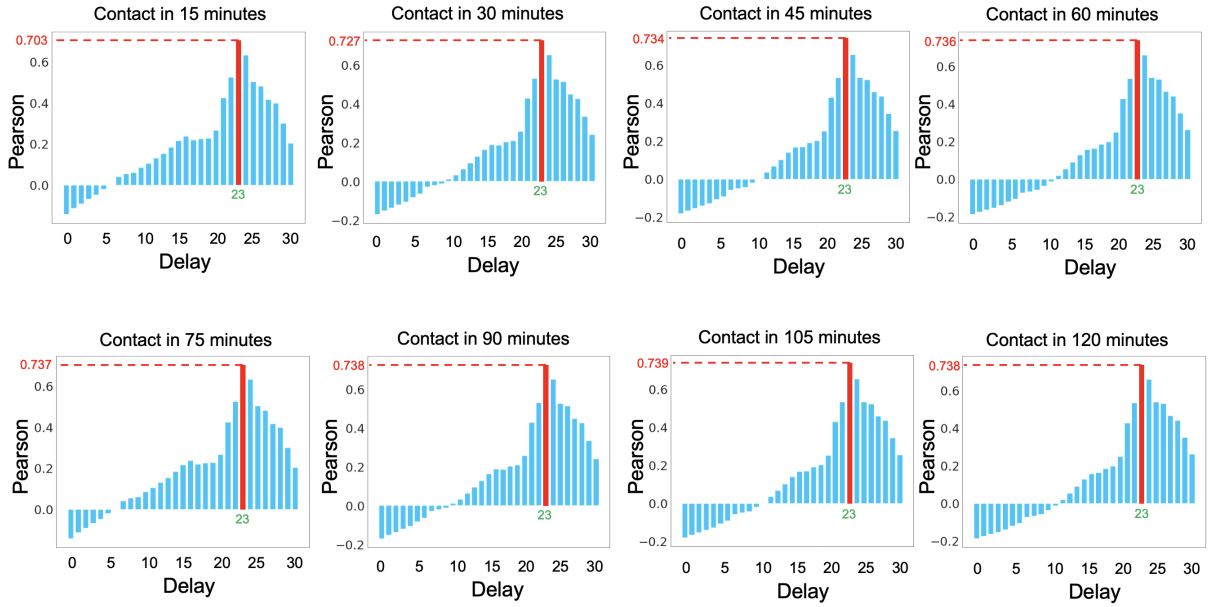
We calculate the Pearson correlations between daily contacts and infectious individuals with delays ranging from 5 to 30 days for different infectious period (Supplementary Figure **8**). Obviously, the experiments show that a 23-day delay results in the best Pearson correlation of 0.78, where the corresponding infectious period equals 17 days in accordance with existing surveys[4–11]. Thus, we consider the infectious period to be 17 days in the article, and an individual is identified as infectious in day t if he/she was confirmed during $[t+1, t+17]$.



Supplementary Figure 8: **Sensitivity for the infectious period.** We calculate the Pearson correlations between the daily contacts the daily confirmed cases with a delay ranging from 0 to 30 days.

We also performed the sensitivity analysis for the contact model by varying the time interval from 15 minutes to 120 minutes and the infectious period from 1 days to 30 days in the contact

model. Specifically, we vary the time interval from 15 minutes to 120 minutes and test the contact models under different time granularities, finding that such a time granularity does not change our conclusion in this article. We find that the Pearson coefficient between the daily number of contacts and delayed daily number of confirmed cases reaches maximum when the delay is 23 days, corresponding to a 17-days infectious period for all values of time interval (Supplementary Figure **9**). This indicates that the proposed contact model is very stable.

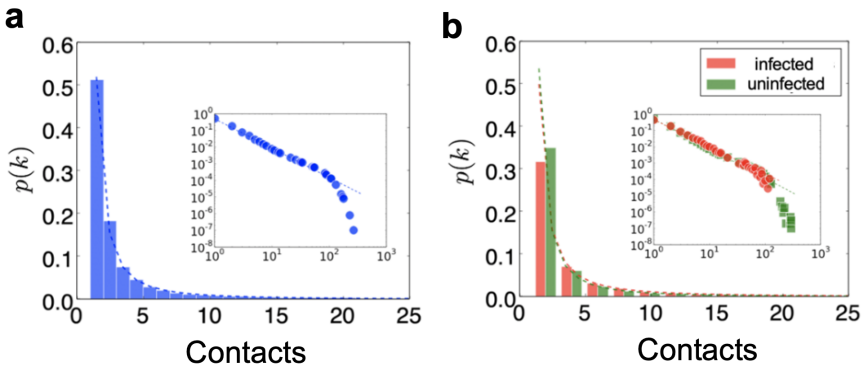


Supplementary Figure 9: **Sensitivity for the time interval T.** The Pearson correlations between daily contacts and daily confirmed cases with a delay ranging from 5 days to 30 days. Each panel corresponds to the time interval T for a contact in the contact model.

KS test

Since the distribution of various type of contacts follow a power-law, here we will discuss in details from a statistical stand point. In this article we have performed the maximum likelihood estimation for the parameters in the model to describe distributions of various types of contacts. In fact, we proposed a probability model based on Bayesian framework to evaluate the risk for every contact. Thus, we described the behaviors by power-law distributions, since we found there is a linear relation under the log-log coordinate. We have performed statistical analysis for our results by the KS test[12]. We calculated the p-value for the distributions, where p-value is exactly the portion of synthetic sequences whose KS-distance is larger than that of the real data. For example, the exponent γ of the distribution of contacts for the whole population displays $\gamma = 1.96$, while

the p-value is 0.53. The exponent γ of that for the infected contacts displays $\gamma = 1.86$, while the p-value is 0.68. The exponent γ of that for the uninfected susceptible displays $\gamma = 2.01$, while the p-value is 0.52 (Supplementary Figure **10**).

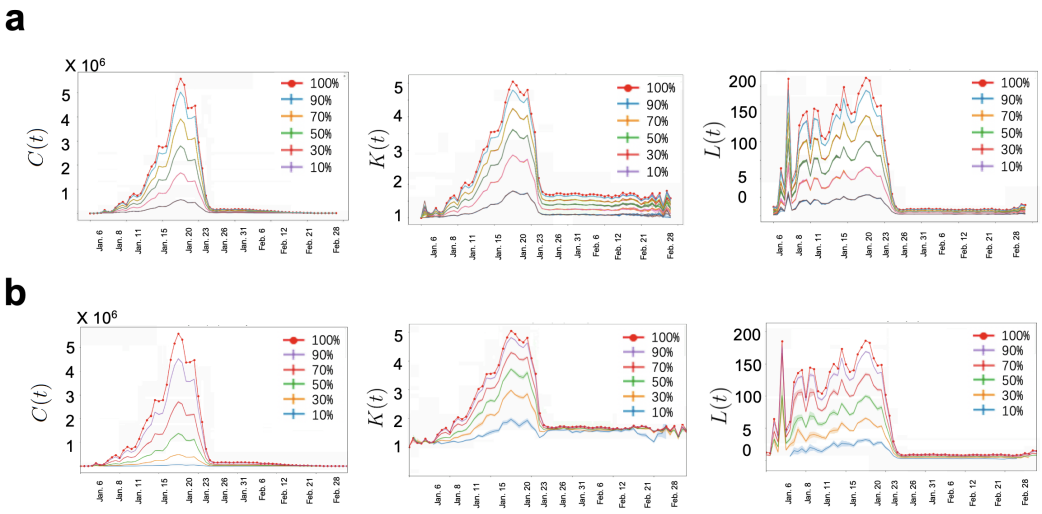


Supplementary Figure 10: **Distributions of behaviors for infected and uninfected individuals. a.** The distribution of contacts for all contacted individuals. **b.** The distribution of contacts for all infected individuals (red bars) and uninfected individuals (green bars).

User involvement

To perform sensitivity analysis for the impacts of users' involvement, we further conduct experiments to simulate different user involvements by randomly selecting $\alpha\%$ users as the voluntary users, and $\alpha\%$ data items each user uploading per day, and evaluate the corresponding performance loss. To show the robustness of our analysis, we repeat 10 times Monte Carlo experiments for a given setting. Specifically, we set user involvement rate to 10%, 30%, 50%, 70%, 90%, and user uploading rate to 10%, 30%, 50%, 70%, 90%. Then, we perform the corresponding experiments. We plot the curves of $K(t)$, $L(t)$ and $C(t)$ calculated from the contact model with different uploading rates from 1 January to 28 February with error bars (Supplementary Figure **11**a). The results show that with the decrease of uploading rates, the values of three parameters decrease obviously, which is because the number of recorded contacts drops along with decreasing upload rates. Notice that the error bars of three parameters are all small, which implies the Monte Carlo experiments are stable to produce similar results. It is, therefore, reasonable to emphasize that the number of user uploads reduced for each Monte Carlo experiment is the same, and the total number of participation does not reduce under user upload analysis. Generally speaking, the Pearson correlation between case of 10% uploading rate and other cases ($\alpha\%$ upload rates) remains high for the analysis of user uploading rate. This phenomenon implies that we might not need a high uploading rate if we are only interested in estimating the trends.

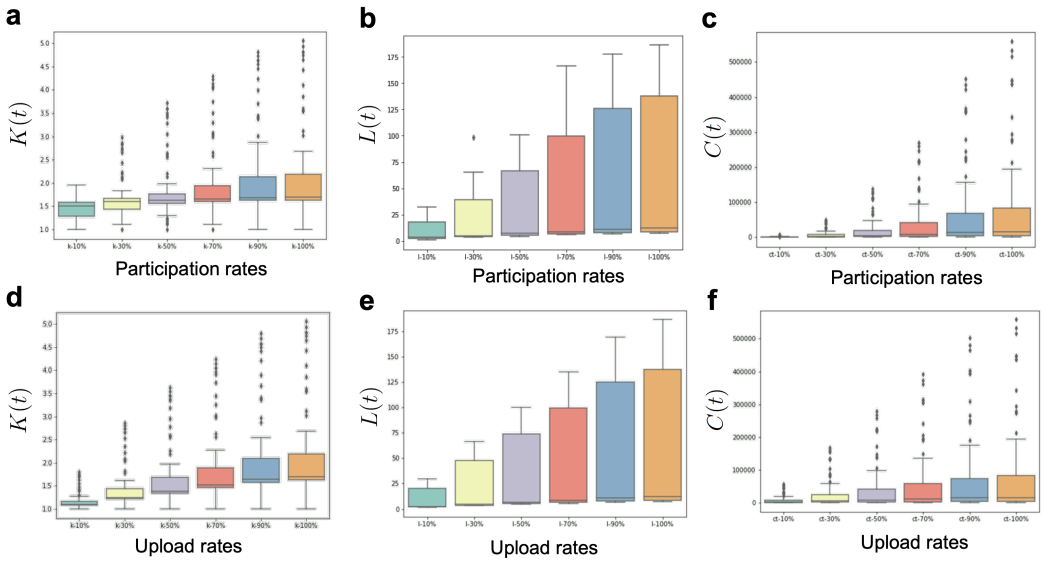
We conduct experiments in the same way as those for uploading rates to evaluate the impact of diverse participation rates (Supplementary Figure **11**b). The corresponding results show that, compared with uploading rate analysis, decreasing the participation rates brings more uncertainty in the Monte Carlo experiments with higher error bars. This is because the participants that are randomly deleted at each time experiment are different. Thus, their impacts on the entire network also vary. Notice that when the participation rate is not very small, the high correlation can still be preserved. However, when it is reduced to 10%, the correlation coefficient decreases significantly. This could be attributed to the power law distribution of the network: since the distribution has an obvious long-tail effect, only when the participation rate is low enough can some key nodes be deleted, thereby affecting the trend of the proposed metrics.



Supplementary Figure 11: **Sensitivity analysis for user involvement. a.** The proposed metrics including $C(t)$, $K(t)$ and $L(t)$ are shown, when simulating the uploading rate ranging from 10% to 90%. **b.** The proposed metrics including $C(t)$, $K(t)$ and $L(t)$ are shown, when simulating the participation rate ranging from 10% to 90%.

We have analyzed the contact model in main body of the paper with six curves showing different user participation rates and upload rates. Here we give the corresponding statistical information such as median and variances of $K(t)$,$L(t)$ and total contacts $C(t)$ shown in Supplementary Figure **12**. It reveals that as $\alpha\%$ of participation and upload rate decreases, the statistical information decreases with the similar trend. This is expected as reduction in either user participation rate or user upload rate decreases the chances of having contacts among users.

Moreover, the practical challenges for real contact tracing apps are: what to do if we have



Supplementary Figure 12: **Performance of different user involvement in the contact model. a-c.** Three box plots show the distribution change of $K(t)$, $L(t)$, and daily number of total contacts Vs. different user participation rates. **d-f.** Three box plots show the distribution change of $K(t)$, $L(t)$, and daily number of total contacts Vs. different user upload rates.

only partial information about contacts and about the real individual health status? As a matter of fact, we have investigated this in this article, indicating the impacts of user involvement on the contact tracing app performance. We have conducted experiments to simulate different user involvements by randomly selecting $\alpha\%$ users as the voluntary users, and evaluate the corresponding performance loss, where $\alpha\% = 10\%, 30\%, 50\%, 70\%, 90\%$. In such a way, both susceptible and infectious individuals could be removed, which simulates partial information case about both susceptible and infectious individuals. One interesting result in our work is even though we only have partial information about the contacts and the conformed cases, we can still have a good performance on estimating the evolving situation of COVID-19 when $\alpha\%$ is not low enough.

Supplementary Note III: Data and real contagion

Data biases

In this article, we analyze the evolving epidemic situation of COVID-19, leveraging the available reported 'confirmed cases' as a proxy of the measure of the extent of the contagion However, many other factors (other than actual COVID19 transmission) may have influenced the recorded number of confirmed cases because of medical resource lacking, incomplete information and so on. We will consider how could these biases in the number of confirmed cases (compared to the number of actual cases) shape the actual results. In fact, the reported number of confirmed cases is typically less than the number of actual cases. Therefore, there is always biases in reality. However, it is impossible to know the actual cases even their distributions. Here, we assume that the reported confirmed cases $Sc(t)$ is proportional to the number of actual cases in reality $Sr(t)$, i.e.,

$$Sc(t) = \beta \cdot Sr(t), \tag{1}$$

where $0 < \beta < 1$ is a constant to quantify the probability for an actual case being confirmed. Then, in this article we perform correlation analysis between the number of daily contacts $C(t)$ and daily confirmed cases $Sc(t)$, and obtain a Pearson coefficient $\rho(C(t), Sc(t)) = 0.78$. Then, we will prove the Pearson coefficient between the daily contacts $C(t)$ and daily infected individuals in reality $Sr(t)$ is the same, i.e.,

$$\rho(C(t), Sr(t)) = \rho(C(t), \beta \cdot Sc(t)) = \rho(C(t), Sc(t)) = 0.78. \tag{2}$$

In other words, although $Sc(t)$ has a bias from $Sr(t)$, i.e., $Sc(t) = \beta \cdot Sr(t)$, it does not affect the correlation analysis in our results.

Recorded contacts and transmission

A primary conceptual focus of the manuscript is how real-world social contacts may relate to COVID-19 contagion patterns. For instance, we define a contact as a co-occurrence within a specific distance (e.g., $15 \times 30m^2$) in the proposed contact model, when in fact it is unlikely that this virus would be transmitted at a range even 10 fold less than this. Therefore, it is necessary to clarify how this might influence the patterns found, or the conclusions drawn. Medically speaking, a close contact is said to occur when two individuals are within a distance of 1.8 meters [13]. Individuals who have close contacts with infectious cases generally have a high probability of getting infected.

However, in practice, it is difficult to decide a close contact in a digital way. Most contact tracing apps exploit Bluetooth and/or GPS to decide a contact when two individuals are in a short distance (e.g., within 20 meters). In fact, one of the current controversial issues lies in whether such type of contacts captured by contact tracing apps is effective since it is not fine-grained enough. For example, the mentioned potential interesting work discussed the importance of network structure and social dynamics in evaluating the potential impact of SARS-CoV-2 control by combing fine-scale data[14]. However, our results show that there is still a clear distinction of contact behaviors between the infected and uninfected contacted individuals under our contact model. Therefore, the frequency of social contacts captured by contact tracing apps actually has a high correlation with contagion patterns.

Risk evaluation

In this article we have considered the risk by their behaviors and features, and we use true/false positive and the ROC curve to analyze the effectiveness of the risk model. Notice that the "positive" in the phrase "true/false positive rates" does not indicate the "positive" in a nucleic acid testing. In fact, we have measured the risk of every contact j, i.e., $P(z_j = 1)$ by the proposed risk model in the Result III. In order to evaluate the risk measured by the model, we study the ROC (receiver operating characteristic) curve. For a threshold $0 < q < 1$, specifically, a contact j is considered to be true positive if $z_j = 1$ and $P(z_j = 1) > q$, while j is considered to be false positive if $z_j = 0$ and $P(z_j = 1) > q$. Then, we can calculate the TPR (true positive rate) by

$$TPR = \frac{\sum_j \mathbf{1}_{\{\mathbf{z_j=1}, \mathbf{P(z_j=1)>q}\}}}{\sum_j \mathbf{1}}, \tag{3}$$

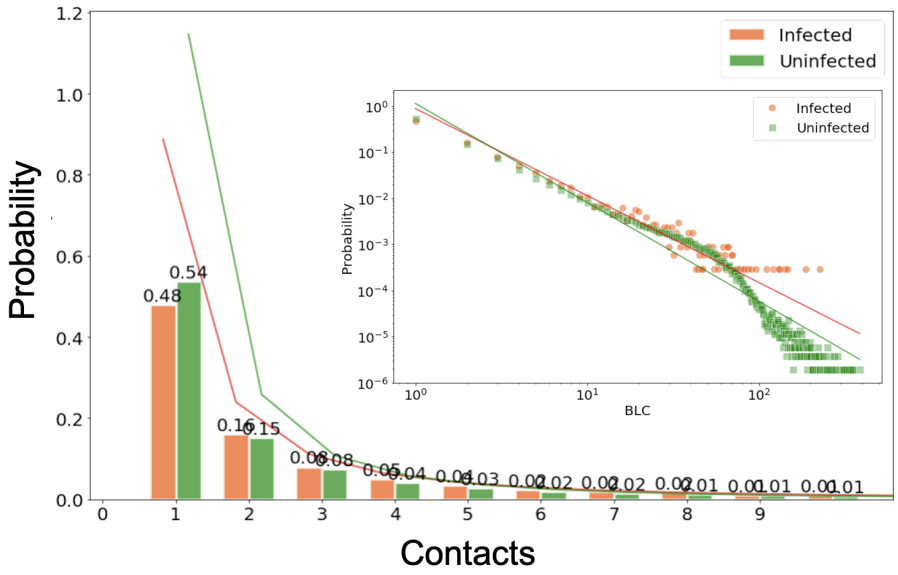
and FPR (false positive rate) by

$$FPR = \frac{\sum_j \mathbf{1}_{\{\mathbf{z_j=0}, \mathbf{P(z_j=1)>q}\}}}{\sum_j \mathbf{1}}. \tag{4}$$

Thus, the ROC curve described by TPR and FPR can well evaluate the risk model for contacted individuals.

We have mentioned that there is a distinction of contact behaviors between the infected and uninfected contacted individuals in the article, where the distinction indicates the difference of the probability distribution for contacts between the infected contacts and the uninfected contacts. Here, we calculated the KL-divergence for the distributions, quantified the difference and showed if the distinction is prominent. Specifically, we calculated the two distributions for the infected

contacts and the uninfected contacts (Supplementary Figure **13**). We also calculated the KL-divergence between the two distributions, where the KL-divergence equals to 0.02. These results all show that there is a slightly distinction of contact behaviors between the infected and uninfected contacted individuals.



Supplementary Figure 13: **Distinction between infected and uninfected individuals.** The distributions of contacts for infected individuals (red bars) and uninfected individuals (green bars)

Tracing apps

In this article, we use the location-related data from tracing apps to calculate if two individuals are close. Based on this, we build the temporal contact graph, which shows the contact relationship between susceptible and infectious individuals. Although the data used in our article is different with the Bluetooth based contact tracing apps, the principles used to define the potential contact are the same. By studying co-location data for 10 millions mobile phone users in Wuhan, our aim is to understand and clarify the potential of contact tracing apps to identify and interrupt transmission chains of the SARS-CoV-2 virus. However, controversies on the contact tracing apps mainly include: 1) the contact tracing apps may have privacy leakage issues and the methods (centralized V.S. decentralized) used to inform the potential contacts with infectious cases about their risks need further investigation; 2) current contact tracing apps utilize location-related information (e.g., Bluetooth, WiFi or GPS) to define a contact and are not fine-grained enough to capture a close contact within a distance of 1.5 meters; 3) such apps may not work for suppressing the transmission of COVID-19 when the participation rate is not high enough. For the first concern,

Google recently showed that the privacy can be protected by using Bluetooth with a proper privacy-preserving protocol. Our work here studies the contact behavior analysis in the transmission, and thereby does not discuss the privacy issues. Note that our approach falls within the category of the centralized way of informing the potential contacts. Though both centralized and decentralized ways can identify individuals having contacts with infectious cases, our results demonstrate that centralized way can provide an abundance of information that can be helpful for prevention and control of COVID-19 (see explanations for the second and third issues as follows). For the second issue, it is true that contact tracing apps cannot accurately define a close contact and may ignore other important factors that impact the spread of COVID-19, e.g., whether a protection measure such as wearing a mask is taken in a contact. However, our results show that there is still a prominent distinction of contact behaviors between the infected and uninfected contacted individuals. Based on this, we designed an infection risk evaluation framework to identify potential infected ones. For the third issue, we evaluated the effect of user involvement and show that user participation rate exerts higher influence on situation evaluation than user upload rate does. Moreover, our results indicate that the contact tracing apps can still be helpful even when user involvement is low. Also, we find that five indicators calculated from the constructed temporal contact graph are informative to understand the actual transmission and evaluate the epidemic situation. In summary, though this article cannot solve all the controversial issues, we provide new evidence that contact tracing apps can be very helpful to the prevention and control of COVID-19.

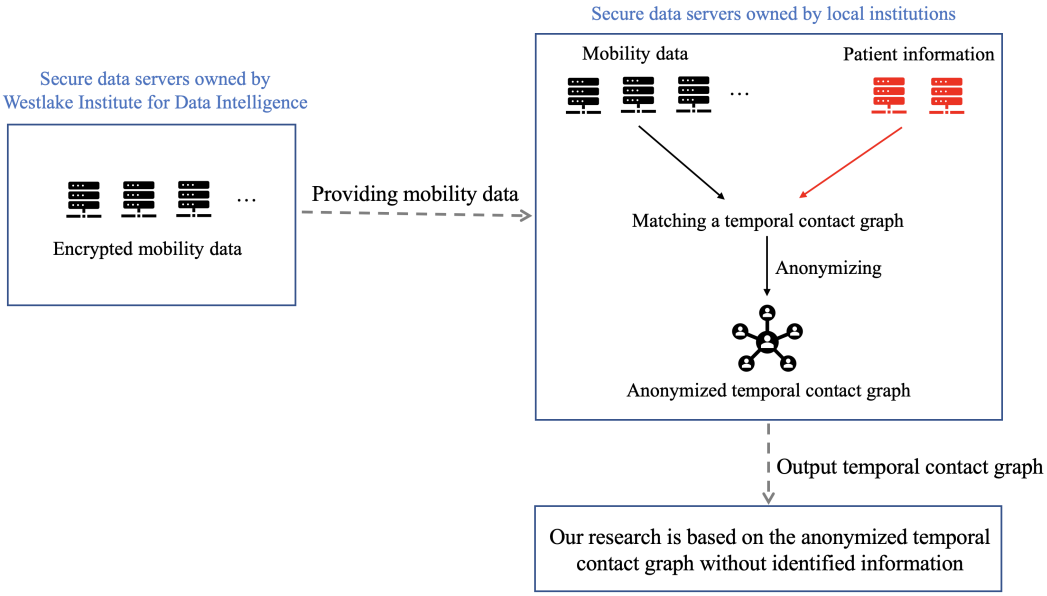
Currently, many other papers in recent times (the company Cuebiq for example provided a lot of similar data to many researchers in Europe and USA), used data collected from GPS, POI as well as wifi. In fact, most contact tracing apps exploit Bluetooth and/or GPS on smartphones to discover nearby devices held by users and identify the contacts between the susceptible and infectious individuals. For example, most contact tracing apps defined a contact by a Bluetooth "handshake", which occurs for two individuals within a distance of 20 meters. Some other apps used additional data such as GPS for contact tracing, e.g., Norway released an app that collects both GPS and Bluetooth information. In this article, a contact is identified for two individuals within about 20 meters. Clearly, our contact model works in a very similar way to those defined by most of the contact tracing apps, though the data is not directly collected from contact tracing apps (which is not available). To summarize, although the data used in our paper is different with the Bluetooth contact tracing apps, the principles used to define the potential contact are just the same.

Ethic and privacy issues

Since many ethic issues have been raised, it is of significance to discuss the guidelines for COVID-19 tracing apps[15]. Local institutions for disease prevention and control have the authority to collect the information on COVID-19 (including the patient information) and they also have the obligation to protect the privacy of patients when processing these data according to the Law on the Prevention and Control of Infectious Diseases of the People's Republic of China. Specifically, Article 12 of the Law on the Prevention and Control of Infectious Diseases of the People's Republic of China stipulates that "All units and individuals within the territory of the People's Republic of China shall **accept the preventive and control measures taken by disease prevention and control institutions and medical agencies for investigation, testing, collection of samples of infectious diseases and for isolated treatment of such diseases, and they shall provide truthful information about the diseases.** Disease prevention and control institutions and medical agencies shall not divulge any information or materials relating to personal privacy". Therefore, since the outbreak of COVID-19, local institutions have collected a large set of data. To protect the privacy of patients, these data are stored on the secure data servers owned by local institutions.

In our project, we identified the confirmed cases based on the phone number information. However, we hereby clarify that we (including authors from Westlake Institute for Data Intelligence) did not obtain any information regarding the exact phone numbers of patients. As mentioned in the first paragraph, these data are stored on the secure data servers and are not available for us. To facilitate our research, Westlake Institute for Data Intelligence uploaded the mobility data contributed by smartphone users and their account information in Wuhan from Jan. 1 to Feb. 28 to the secure data servers owned by local institutions. It is noteworthy that the phone number information here refers to the encrypted (i.e., hashed) information from the original phone number. To protect users from privacy leakage and malicious attack, their phone numbers were mapped to hash values. This procedure was irreversible and would be completed once the account was created. This is actually a popular approach adopted by many location-based service providers. The phone numbers of patients were also mapped to hash values by the same hash function on the secure data servers by local institutions. Since a phone number yielded a unique hash value, we could leverage such a connection to identify the patients. The temporal contact graph encoding the contact among the infectious and the susceptible individuals was then constructed.

A schematic diagram of the above procedure is shown in Figure **14**. The whole process of constructing the temporal contact graph was conducted on the secure data servers of the local institutions. In the temporal contact graph, each node (a user) is further anonymized and thus cannot be traced back to his/her phone number information (i.e., hashed values) anymore. This means



Supplementary Figure 14: **The procedure of constructing a temporal contact graph.**

that we were unable to gain access to any identifiable phone number or hash value about each node in the temporal contact graph even though some nodes were identified as confirmed cases. Moreover, the mobility data uploaded by Westlake Institute for Data Intelligence were destroyed on the secure data server, and the temporal contact graph was provided for us offline only for research purpose. Since Westlake Institute for Data Intelligence does not own the phone numbers of confirmed cases and the temporal contact graph does not contain any private information, we do not have the authority issue. We agree that the phone number is a unique identifier. However, it was well protected in our research, since the phone numbers of confirmed cases would be hashed before they were used on the secure data servers. Furthermore, such information was securely stored on the data servers of local institutions and it was not available for us. We merely utilized the temporal contact graph to obtain relevant results in our current research. This is what we meant by saying "did not include identified individual-level data". As mentioned above, both the personal phone number data from Westlake Institute for Data Intelligence and those from the local institutions were preprocessed by the same pseudonymization mechanism (i.e., hash function). Besides, the final output (i.e., the temporal contact graph) was further pseudonymized to avoid data restoration. Thus, the hashed phone numbers do not allow individuals to be identified as covid-19 patients or individually identified. In a word, there is no private identifying information about the individuals accessible to us (researchers) and no interaction (or intervention) between the individuals and us. In addition, we note that the whole procedure is amenable to laws in both China and EU as follows. Article 1038 of the Civil Code of the People's Republic of China, stipulates that "Information pro-

cessors shall not disclose or tamper with the personal information they collect or store; and without the consent of the data subject, information processors shall not illegally provide any other person with his personal information, **except for the processed information cannot be identified with any specific person and restored.**" Article 89 of the General Data Protection Regulation (GDPR) enforced by the European Union stipulates that "Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject. Those safeguards shall ensure that technical and organizational measures are in place in particular in order to ensure respect for the principle of data minimization. **Those measures may include pseudonymization provided that those purposes can be fulfilled in that manner. Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner.**"

Another key point in the debate on the contact tracing apps is about the concerns that the contact tracing apps may have privacy leakage issues and the methods (centralized V.S. decentralized) used to inform the potential contacts with infectious cases about their risk need further investigation. For this issue, Google recently showed that the privacy can be protected by using Bluetooth with a proper privacy-preserving protocol[16]. Also, there are some other works on this direction[17–19]. Our work here studies the contact analysis in the transmission, and thereby does not discuss the privacy issues. Note that our approach falls within the category of the centralized way of informing the potential contacts. Though both centralized and decentralized ways can identify individuals having contacts with infectious cases, our results demonstrate that centralized way can provide an abundance of information that can be helpful for prevention and control of COVID-19. As indicated by the title of our article, our main focus is to reveal the evolving situation of COVID-19 by exploiting the temporal contact graph. The study of the effects that an app will have on people's behaviour is another important topic. There are also some works on this direction[20,21]. This topic is a new and independent topic and therefore was not considered in this work.

Supplementary References

1. Kraemer, M. U. *et al.* The effect of human mobility and control measures on the covid-19 epidemic in china. *Science* **368**, 493–497 (2020).

2. Jia, J. S. *et al.* Population flow drives spatio-temporal distribution of COVID-19 in china. *Nature* (to appear).

3. Aleta, A. *et al.* Modeling the impact of social distancing, testing, contact tracing and household quarantine on second-wave scenarios of the covid-19 epidemic. *medRxiv* (2020).

4. Lauer, S. A. *et al.* The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine* **172**, 577–582 (2020).

5. Sohrabi, C. *et al.* World health organization declares global emergency: A review of the 2019 novel coronavirus (covid-19). *International Journal of Surgery* (2020).

6. Li, Q. *et al.* Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England Journal of Medicine* (2020).

7. Linton, N. M. *et al.* Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. *Journal of clinical medicine* **9**, 538 (2020).

8. Bi, Q. *et al.* Epidemiology and transmission of covid-19 in 391 cases and 1286 of their close contacts in shenzhen, china: a retrospective cohort study. *The Lancet Infectious Diseases* (2020).

9. Cao, M. *et al.* Clinical features of patients infected with the 2019 novel coronavirus (covid-19) in shanghai, china. *MedRxiv* (2020).

10. Chen, J. *et al.* Clinical progression of patients with covid-19 in shanghai, china. *Journal of Infection* (2020).

11. Cheng, Y. *et al.* Kidney disease is associated with in-hospital death of patients with covid-19. *Kidney international* (2020).

12. Clauset, A., Shalizi, C. R. & Newman, M. E. Power-law distributions in empirical data. *SIAM review* **51**, 661–703 (2009).

13. Centers for Disease Control and Prevention. Contact tracing for covid-19. https://www.cdc.gov/coronavirus/2019-ncov/php/contact-tracing/contact-tracing-plan/contact-tracing (2020).

14. Firth, J. A. *et al.* Combining fine-scale social contact data with epidemic modelling reveals interactions between contact tracing, quarantine, testing and physical distancing for controlling covid-19. *medRxiv* (2020).

15. Morley, J., Cowls, J., Taddeo, M. & Floridi, L. Ethical guidelines for covid-19 tracing apps (2020).

16. Gvili, Y. Security analysis of the covid-19 contact tracing specifications by apple inc. and google inc. *IACR Cryptol. ePrint Arch.* **2020**, 428 (2020).

17. Singapore Government. Tracetogether, safer together. https://www.tracetogether.gov.sg/ (2020).

18. Australia Government Department of Health. Covidsafe app. https://www.health.gov.au/resources/apps-and-tools/covidsafe-app (2020).

19. National Cyber Security Centre. Nhs covid-19: the new contact-tracing app from the nhs. https://www.ncsc.gov.uk/information/nhs-covid-19-app-explainer (2020).

20. There are many reasons why COVID-19 contact-tracing apps may not work. Adam vaughan. https://www.newscientist.com/article/2241041/ (2020).

21. Hinch, R. *et al.* Effective configurations of a digital contact tracing app: A report to nhsx. *en. In:(Apr. 2020). Available here. url: https://github. com/BDI-pathogens/covid-19_instant_tracing/blob/master/Report* (2020).