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

Mincheng Wu
Zhejiang University
Chao Li
Zhejiang University
Zhangchong Shen
Zhejiang University
Shibo He (s18he@zju.edu.cn)
Zhejiang University
Lingling Tang
Zhejiang Shuren University
Jie Zheng
Zhejiang Institute of Medical-care Information Technology
Yi Fang
Westlake Institute for Data Intelligence
Kehan Li
Zhejiang University
Yanggang Cheng
Zhejiang University,
Zhiguo Shi
Zhejiang University,
Guoping Sheng
Zhejiang Shuren University
Yu Liu
Westlake Institute for Data Intelligence
Jinxing Zhu
Westlake Institute for Data Intelligence
Xinjiang Ye
Westlake Institute for Data Intelligence
Jinlai Chen
Westlake Institute for Data Intelligence
Wenrong Chen
Westlake Institute for Data Intelligence
Lanjuan Li

Zhejiang University

Youxian Sun

Zhejiang University

Jiming Chen

Zhejiang University

Article

Keywords: Digital, prevention, level, infected

Posted Date: October 18th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-31777/v2

Version of Record: A version of this preprint was published at Communications Physics on November 4th, 2022. See the published version at https://doi.org/10.1038/s42005-022-01045-4.

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.*

[*] To whom correspondence should be addressed. E-mail:s18he@zju.edu.cn, ljli@zju.edu.cn, cjm@zju.edu.cn. [†] These authors contributed equally: Mincheng Wu, Chao Li.

Abstract

Digital contact tracing has been recently advocated by China and many countries as part of digital prevention measures on COVID-19. Controversies have been raised about their effectiveness in practice as it remains open how they can be fully utilized to control COVID-19. In this article, we show that an abundance of information can be extracted from digital contact tracing for COVID-19 prevention and control. Specifically, we construct a temporal contact graph that quantifies the daily contacts between infectious and susceptible individuals by exploiting a large volume of location-related data contributed by 10,527,737 smartphone users in Wuhan, China. The temporal contact graph reveals five time-varying indicators can accurately capture actual contact trends at population level, demonstrating that travel restrictions (e.g., city lockdown) in Wuhan played an important role in containing COVID-19. We reveal a strong correlation between the contacts level and the epidemic size, and estimate several significant epidemiological parameters (e.g., serial interval). We also show that user participation rate exerts higher influence on situation evaluation than user upload rate does. At individual level, however, the temporal contact graph plays a limited role, since the behavior distinction between the infected and uninfected contacted individuals are not substantial. The revealed results can tell the effectiveness of digital contact tracing against COVID-19, providing guidelines for governments to implement interventions using information technology.

COVID-19, caused by SARS-CoV-2, has rapidly spread to most of the countries in the past year. As of 11 AM CEST, 6 October, 2021, world-wide confirmed cases of COVID-19 has reached 235,175,106, among which 4,806,841 patients died[1]. It has been overwhelming the medical systems of many countries with large case counts and threatening to infect an extremely large population, but it is still too early to tell its disappearance[2]. Currently, many countries (e.g., the U.S.A., the U.K., Australia, etc.) have been cooperating together to prevent and control such an unprecedented disease via a variety of ways[3–5].

As is known, contact tracing is one of the most effective ways to find the high-risk individuals who may be infected, while it costs the expense in effort, time, and financial. Recently, digital contact tracing using information technology has been widely advocated to replace traditional labor-intensive contact surveys[5–7]. The main idea is to exploit Bluetooth/positioning sensors on smartphones to discover nearby devices held by users and identify the contacts with the infectious individuals[8,9]. On one hand, about 28 countries such as China, Switzerland, Spain, the United Kingdom, Australia, Singapore and Germany have implemented various measures using informa-

tion technology (e.g., launching digital contact tracing apps)[10–14]. On the other hand, however, recent works have revealed that digital contact tracing contributes little to contain outbreaks, principally because of low participation rates and low engagement of participants[15,16]. As many controversial issues of digital contact tracing have been raised, it is urgent to review empirical evidence for the effectiveness of this measure against a pandemic spreading from different aspects[17–19].

In this article, we take a pioneering and in-depth investigation into this issue, and show that an abundance of information can be extracted from digital contact tracing for COVID-19 prevention and control. We construct a temporal contact graph (Fig. 1a and 1d) that quantifies the daily contacts between infectious and susceptible individuals by exploiting a large volume of location-related data contributed by more than 10,527,737 smartphone users in Wuhan, China. We demonstrate that such a temporal contact graph has many applications, e.g., to estimate the individual-level contact trend, analyze the dynamic contact behavior (Fig. 1b), identify the potential infected contacted individuals (Fig. 1c), estimate the possible number of confirmed cases in the near future (e.g., cases in the next week) (Fig. 1e), and assist the decision-making of control measures. This is different from previous studies which focused on integrating mathematical models and available statistical data of confirmed cases to characterize the transmission of epidemic diseases[20–36], or those which utilized individual mobility traces (with the information of confirmed cases) to simulate the spreading process[7,37], opening up a new perspective to understand the spreading of COVID-19 from the aspect of digital contact tracing.

Since contact tracing measures are essentially based on crowdsourcing[38,39], their performance highly relies on the involvement of voluntary smartphone users. Due to potential privacy leakage and cost incurred during crowdsourcing process, voluntary users are reluctant to participate and contribute their personal data at a fine-grained scale[38,39]. It is, therefore, challenging to fully utilize sparse and noisy crowdsourced data of contact information from voluntary users to capture the intrinsic transmission characteristic of COVID-19. In this article, we introduce five time-varying indicators that are validated to have the capability of accurately capturing actual contact trends at individual and population level in Wuhan, providing a data-driven evidence that the travel restrictions in Wuhan significantly reduced the chance of susceptible individuals having contacts with the infectious and thus played an important role in containing COVID-19. We reveal a strong correlation (Pearson coefficient 0.77) between the number of daily symptomatic cases and daily total contacts with a 12-days delay, and estimate several significant epidemiological parameters such as the serial interval. We study the effect of user involvement on the effectiveness of digital contact tracing measures, finding that user participation rate exerts higher influence on

situation evaluation than user upload rate does. We also show that the distinction of contact behaviors between the infected and uninfected contacted individuals are not prominent, which is more substantial than the sex character but less substantial than the age characteristics. By designing an infection risk evaluation framework, the area under the receiver operating characteristic curve reaches 0.57, indicating it only performs a limited role in identifying high-risk contacted individuals. This indicates that it is not highly effective to narrow down the search of high risk contacted individuals for quarantine by the distinction of contact behaviors. The empirical results can offer a

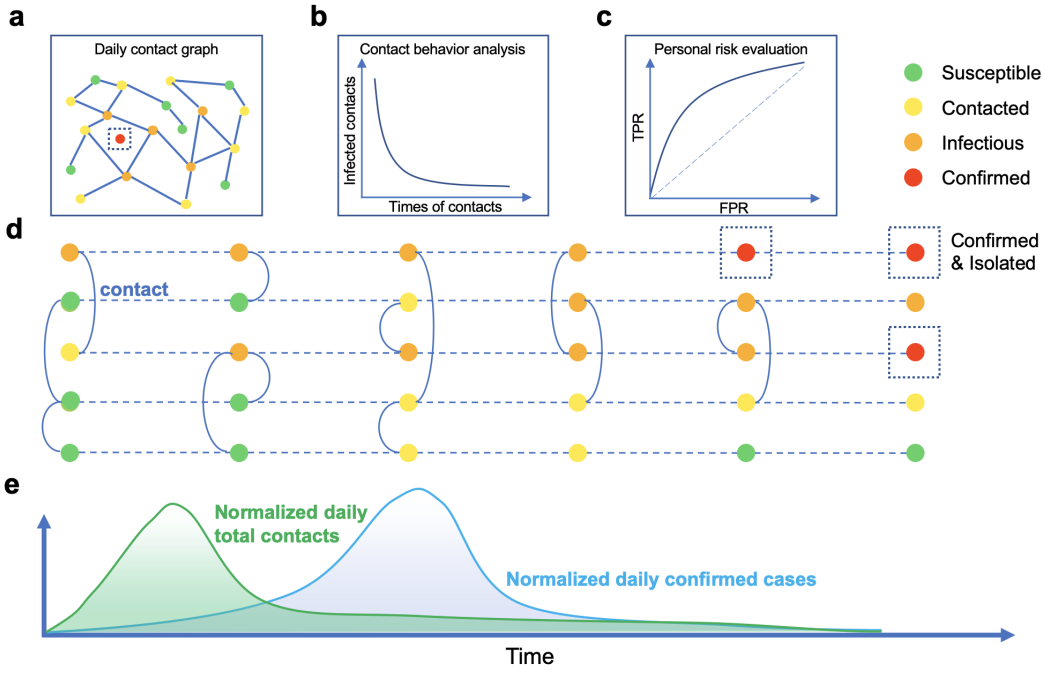

Figure 1: **Temporal contact graph and schematics for its potential applications.** An individual has four status: susceptible, contacted, infectious and confirmed. The status 'susceptible' turns to 'contacted' when an individual had at least one contact with infectious individuals. A contacted individual may be infected or stay healthy. The status 'infectious' changes to 'confirmed' when confirmation is made. A confirmed case will be quarantined for treatment in China and no longer infectious to others. **a.** Daily contact graph. **b.** The analysis for contact behaviors shows the distributions of contact counts between infected and uninfected contacted individuals. **c.** The personal risk evaluation based on contact behaviors. **d.** Contact history and status of individuals. A node denotes an individual and different colors indicate different status. A dashed line means the status evolution of a single individual in timeline, and a solid curve between two individuals means a contact. **e.** The correlation between normalized daily total contacts and daily confirmed cases.

promising way to evaluate and predict the evolving epidemic situation of COVID-19, and provide guidelines for governments to implement digital contact tracing measures.

Results

Characteristics of informative indicators. We leverage a large volume of location-related data set contributed by 10,527,737 smartphone users in Wuhan, China. Each item in the data set includes a geohash encoded meshed area, a timestamp and an anonymized identity. We build a contact model, in which a contact between two individuals is said to occur when they are reported within a certain spatial area and a temporal interval (see the Method section for more details). By collaborating with local authority, we obtain the information whether and when each anonymous individual was confirmed. With the information of 16,647 confirmed cases and the contact model we build, we identify 562,280 contacted individuals, with which we are able to construct a temporal contact graph (consisting of 3.7 million contacts) between infectious and susceptible individuals. We introduce five informative indicators (t denotes a day): 1) $C(t)$, the daily total number of contacts between infectious and contacted individuals, i.e. the number of edges in the constructed temporal contact graph; 2) $S(t)$, the daily number of infectious individuals who had encountered with contacted individuals at least once, i.e. the number of infectious nodes; 3) $N(t)$, the daily number of contacted individuals who had encountered the infectious at least once, i.e. the number of susceptible nodes; 4) $L(t)$, the daily average contacts of infectious individuals associating with contacted individuals, i.e., the average degree of infectious nodes, and 5) $K(t)$, the daily average contacts of contacted individuals associating with infectious individuals, i.e., the average degree of susceptible nodes.

The five indicators at the beginning of 2020 are shown along with a series of implements (Fig. 2a). The daily total contacts between infectious and susceptible individuals $C(t)$ can reflect the potential transmission. We find that $C(t)$ increased dramatically first from 4 to 20 January, 2020, due to the fast increasing infectious individuals, and then dropped after 20 January. As we know, the Chinese authority announced the outbreak of COVID-19 and confirmed its infection among people on 20 January, which explains the decline of $C(t)$. Obviously, $C(t)$ decreased sharply around 23 January when the lockdown was implemented in Wuhan, and tended to zero around 28 February.

From a macroscopic view, $N(t)$ and $S(t)$ describe population-level contacts trend in Wuhan. Notice that $N(t)$ had a minor bouncing back after 26 January, 2020, which is possibly due to

the number of confirmed cases quickly increased after 23 January, and people in Wuhan could still move within the city (their mobility increased due to the approaching of Chinese New Year). Then, $N(t)$ began to decline on 4 February, and approached zero around 28 February. Compared to $N(t)$, however, $S(t)$ performs a different characteristic from the other ones. Initially, $S(t)$ quickly

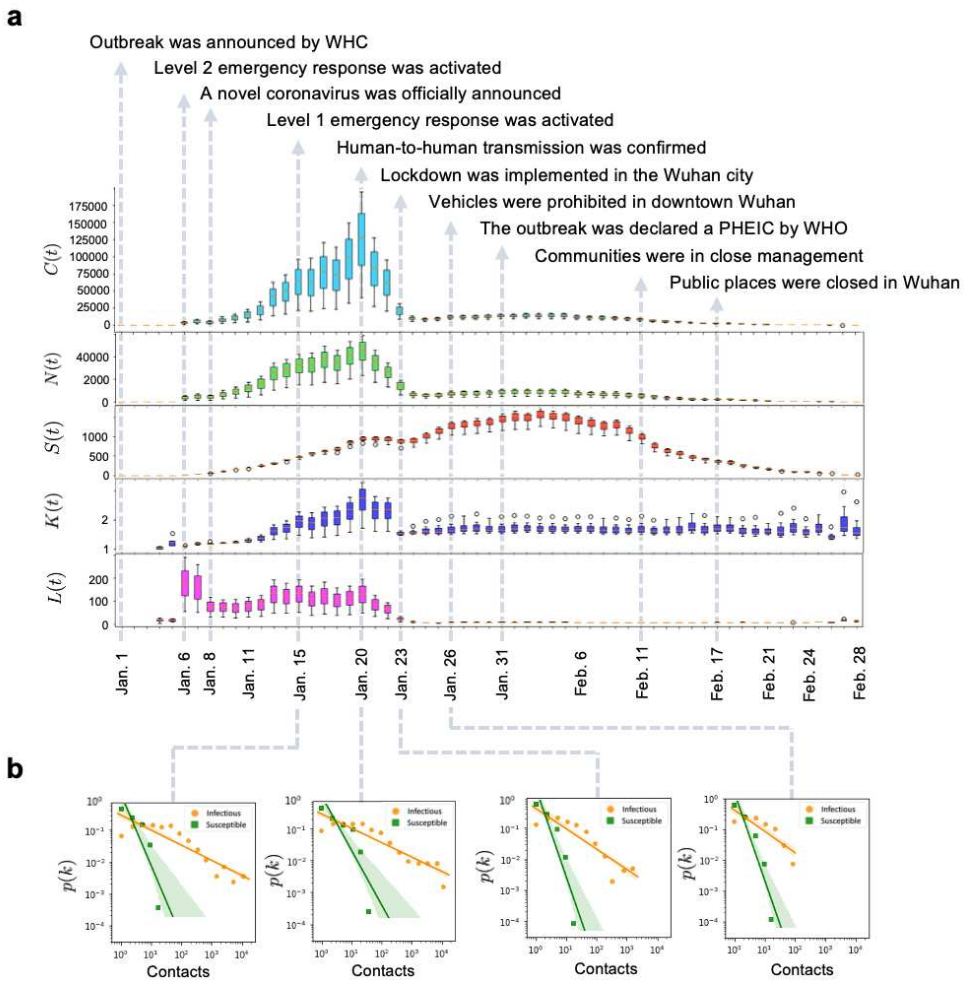

Figure 2: **Daily characteristics of five indicators. a.** $C(t)$, the daily total number of contacts between infectious and contacted individuals. $N(t)$, the daily total number of contacted individuals who had encountered the infectious at least once. $S(t)$, the daily total number of infectious individuals who had encountered with contacted individuals at least once. $K(t)$, the daily average number of infectious individuals that each susceptible individual encountered. $L(t)$, the daily average number of susceptible individuals that each infectious individual contacted. **b.** The distributions of the daily number of contacts by all contacted individuals and the distributions of the daily number of contacts by all confirmed cases on four specific days.

increased with the number of confirmed cases as few of them are under quarantine. It began to drop on 20 January upon the official announcement and reached the local minimum on 23 January, after which it had a duration of increase. It decreased again on 3 February and eventually approached to zero around 28 February. The main reason is that the confirmed cases increased fast after 20 January and the chance of meeting an infectious individual remained high as many of them were not hospitalized due to test capacity constraint.

Further evidence can be observed from the indicators $K(t)$ and $L(t)$. $K(t)$ performed a similar behavior as $N(t)$, while $L(t)$ displays a more distinct fluctuation in the early January, 2020, since the infected are not isolated, and they contacted the susceptible as usual in the incubation period. On account of the small proportion of the infected and the randomness of their movements, the two indicators were not not stable during 6 to 20 January. For example, they first dropped a bit around 10 January, which may be due to the mobility reduction caused by the sudden drop of temperature (Supplementary Fig. 4). The dynamic $K(t)$ and $L(t)$ accurately describe the actual individual-level contacts trend in Wuhan, providing data-driven evidence that travel restrictions in Wuhan significantly reduced the chance of a susceptible individual having contacts with the infectious individuals and thus played an important role in containing COVID-19.

From the perspective of the infectious, the distribution of daily contacts also follows a power-law distribution (Fig. 2b), and has a prominent long tail especially when the exponent coefficient is small before 23 January. The long tails indicate that there were some super active cases who had contacted with hundreds of susceptible individuals. Identifying and quarantining them helps mitigate the fast transmission. From 15 to 26 January, specifically, the power exponent increased first and then decreased, having the same evolving pattern as $K(t)$. Therefore, $C(t)$, $N(t)$, $S(t)$, $K(t)$ and $L(t)$, characterized the spread of COVID-19 from dimensions of susceptible individuals, confirmed cases and overall contacts, which were informative for COVID-19 prevention and control (see Supplementary Fig. 7 for more sensitivity analyses).

A strong situation correlation revealed by digital contact tracing. The temporal contact graph shows the potential group of contacted individuals at high infection risk. Intuitively, more contacts between infectious and contacted individuals are likely to cause more confirmed cases in the future. We proceed to investigate the correlation between the daily number of contacts $C(t)$ and the symptomatic cases reported by authority[40].

The curves of daily number of contacts $C(t)$ (in blue) and daily symptomatic cases (in red)

with normalization (i.e., normalized by the maximum) in Wuhan are shown in Fig. 3a, from which we observe a prominent delay between them. By moving points in the time series of daily number of total symptomatic cases ahead (in yellow), these two curves present more similar trends. To find the proper delay that results in the best similarity in trends between the curves of daily number of contacts and confirmed cases, we alter the delays ranging from 0 to 17 days according to existing surveys[41,42]. The experiments show that a 12-days delay results in the best Pearson correlation of 0.77 (Fig. 3b) in accordance with recent works[42–49]. As for the cumulative correlation analysis, the curves of cumulative contacts (in blue) and cumulative symptomatic cases (in red) with normalization (i.e., normalized by the maximum) in Wuhan are shown in Fig. 3c, where we find a strong correlation between the number of cumulative contacts and the cumulative confirmed cases with 12 days ahead (in yellow). The delay from being contacted to symptom onset may vary for different individuals, while analyzing the cumulative correlation would weaken these variations, reaching a higher Pearson correlation. Specifically, the Pearson correlation reaches 0.99 when there is a 12-days delay between normalized cumulative contacts and normalized cumulative symptomatic cases (Fig. 3d). Since the correlation between cumulative contacts and cumulative symptomatic cases is higher than that between daily contacts and daily symptomatic cases. Thus, the number of cumulative contacts can reflect and estimate the number of symptomatic cases with higher accuracy, having a better predictability of the number of symptomatic cases than the number of daily contacts does. In summary, indicator $C(t)$ provides a new way to evaluate and predict the epidemic situation of COVID-19.

Furthermore, we also explore several significant epidemiology parameters including the contacting period, incubation period, and serial interval (Fig. 3e). Specifically, the contacting period indicates the interval from the first possible contact to the last possible contact, which is estimated to be 2.3 days (95$\%$CI, 0.4 to 6.7 days) (Fig. 3f). The incubation period indicates the interval from the last possible contact to symptom onset, which is estimated to be 7.3 days (95$\%$CI, 1.2 to 14.1 days) (Fig. 3g). The serial interval indicates the interval from symptom onset of A to symptom onset of B who is infected by A, which is a proxy of generation period from the infection of A to the infection of B who is infected by A. Notice that the serial interval could be negative because of asymptomatic transmissions, and it is estimated to be 2.5 days (95% CI, -9.2 to 13.9 days) (Fig. 3h). These estimations are in accordance with most existing survey[42–49], demonstrating the effectiveness of revealing epidemic situation at population level by digital contact tracing.

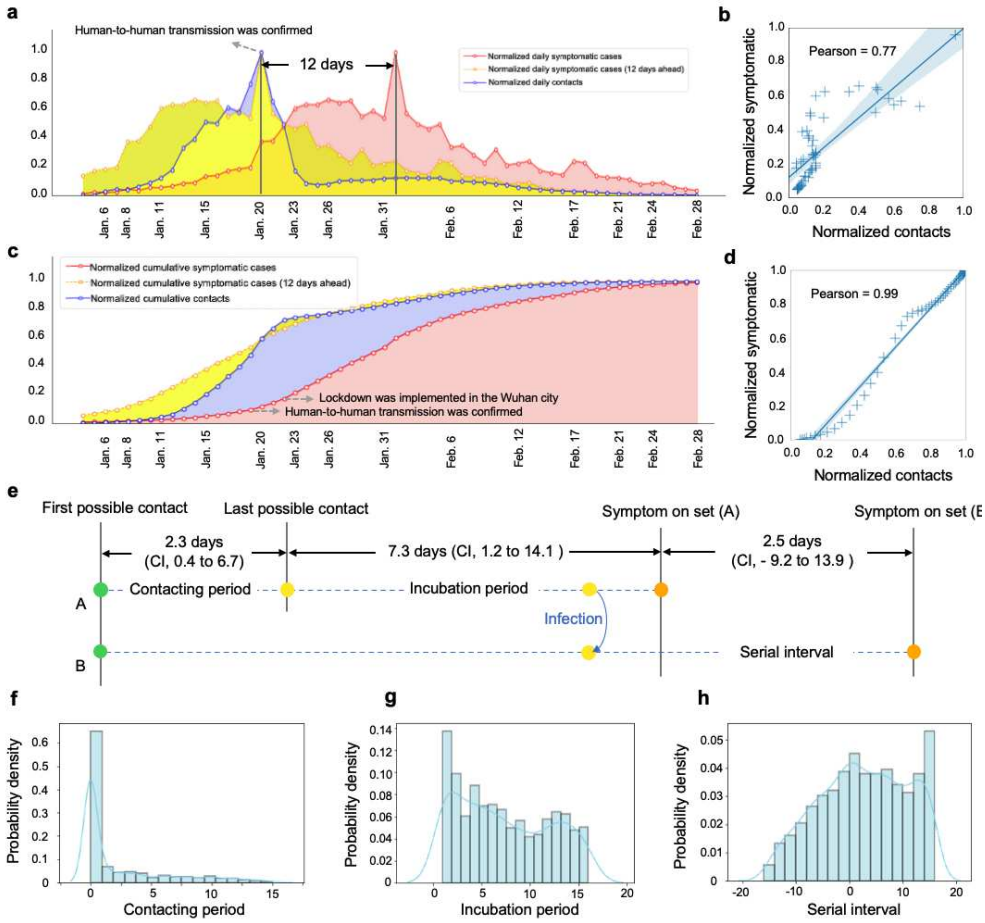

Figure 3: **Daily and cumulative correlation analysis. a.** Historical time series of the number of daily contacts (in blue), daily reported symptomatic cases in Wuhan (in red) and daily reported symptomatic cases ahead 12 days (in yellow). **b.** The reached maximum Pearson correlation (0.77) between normalized daily contacts and normalized daily confirmed cases with a 12-days delay. **c.** Historical time series of the number of cumulative contacts (in blue), cumulative symptomatic cases in Wuhan (in red) and cumulative reported symptomatic cases with a 12-days delay (in yellow), where the Pearson correlation reaches 0.99. **d.** The reached Pearson correlation (0.99) between normalized cumulative contacts and normalized cumulative symptomatic cases with a 12-days delay. **e.** The timeline displays the contact period, incubation period, and serial interval inferred by digital contact measure. **f.** The distribution of the duration from the first possible contact to the last possible contact (mean 2.3 days, 95%CI 0.4 to 6.7 days). **g.** The distribution of the duration from the last possible contact to symptom onset (mean 7.3 days, 95%CI 1.2 to 14.1 days). **h.** The distribution of the duration from the symptom onset of A to the symptom onset if B who is infected by A (mean 2.5 days, 95%CI -9.2 to 13.9 days).

The impacts of user involvement on the contact tracing performance. Clearly, digital contact tracing is based on crowdsouring. Individual smartphone users are voluntary to participate in the process and upload their contact information. It remains open to tell how the performance of contact tracing (e.g., estimating $K(t)$ and $L(t)$ and daily confirmed cases) is affected by user involvement, raising the question on whether contact tracing measures can really work in practice. We study on this issue by taking into account two types of user involvement: user participation rate (the proportion of users in the whole population) and data uploading rate (their data reporting frequency per day). To simulate user involvement, we randomly choose $\alpha\%$ users as the voluntary users, and $\alpha\%$ data items each participating user uploading per day, and evaluate the corresponding performance loss.

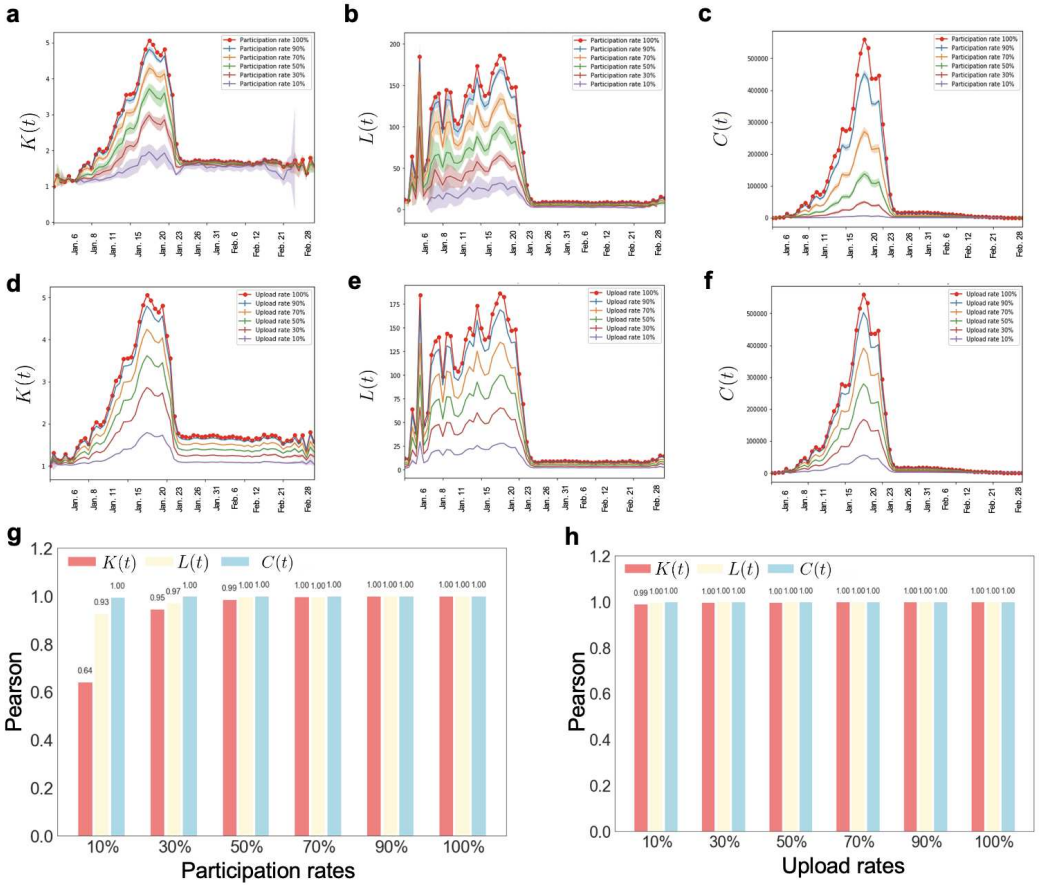

Figure 4: **The Performance of contact tracing under different user involvements. a-c.** Three figures show the change of daily $K(t)$, $L(t)$, and $C(t)$ from Jan. 1st to Feb. 28th with error bars Vs. different user participation rates. **d-f.** Three figures show the change of daily $K(t)$, $L(t)$, and $C(t)$ from Jan. 1st to Feb. 28th with error bars Vs. different user upload rates. **g-h.** The Pearson correlations Vs. different user participation rates and user upload rates.

We conduct extensive explorations by varying the values of α, and repeat ten times of Monte Carlo experiments at each involvement level to make our experiments more credible. At a specific α, we plot the time series with error bars of $K(t)$, $L(t)$ and total contacts $C(t)$ for both scenarios of user participation rate and user upload rate, ranging from Jan. 1st, 2020 to Feb, 28th, 2020. It is shown that, as α decreases, corresponding time series decrease with the similar trend (Fig. 4a-4f). This is expected as reduction in either user participation rate or user upload rate decreases the chances of having contacts among users. To see if the reduction has influence on capturing the evolving trends, we calculate the Pearson correlations between the time series under $\alpha\%$ and full (100%) participation rate/data upload rate case (Fig. 4g and 4h).

We get the following observations. 1) Decreasing the user upload rate or participation rate results in the lower values of $K(t)$, $L(t)$ and $C(t)$. 2) User participation rate and data upload rate have minor effects on the evaluation of evolving pattern of $C(t)$, whose error bars are not as obvious as another two variables. The above observations indicate that $C(t)$ is more robust than $K(t)$ and $L(t)$ when user involvement changes. 3) $K(t)$ is more sensitive to the change of user involvement α than $L(t)$. This is because the number of susceptible individuals is much larger than that of the infectious. 4) User participation rate exerts higher influence on the three indicators than user upload rate does according to Fig. 4g and Fig. 4h. Therefore, we should encourage more user participation to obtain a better performance in practice. Considering their privacy and cost concerns, it would be a good strategy to allow voluntary smartphone users having a relatively low data upload rate. 5) For the participation rates analysis, when the participation rate reduces to 10%, the correlation coefficient reduces significantly according to Fig. 4g, which can be attributed to the characteristics of the overall power-law distribution of the network, which 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 entire network. We note that the performance of individual-level infection risk evaluation will be impacted when user participation rate or upload rate drops since we may miss many contacts with infectious cases in such case and make an incorrect evaluation.

Individual-level infection risk evaluation by contact behavior discrimination. In a spreading process, contacted individuals have chance of being infected, or staying healthy. We proceed to study the contact behaviors between the infected and uninfected contacted individuals, based on which we can obtain an individual-level infection risk evaluation. We count the number of contacts each contacted individual had with the infectious in recent 17 days, i.e., the infectious period (see Supplementary Information for more sensitivity analysis), and calculate the probability $p(k)$ that a contacted individual had k contacts for infected and uninfected contacted individuals, respectively. Power-law behaviors are found for both types, while the parameters are mildly different. The infected contacted individuals have a power-law distribution with an average $< k >= 5.93$ and an exponent $\gamma = 1.66$, while the uninfected contacted individuals have a power-law distribution with $< k >= 5.38$ and $\gamma = 1.81$ (Fig. 5a).

Further, we count the number of infectious individuals who had contacts with any contacted individual in recent 17 days, and calculate the probability $p(k)$ that a contacted individual have associated k infectious individuals for infected and uninfected contacted individuals, respectively. The infected contacted individuals follows a power-law distribution with $< k >= 3.95$ and $\gamma = 1.33$, and the uninfected contacted individuals follows a power-law distribution with $< k >= 2.89$ and $\gamma = 1.79$ (Fig. 5b). We count the number of days when contacted individuals had contacts with any infectious individual. The probability $p(k)$ that a contacted individual have encountered any infectious individual for k days in recent 17 day for infected and uninfected contacted individuals, respectively. The infected contacted individuals follows a power-law distribution with $< k >= 2.27$ and an exponent $\gamma = 1.94$, while the uninfected contacted individuals follows a power-law distribution with $< k >= 2.03$ and $\gamma = 2.22$ (Fig. 5c). These distributions are different in terms of the expectations and the power exponents: the infected contacted individuals have more contacts than uninfected contacted individuals and the corresponding distribution has a fatter tail. This indicates that there are an appreciable quantity of infected contacted individuals with a large amount of contacts.

Based on these contact behavior discriminations, we proceed to perform an individual-level infection risk evaluation for each contacted individual. We propose a risk evaluation method based on the Bayesian framework by calculating the posterior infected probability for any contacted individual[50]. We first introduce a variable z_j to represent the health status for contacted individual j, i.e., $z_j = 1$ if j is infected and $z_j = 0$ otherwise. Then, the infection risk for j is determined by the posterior probability $P(z_j = 1|B_j, F_j)$:

$$P(z_j = 1|B_j, F_j) = \frac{P(B_j, F_j|z_j = 1) \cdot P(z_j = 1)}{P(B_j, F_j)}, \quad (1)$$

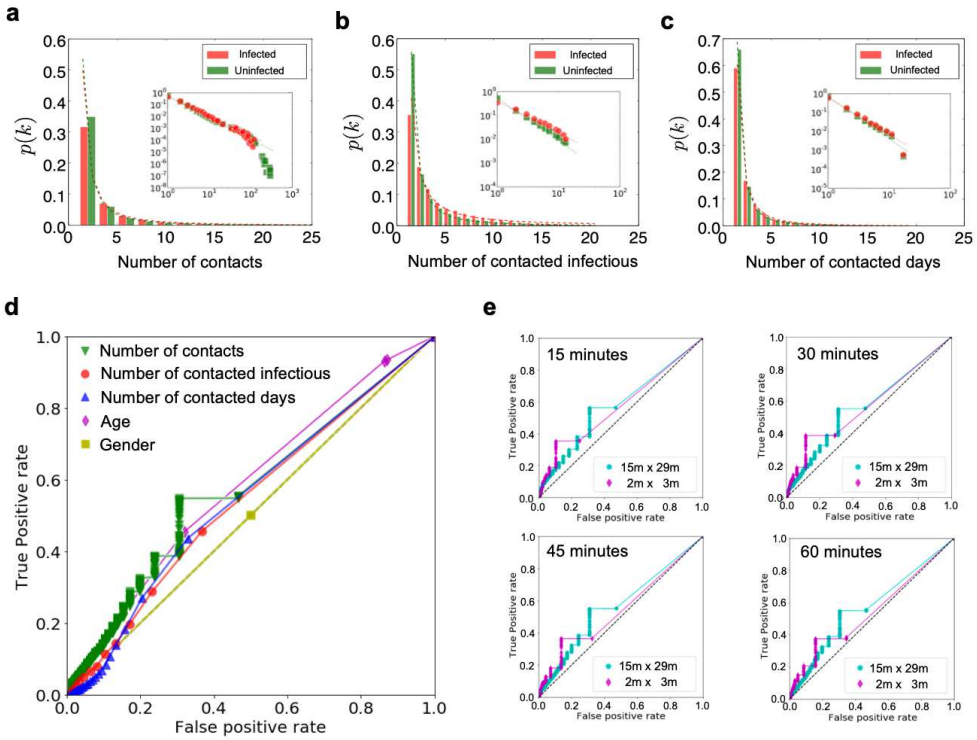

Figure 5: **Infection risk evaluation based on the Bayesian framework. a.** The distributions of the times of contacts with the infectious by infected and uninfected contacted individuals, respectively. **b.** The distributions of the numbers of the infectious by infected and uninfected contacted individuals, respectively. **c.** The distributions of the days of contacts with the infectious by infected and uninfected contacted individuals, respectively. **d.** The ROC curves for the risk evaluation. Here the x-axis denotes the false positive rate and the y-axis denotes the true positive rate, where a random guess gives a point along the dashed diagonal line. **e.** The ROC curves for the risk evaluation with different temporal and spatial granularities.

where B_j denotes the contacts for j, and F_j denotes the individual feature, indicating the age, gender, and etc. The term $P(B_j, F_j|z_j = 1)$ is the likelihood, and $P(z_j = 1)$ indicates the infected probability for any contacted individual j a prior, which is taken as a constant (see the Methods section for more details).

After calculating the infection risk of every contacted individual, we vary the positive threshold from 0 to 1 and display the ROC (receiver operating characteristic) curve. The ROC space is defined by plotting the false positive rate in x-axis and the true positive rate as y-axis, indicating the relative trade-offs between false positive (costs) and true positive (benefits) (Fig. 5d). Increasing the threshold results in fewer true positives and false positives. However, the true positive is

larger than false positives, indicating the infection risk model is effective. Above an appropriate threshold, for example, we can find about 50% of the infected contacted individuals with 30% false report of the uninfected contacted individuals, where the AUC (area under the ROC curve) reaches 0.57 by using the contact graph. The feature of gender did not contains any information to distinguish infected ones, where the AUC is 0.5, while the AUC with the feature of age reaches 0.59. Generally, a high AUC can help narrow down high risk contacted individuals for quarantine in practice. Obviously, information of the age provides a more accurate discrimination to identify the infected contacted individuals, while there is nearly no distinction by gender. The results indicate that the distinction of contact behaviors between the infected and uninfected contacted individuals are not prominent, which is more substantial than the sex character but less substantial than the age characteristics. To perform a sensitivity analysis for the temporal and spatial granularities, we vary the time interval and spatial area in the contact model. Specifically, the time interval is ranging from 15 minutes to 60 minutes, and the contact distance is randing from an area of $2m \times 3m$ to $15m \times 29m$. The ROC curves shows the parameters are not sensitive, indicating a stable analytical result (Fig. 5e).

Discussions

Since the emergence of COVID-19, researchers have proposed many mathematical models to characterize the transmission of COVID-19[20–23,51]. As digital contact tracing has been advocated by many countries, it rises the pressing issue of how to fully utilize such a new approach to contain COVID-19. Here, we provide the first collection of results that accurately characterize the evolving epidemic situation of COVID-19 by exploiting the temporal contact graph. Our approach offers a new data-driven approach to evaluate and predict the evolving epidemic situation of COVID-19. Clearly, our data-driven approach and the traditional model-based approaches are complementary to characterize the transmission of COVID-19.

As the contact tracing data are still unavailable, their performance on COVID-19 prevention and control can not be directly evaluated. Some previous studies utilized the mobility data of smartphone users to capture the contacts and simulate the infection process due to the unavailability of user infection status[2,7]. Here, we leverage a large amount of location-related data contributed by 10,527,737 voluntary users to study such an issue. As we know the health status of smartphone users, we construct a temporal contact graph between susceptible and infectious individuals, which can be directly used to characterize the transmission of COVID-19. This distinguishes our work from most of the previous studies. We show that we can obtain a good performance in estimating and evaluating the epidemic situation even when user participation rate and data upload rate are low. We also demonstrate that user participation rate has a bigger impact than data upload rate on the estimations of the proposed indicators. Our results can provide guidelines for governments to practically deploy digital contact tracing measures.

Methods

Method I: Data Collection and Contact Model The data are contributed by 10,527,737 voluntary users in Wuhan, China, and collected by crowdsourcing platforms from our industry partners. The location-related information was authorized and uploaded every time smartphone users are using location-based services. Privacy protection mechanisms such as perturbation and pseudonymization are adopted during data collection. The location-related information, including POI, GPS, geomagnetic, etc., is projected into meshed area. The confirmed cases from 18 January to 28 February, 2020, serve as the sources of the infection. They are linked to the status of smartphone users by their phone number, which is validated by the local authorities.

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. This article only utilizes the temporal contact graph that is derived from the raw data.

We propose a contact model based on the crowdsourced dataset: a contact between two smartphone users is said to occur when they report the identical geohash within a given time interval. As aforementioned, the geohash can be projected into a mesh area of a certain meshed area (e.g., $15 \times 29m^2$). This means that a contact is characterized when the distance between two smartphone users is within 18 meters averagely. Such a definition is similar to that adopted by most contact tracing apps which exploit Bluetooth or GPS to decide a contact when two users are in a short distance. As smartphone users report data in a very low and irregular frequency, the contributed data are typically sparse. We would miss many contacts if we only count those where two smartphone users are reporting identical information simultaneously. Considering the data sparsity, we define a contact occurring when two users upload the same geohash with time interval T. We vary T from fifteen minutes to two hours and perform the sensitivity analyses (see Supplementary Fig. 7 for more details). In the article, we present the results when T equals to two hours for an illustration.

Method II: The construction of the Temporal Contact Graph An individual has four status: susceptible, contacted, infectious and confirmed. The status 'susceptible' turns to 'contacted' when an individual had at least one contact with infectious individuals. A contacted individual may be infected or stay healthy. The status 'infectious' changes to 'confirmed' when confirmation is made.

A confirmed case will be quarantined for treatment in China and no longer infectious to others.

Recent results indicated that an infected individual can turn to infectious before and after the symptom onset, known as pre-symptomatic transmission and symptomatic transmission. Taking into account both types of transmission, we define the infectious period from the time when an infected individual becomes infectious to the time when he/she is removed (recovered or quarantined for treatment). We analyze the range of this period, finding that 17 days is the best choice (see Supplementary Figs. 8 and 9 for sensitivity analyses).

By using the contact model, we identify 562,280 susceptible individuals having contacts with 16,647 infectious individuals who turn to confirmed status later. The daily temporal contact graph is constructed as a temporal undirected weighted bipartite graph where the vertices represent contacted susceptible individuals or infectious individuals and the weight represents the number of contacts between them in a single day. An illustration is provided in Fig. 1. This bipartite temporal graph is used in all the analysis in this article.

Method III: Bayesian Framework We calculate the posterior probability $P(Z|B,F)$ under the Bayesian framework, where we denote the behavior events by B and denote the feature events by F. Specifically, $b_j^{(u)}$ indicates the times of contact u for any contacted individual j, and $f_j^{(v)}$ indicated the category of feature v for j. To measure the infection risk of a contacted individual j, we employ the Bayesian formula

$$P(z_j = 1|\mathbf{B}_j, \mathbf{F}_j) = \frac{P(\mathbf{B}_j, \mathbf{F}_j|z_j = 1) \cdot P(z_j = 1)}{P(\mathbf{B}_j, \mathbf{F}_j)}. \tag{2}$$

The term $P(\mathbf{B}_j, \mathbf{F}_j|z_j = 1)$ is called the likelihood, indicating the distributions of behaviors and features for any infected individual j. Assuming the behaviors and features are independent[52], we have

$$P(\mathbf{B}_j, \mathbf{F}_j|z_j = 1) = \prod_u P(B_j^{(u)}|z_j = 1) \cdot \prod_v P(F_j^{(v)}|z_j = 1). \tag{3}$$

Since we have found that the probabilities for various contacts follow power-law distributions, i.e.,

$$P(b_j^{(u)} = k|z_j = 1) = c^{(u)} \cdot k^{-\gamma^{(u)}},\ k = 1, 2, \cdots, \tag{4}$$

where coefficient $c^{(u)}$ is the normalizing constant, satisfying

$$c^{(u)} = \frac{1}{\int_{k=1}^{\infty} k^{-\gamma^{(u)}} dk} = \gamma - 1,\ \gamma > 1. \tag{5}$$

We next try to compute the values of c and γ by maximum likelihood estimate[53]. Supposing we have N infected samples $b_1, b_2, \cdots, b_N$, we obtain the likelihood function

$$l(\gamma) = \ln P(b_1, b_2, \cdots, b_N|\gamma) = \ln \prod_{j=1}^{N} (\gamma - 1) \cdot b_j^{-\gamma} = (-\gamma) \cdot \sum_{j=1}^{N} \ln b_j + N \cdot \ln(\gamma - 1). \quad (6)$$

Then,

$$\frac{\partial l(\gamma)}{\partial \gamma} = -\sum_{j=1}^{N} \ln b_j + N \cdot \frac{1}{\gamma - 1}. \quad (7)$$

Holding $\frac{\partial l(\gamma)}{\partial \gamma} = 0$, we can obtain

$$\hat{\gamma} = 1 + \frac{N}{\sum_{j=1}^{N} \ln b_j}. \quad (8)$$

As $P(F_j^{(v)}|z_j = 1)$ indicates the features for any infected individual j such as gender or age, we assume the distributions are multinomial, i.e.,

$$P(f_j^{(u)} = k|z_j = 1) = Q^{(u)}(k). \quad (9)$$

Specifically, supposing we have M infected samples $f_1, f_2, \cdots, f_M$, the multinomial distribution $Q(k)$ is estimated by

$$\hat{Q}(k) = \frac{\mathbf{1}_{\{f_j = k\}}}{M}. \quad (10)$$

Notice that there is difference between the behaviors of the infected contacted individuals and the uninfected contacted individuals. We thus denote the estimations from the infected samples by $\hat{\gamma}_I^{(u)}$ for behavior u and $\hat{Q}_I^{(v)}$ for feature v, while we denote the estimations from the uninfected samples by $\hat{\gamma}_U^{(u)}$ for behavior u and $\hat{Q}_U^{(v)}$ for feature v. Substituting Eq. (8) and Eq. (10) into Eq. (2), we can calculate the posterior probability

$$\begin{aligned} &P(z_j = 1|\mathbf{B}_j, \mathbf{F}_j) \\ =&\frac{\prod_u (\hat{\gamma}_I^{(u)} - 1) \cdot b_j^{-\hat{\gamma}_I^{(u)}} \cdot \prod_v \hat{Q}_I^{(v)}(f_j^{(v)}) \cdot \rho}{\prod_u (\hat{\gamma}_I^{(u)} - 1) \cdot b_j^{-\hat{\gamma}_I^{(u)}} \cdot \prod_v \hat{Q}_I^{(v)}(f_j^{(v)}) \cdot \rho + \prod_u (\hat{\gamma}_U^{(u)} - 1) \cdot b_j^{-\hat{\gamma}_U^{(u)}} \cdot \prod_v \hat{Q}_U^{(v)}(f_j^{(v)}) \cdot (1 - \rho)}, \end{aligned} \quad (11)$$

where ρ can be obtained by the proportion of the infectious among the population.

Data availability The temporal contact graph and other key statistical information used in all the analyses will be made available upon publication. The daily symptomatic cases are referred to the Ref. [40].

Competing interests The authors declare no competing interests.

References

1. World Health Organization. Coronavirus disease (COVID-19) outbreak situation. https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (2020).

2. Kissler, S. M., Tedijanto, C., Goldstein, E., Grad, Y. H. & Lipsitch, M. Projecting the transmission dynamics of sars-cov-2 through the postpandemic period. *Science* **368**, 860–868 (2020).

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

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

5. Nature Editorial. Show evidence that apps for COVID-19 contact-tracing are secure and effective. https://www.nature.com/articles/d41586-020-01264-1 (2020).

6. Ferretti, L. *et al.* Quantifying sars-cov-2 transmission suggests epidemic control with digital contact tracing. *Science* (2020).

7. 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).

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

9. Hinch, R. *et al.* Effective configurations of a digital contact tracing app: A report to nhsx. *Technical report* (2020).

10. Ting, D. S. W., Carin, L., Dzau, V. & Wong, T. Y. Digital technology and covid-19. *Nature medicine* **26**, 459–461 (2020).

11. Ballouz, T. *et al.* Digital proximity tracing app notifications lead to faster quarantine in non-household contacts: results from the zurich sars-cov-2 cohort study. *medRxiv* (2020).

12. Rodríguez, P. *et al.* A population-based controlled experiment assessing the epidemiological impact of digital contact tracing. *Nature communications* **12**, 1–6 (2021).

13. Fancourt, D., Bu, F., Mak, H. W. & Steptoe, A. Covid-19 social study. *Results release* **22** (2020).

14. Menges, D., Aschmann, H. E., Moser, A., Althaus, C. L. & von Wyl, V. The role of the swisscovid digital contact tracing app during the pandemic response: results for the canton of zurich. *medRxiv* (2021).

15. Burdinski, A., Brockmann, D. & Maier, B. F. Digital contact tracing contributes little to covid-19 outbreak containment. *medRxiv* (2021).

16. Lewis, D. Why many countries failed at covid contact-tracing-but some got it right. *Nature* 384–387 (2020).

17. Cebrian, M. The past, present and future of digital contact tracing. *Nature Electronics* **4**, 2–4 (2021).

18. Akinbi, A., Forshaw, M. & Blinkhorn, V. Contact tracing apps for the covid-19 pandemic: a systematic literature review of challenges and future directions for neo-liberal societies. *Health Information Science and Systems* **9**, 1–15 (2021).

19. Elmokashfi, A. *et al.* Nationwide rollout reveals efficacy of epidemic control through digital contact tracing. *medRxiv* (2021).

20. Chinazzi, M. *et al.* The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* (2020).

21. Kraemer, M. U. G. *et al.* The effect of human mobility and control measures on the COVID-19 epidemic in china. *Science* (2020).

22. Wu, J. T. *et al.* Estimating clinical severity of COVID-19 from the transmission dynamics in wuhan, china. *Nature Medicine* **26**, 506–510 (2020).

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

24. Tong, Z.-D. *et al.* Potential presymptomatic transmission of sars-cov-2, zhejiang province, china. *Emerging Infectious Diseases* (2020).

25. Ba, Y. *et al.* Asymptomatic carrier transmission of COVID-19. *JAMA* (2020).

26. Zhao, S. *et al.* Preliminary estimation of the basic reproduction number of novel coronavirus (2019-ncov) in china, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *International Journal of Infectious Diseases* 214–217 (2020).

27. He, X. *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine* (2020).

28. Chowell, G., Cleaton, J. M. & Viboud, C. Elucidating transmission patterns from internet reports: Ebola and middle east respiratory syndrome as case studies. *The Journal of Infectious Diseases* S421–S426 (2020).

29. Wu, J. T., Leung, K. & Leung, G. M. Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study. *The Lancet* 689–697 (2020).

30. Lipsitch, M. *et al.* Transmission dynamics and control of severe acute respiratory syndrome. *Science* 1966–1970 (2003).

31. Riley, S. *et al.* Transmission dynamics of the etiological agent of sars in hong kong: Impact of public health interventions. *Science* 1961–1966 (2003).

32. Shaman, J., Karspeck, A., Yang, W., Tamerius, J. & Lipsitch, M. Real-time influenza forecasts during the 2012–2013 season. *Nature communications* 1–10 (2013).

33. Kucharski, A. J. *et al.* Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The Lancet Infectious Diseases* 1–7 (2020).

34. Pastor-Satorras, R. & Vespignani, A. Epidemic spreading in scale-free networks. *Physical review letters* (2001).

35. Kitsak, M. *et al.* Identification of influential spreaders in complex networks. *Nature Physics* 888–893 (2010).

36. Wu, M. *et al.* A tensor-based framework for studying eigenvector multicentrality in multilayer networks. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* **116**, 15407–15413 (2019).

37. 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). URL `https://www.medrxiv.org/content/early/2020/05/18/2020.05.06.20092841. https://www.medrxiv.org/content/early/2020/05/18/2020.05.06.20092841.full.pdf`.

38. Sun, K., Chen, J. & Viboud, C. Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: a population-level observational study. *The Lancet Digital Health* (2020).

39. He, S., Shin, D.-H., Zhang, J. & Chen, J. Near-optimal allocation algorithms for location-dependent tasks in crowdsensing. *IEEE Transactions on Vehicular Technology* 3392–3405 (2017).

40. Hao, X. *et al.* Reconstruction of the full transmission dynamics of covid-19 in wuhan. *Nature* **584**, 420–424 (2020).

41. He, X. *et al.* Temporal dynamics in viral shedding and transmissibility of covid-19. *Nature medicine* **26**, 672–675 (2020).

42. 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).

43. 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).

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

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

46. 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).

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

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

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

50. Bernardo, J. M. & Smith, A. F. *Bayesian theory*, vol. 405 (John Wiley & Sons, 2009).

51. Giordano, G. *et al.* Modelling the COVID-19 epidemic and implementation of population-wide interventions in italy. *Nature Medicine* (2020).

52. Flach, P. A. & Lachiche, N. Naive bayesian classification of structured data. *Machine Learning* **57**, 233–269 (2004).

53. Bauke, H. Parameter estimation for power-law distributions by maximum likelihood methods. *The European Physical Journal B* **58**, 167–173 (2007).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation10.5.pdf