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

Mincheng Wu
College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China.
Chao Li
College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China.
Zhangchong Shen
College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China.
Shibo He (s18he@zju.edu.cn)
College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China.
Lingling Tang
Shulan (Hangzhou) Hospital Affiliated to Shulan International Medical College, Zhejiang Shuren University, Hangzhou, China.
Jie Zheng
Zhejiang Institute of Medical-care Information Technology, Hangzhou, China
Yi Fang
Westlake Institute for Data Intelligence, Hangzhou, China.
Kehan Li
College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China
Yanggang Cheng
College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China
Zhiguo Shi
College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China.
Guoping Sheng
Shulan (Hangzhou) Hospital Affiliated to Shulan International Medical College, Zhejiang Shuren University, Hangzhou, China.
Yu Liu
Westlake Institute for Data Intelligence, Hangzhou, China.
Jinxing Zhu
Westlake Institute for Data Intelligence, Hangzhou, China.
Xingjiang Ye
Westlake Institute for Data Intelligence, Hangzhou, China.
Jinlai Chen
Westlake Institute for Data Intelligence, Hangzhou, China.
Wenrong Chen

Westlake Institute for Data Intelligence, Hangzhou, China.

Lanjuan Li (ljli@zju.edu.cn)

State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, Zhejiang University, Hangzhou, China.

Youxian Sun

College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China.

Jiming Chen (cjm@zju.edu.cn)

College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China.

Research Article

Keywords: contact tracing applications, temporal contact graph, time-varying indicators

Posted Date: May 27th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-31777/v1

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

Contact tracing APPs have been recently advocated by many countries (e.g., the United Kingdom, Australia, etc.) as part of control measures on COVID-19. Controversies have been raised about their effectiveness in practice as it still remains unclear how they can be fully utilized to fuel the fight against COVID-19. In this article, we show that an abundance of information can be extracted from contact tracing for COVID-19 prevention and control, providing the first data-driven evidence that supports the wide implementation of such APPs. 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. Five time-varying indicators we introduce can accurately capture actual contact trends at individual and population levels, demonstrating that travel restriction in Wuhan played an important role in containing COVID-19. We reveal a strong correlation (Pearson coefficient 0.929) between daily confirmed cases and daily total contacts, which can be utilized as a new and efficient way to evaluate and predict the evolving epidemic situation of COVID-19. Further, we find that there is a prominent distinction of contact behaviors between the infected and uninfected contacted individuals, and design an infection risk evaluation framework to identify infected ones. This can help narrow down the search of high risk contacted individuals for quarantine. Our results indicate that user involvement has an explicit impact on individual-level contact trend estimation while minor impact on situation evaluation, offering guidelines for governments to implement contact tracing APPs.

COVID-19, caused by SARS-CoV-2, has quickly spread to most of the countries in the last four months, and was characterized as a pandemic by World Health Organization on 11 March, 2020. As of 8 May, world-wide confirmed cases of COVID-19 has reached 3.77 million, among which about 259,593 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. Currently, many nations have been cooperating together to fight for such an unprecedented disease, and it is still too early to tell its disappearance.

Recently, contact tracing APPs have been widely advocated as part of control measures on COVID-19 pandemic[2,3]. The main idea is to exploit Bluetooth on smartphones to discover nearby devices held by users and identify the contacts with the infectious individuals. Google and Apple announced their collaboration on 10 April that aims to provide an operation-system-transparent APP for contact tracing[4]. Australia and Singapore have launched two APPs for contact tracing[5,6],

which have attracted 5.1millions and 1.4 millions smartphone users, respectively. The United Kingdom just released a trial on Isle of Wight[7] and Germany is reported to deploy such type of APPs soon[2]. As the epidemic situation evolves, many more contact tracing APPs could be implemented in the future. Controversies[8,9], however, have been raised about their effectiveness on COVID-19 prevention and control since it still remains unclear how they can be fully utilized to fuel the fight against COVID-19.

In this article, we take a pioneering and in-depth investigation into this issue and show that

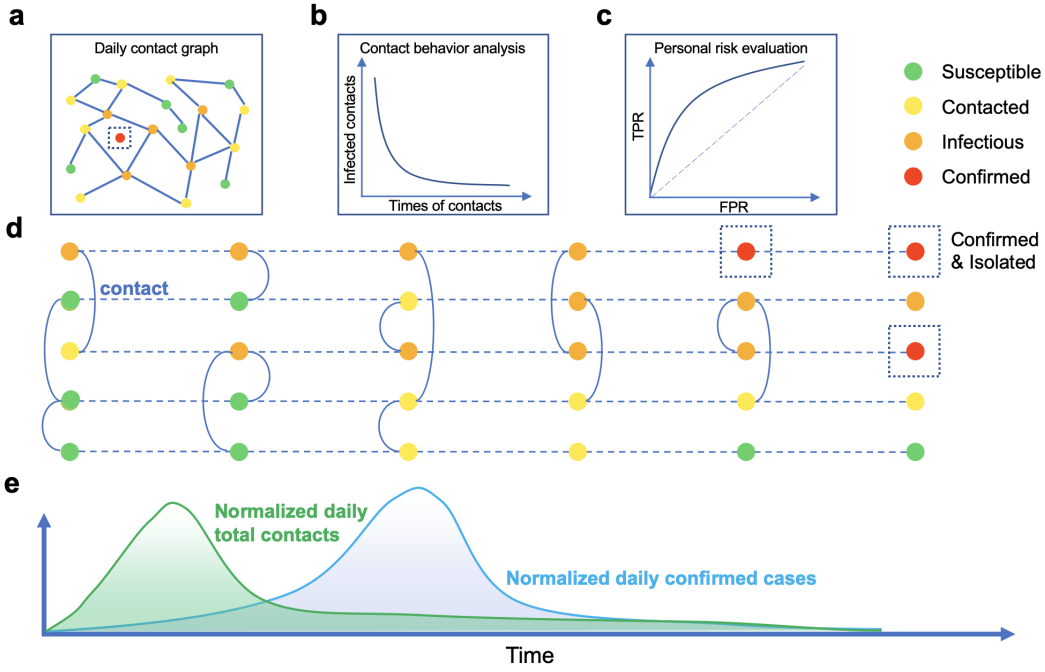

Figure 1: **Temporal contact graph and schematics for its potential applications.** An individual has four states: susceptible, contacted, infectious and confirmed. State susceptible turns to contacted when an individual had at least one contact with infectious individuals. A contacted individual may be infected or stay healthy. The state 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 healthy contacted individuals. **c.** The personal risk evaluation based on contact behaviors. **d.** Contact history and state of individuals. A circle denotes an individual and different colors indicate different states.A dashed line means the state evolution of a single individual in timeline. A solid curve between two individuals means a contact. **e.** The correlation between normalized daily total contacts and daily confirmed cases.

an abundance of information can be extracted from 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 mainly focused on integrating mathematical models and available statistical data of confirmed cases to characterize the transmission of epidemic diseases[10–26], opening up a new perspective to understand the spread of COVID-19.

Since contact tracing APPs are essentially based on crowdsourcing[27,28], their performance highly relies on the involvement of voluntary smartphone users. Due to potential privacy leak and cost incurred during crowdsourcing process, voluntary users are reluctant to participate and contribute their personal data at a fine-grained scale. 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 restriction 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 0.929) between the number of daily confirmed cases and daily total contacts 14 days ago, offering a promising way to evaluate and predict the evolving epidemic situation of COVID-19. We find that there is a prominent distinction of contact behaviors between the infected and uninfected contacted individuals, and design an infection risk evaluation framework to identify infected ones. This can help narrow down high-risk contacted individuals for quarantine and greatly reduce the cost of random testing and clinical diagnosis. We also study the effect of user involvement on the effectiveness of contact tracing APPs, providing guidelines for governments to implement contact tracing APPs. Our results provide the first data-driven evidence that supports the wide implementation of contact tracing APPs.

Results

Characteristics of Informative Indicators. Based on the contact history data provided by Westlake Institute for Data Intelligence, we build a contact model (see the Method section for more details) consisted of two types of the contacts: 1) building-level location-based contact (BLC), and 2) room-level anchor-based contact (RAC). Without loss of generality, we focus on the contact history data contributed by 10,527,737 smartphone users in Wuhan, China. Based on a set 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 is a day index): 1) $C(t)$, the daily total times of BLCs between infectious and contacted individuals; 2) $K(t)$, the daily average times of BLCs for contacted individuals associating with

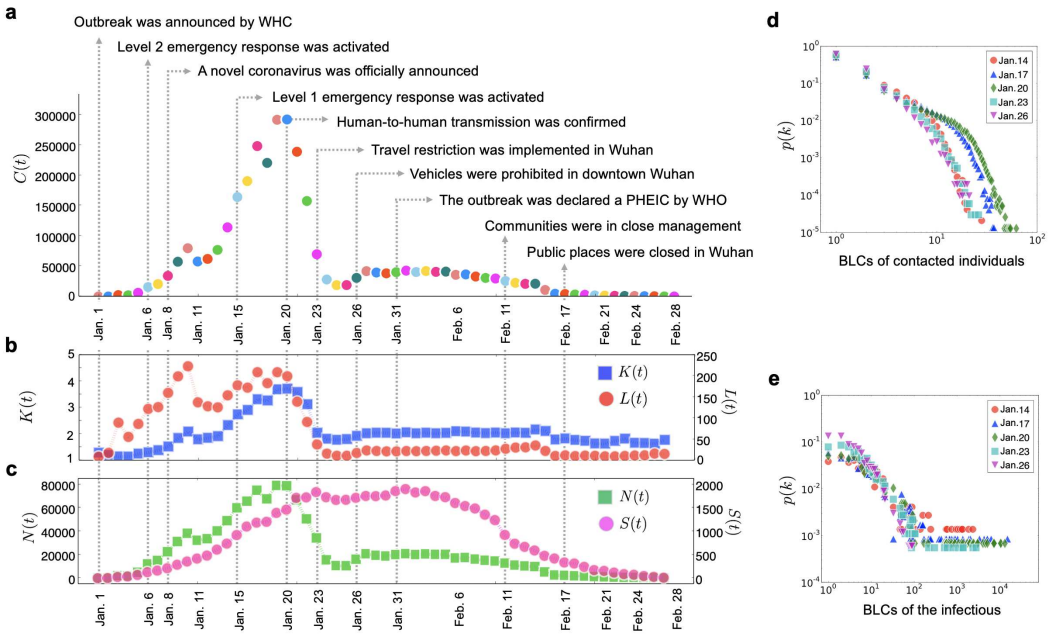

Figure 2: **Daily characteristics of building-level location-based contacts (BLCs). a.** $C(t)$, the daily total times of BLCs between infectious and contacted individuals. **b.** $K(t)$, the daily average times of BLCs for contacted individuals associating with infectious individuals. $L(t)$, the daily average times of BLCs for infectious individuals associating with contacted individuals. **c.** $N(t)$, 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. **d.** The distributions of the daily times of BLCs by all contacted individuals. **e.** The distributions of the daily times of BLCs by all confirmed cases.

infectious individuals; 3) $L(t)$, the daily average times of BLCs for infectious individuals associating with contacted individuals; 4) $N(t)$, the daily total number of contacted individuals who had encountered the infectious at least once under BLC model, and 5) $S(t)$, the daily total number of infectious individuals who had encountered with contacted individuals at least once under BLC model.

The daily total contacts between infectious and susceptible individuals $C(t)$ can reflect the daily overall transmission (Fig. 2a). We find that $C(t)$ increased dramatically first from 1 to 19 January 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 travel restriction was implemented in Wuhan, and tended to zero around 28 February. Similar evidence can be observed for indicators $K(t)$ and $L(t)$, while $L(t)$ displays a more distinct fluctuation in the early January since the infected are not isolated, and they contacted susceptible as usual in the incubation period (Fig. 2b). Because the small proportion of the infected and the randomness of their movements, the two indicators were not not stable during 10 to 20 January. For example, they first dropped a bit from 10 to 12 January, which may be due to the mobility reduction caused by the sudden drop of temperature (Supplementary Fig. 5). The dynamic $K(t)$ and $L(t)$ accurately describe actual individual-level contact trend in Wuhan, providing a data-driven evidence that travel restriction 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 a macroscopic view, $N(t)$ and $S(t)$ describe population-level contact trend in Wuhan (Fig. 2c). Notice that $N(t)$ had a similar behavior as $K(t)$ except a minor bouncing back after 23 January, when the travel restriction was implemented in Wuhan. This is possibly due to the number of confirmed cases quickly increased after 20 January and people in Wuhan could still move within the city (their mobility increased due to the approaching of Chinese New Year). Contacted individuals 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 four indicators. Initially, $S(t)$ quickly 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 small duration of increase. It decreased again on 4 February and eventually approached to 0 around 28 February. The main reason is that the confirmed cases increased fast after 20 January and the chance of meeting a infectious individual remained high as

many of them were not hospitalized due to test capacity constraint.

Fig. 2d shows the detailed distributions of the daily contacts, which all have power-law tails. From 1 to 20 January, specifically, the power exponent increased first and then decreased, having the same evolving pattern as $K(t)$. From the perspective of the infectious, the distribution of daily contacts also follows a power-law distribution (Fig. 2e), and have a prominent long tail especially when the exponent coefficient is small before 23 January. The long tails indicate that there are some super active cases that had contact with hundreds of susceptible individuals. Identifying and quarantining them help mitigate the fast transmission. Therefore, $K(t)$, $L(t)$, $C(t)$ ($N(t)$ and $S(t)$) characterize the spread of COVID-19 from dimensions of susceptible individuals, confirmed cases and overall contacts and is informative for COVID-19 prevention and control (see Supplementary Figs. 6, 7 for more analysis of RACs).

A strong correlation between daily number of contacts and confirmed cases. 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. Noticing that we have two types of contacts: RAC and BLC, we investigate the correlation between the daily total number of contacts $C(t)$ and confirmed cases under these two types of contacts, respectively.

The curves of daily total contacts under RAC and BLC, as well as the daily confirmed cases with log normalization (i.e., first we apply logarithm for the time series and then normalize it) 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 confirmed cases ahead, these curves present more similar trends. To find the proper delay that results in the best similarity in trends between the curves of daily total contacts and confirmed cases, we alter the delays range from 10 to 20 days (see Supplementary Fig. 9 for more details). The experiments show that a 14-day delay results in the best Pearson correlation[29] under both RAC and BLC models. This also implies that it takes an average duration of 14 days for a contacted individual to be confirmed after being infected.

More specifically, the Pearson's correlation achieves 0.9292 and 0.8016 (Fig. 3b and 3c) under the RAC and BLC model, respectively. Also, the results show that RAC model has a better performance in estimating the confirmed cases than BLC model does. This illustrates that the RAC model is more accurate to characterize an effective contact between an infectious and a susceptible individual, because: 1) BLC may categorize many passersby in building level as contacts who had

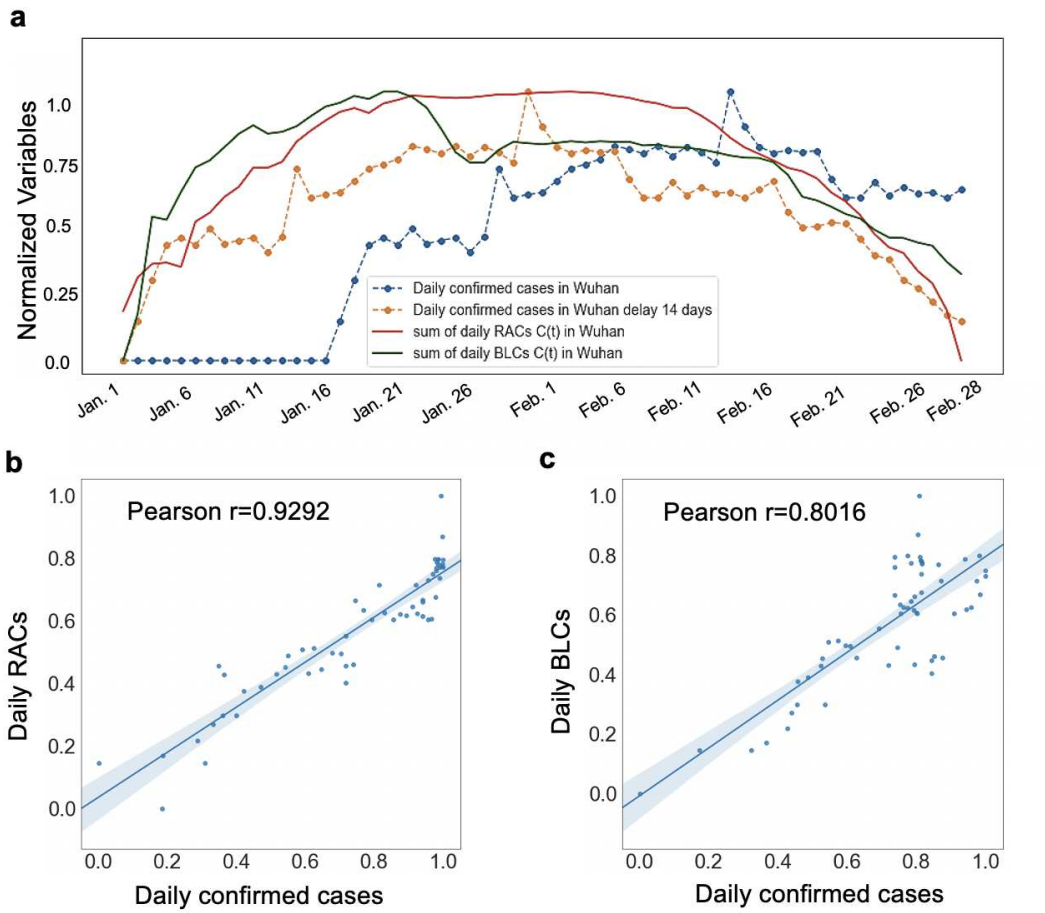

Figure 3: **a.** Historical time series of daily total number of contacts under RAC and BLC models, daily number of confirmed cases in Wuhan and daily number of confirmed cases with 14-day delay. **b.** Correlation analysis between daily number of RAC contacts and daily number of confirmed cases. **c.** Correlation analysis between daily number of BLC contacts and daily number of confirmed cases.

not actually had close contact with infectious individuals, 2) the chance of getting infected is higher between close relationships in indoor environments described by RAC. In summary, indicator $C(t)$ determines the confirmed case counts in the near future, which provides a new way to evaluate and predict the epidemic situation of COVID-19.

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 twenty days, and calculate the probability $p(k)$ that a contacted individual have k times of RACs. The probability distribution decays as a power-law (Fig. 4a), following $p(k) \propto k^{-\gamma}$. The average number of RACs equals to 8.34 and the estimated $\gamma = 1.66$. We further study the contact behaviors for infected and uninfected contacted individuals, respectively. Again, power-law behaviors are found for both types, while the parameters are prominently different. The infected contacted individuals have a power-law distribution with an average $< k >= 13.09$ and an exponent $\gamma = 1.21$, while the uninfected contacted individuals have a power-law distribution with $< k >= 8.17$ and $\gamma = 1.68$ (Fig. 4b). Clearly, these two distributions are of significant differences in terms of the expectations and the power exponents: the infected contacted individuals have more RACs than uninfected contacted individuals and the distribution has a fatter tail. This indicates that there are an appreciable quantity of infected contacted individuals with a large amount of RACs (see Supplementary Fig. 8 for analysis of more contact behaviors).

For the contact behavior under BLC model, similar characteristics are found. Specifically, the probability $p(k)$ that a contacted individual has k times of BLCs with the infectious in recent twenty days also decays as power-law, with $< k >= 5.39$ and $\gamma = 1.96$ (Fig. 4c). Among all the contacted individuals, the infected ones have a power-law distribution with an average $< k >= 5.93$ and an exponent $\gamma = 1.86$, while the uninfected ones with $< k >= 5.38$ and $\gamma = 2.01$ (Fig. 4d). The contact behaviors between the infected and the uninfected contacted individuals are also differentiable. By comparing the results under both RACs and BLCs, we conclude that there is a clear contact behavior discrimination between infected and uninfected contacted individuals and the RAC model yields more prominent discrimination than the BLC model.

Based on the contact behavior discrimination, we proceed to perform individual-level infection risk evaluation for each contacted individual. We propose a risk evaluation method based on the Bayesian framework that calculates the posterior probability for any contacted individual being infected[30]. We first introduce a variable z_j to represent the health state 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 $P(z_j = 1|B_j, F_j)$:

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

where B_j denotes contact counts under RAC and BLC model for j, and F_j denotes the individual feature, indicating the age, residence, and etc. (Fig. 4e). The term $P(B_j, F_j|z_j = 1)$ is the likelihood, and $P(z_j = 1)$ indicates the probability for any contacted individual j being infected a

prior, which is taken as a constant (see the Methods section for more details).

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. 4f). 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 60% of the infected contacted individuals with 20% false report of the uninfected contacted individuals by using the contact and feature information. This can help narrow down high risk contacted individuals for quarantine in practice. Obviously, infor-

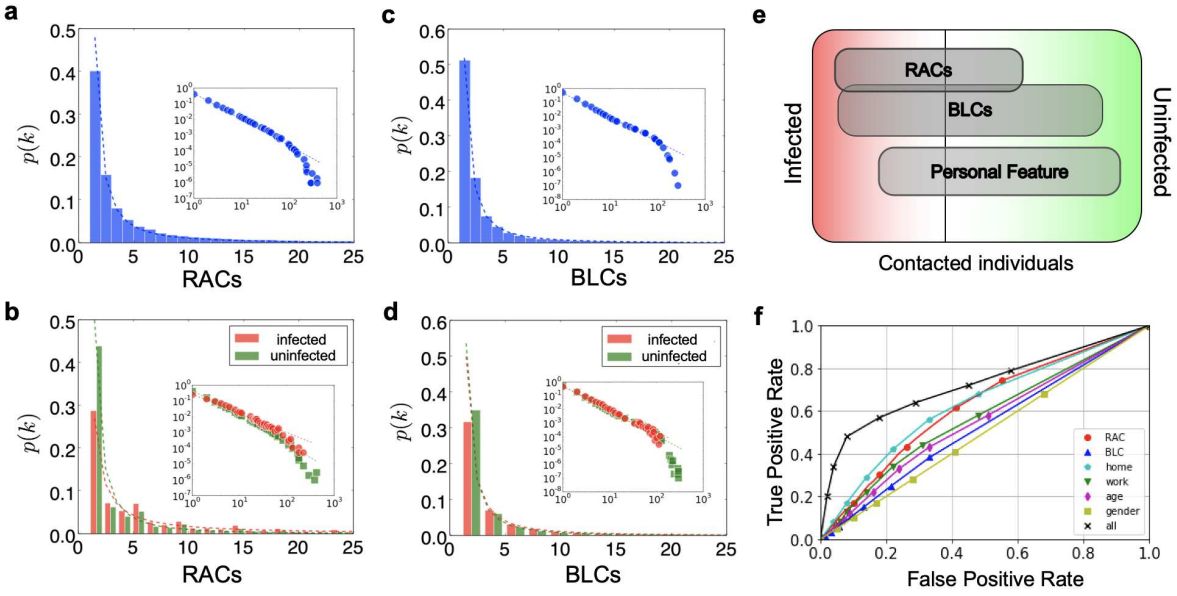

Figure 4: **Infection risk evaluation based on the Bayesian framework. a.** The distribution of the times of RACs with the infectious by all contacted individuals. **b.** The distributions of the times of RACs with the infectious by infected and uninfected contacted individuals, respectively. **c.** The distribution of the times of BLCs with the infectious by all contacted individuals. **d.** The distributions of the times of BLCs with the infectious by infected and uninfected contacted individuals, respectively. **e.** A Venn diagram describing the relationship between events and the posterior probability. **f.** 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.

mation of RACs, residence and work address provides a more accurate discrimination to identify the infected contacted individuals, while there is nearly no distinction by gender. The result confirms to our behavior analysis between the two types of contacts and features, demonstrating that a more prominent discrimination yields a better performance to distinguish infected from uninfected contacted individuals.

The impacts of user involvement on the contact tracing APP performance. Clearly, contact tracing APPs are based on crowdsouring. Individual smartphone users are voluntary to participate in the process and upload their contact information. Many governments are about to launch such APPs, however, it remains open to tell how their performance (e.g., estimating $K(t)$ and $L(t)$ and daily confirmed cases) is affected by user involvement, raising questions on whether such APPs 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 using a tracing APP among 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.

We conduct extensive experiments by varying the values of α under the BLC model. At a specific α, we plot the statistical information such as median and variance of the time series of $K(t)$, $L(t)$ and total contacts $C(t)$ for both scenarios of user participation rate and user upload rate with box plots. It is shown that, as α decreases, the statistical information decreases with the similar trend (Fig. 5a-5f). 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 correlation between the time series under $\alpha\%$ and full (100%) participation rate/data upload rate case (Fig. 5g and 5h). 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 $C(t)$. Note that $C(t)$ is a population-level statistical indicator while $K(t)$ and $L(t)$ are individual-level statistical indicators. The above observations indicate that population-level indicators are more robust than individual-level indicators 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. 5g and Fig. 5h. 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

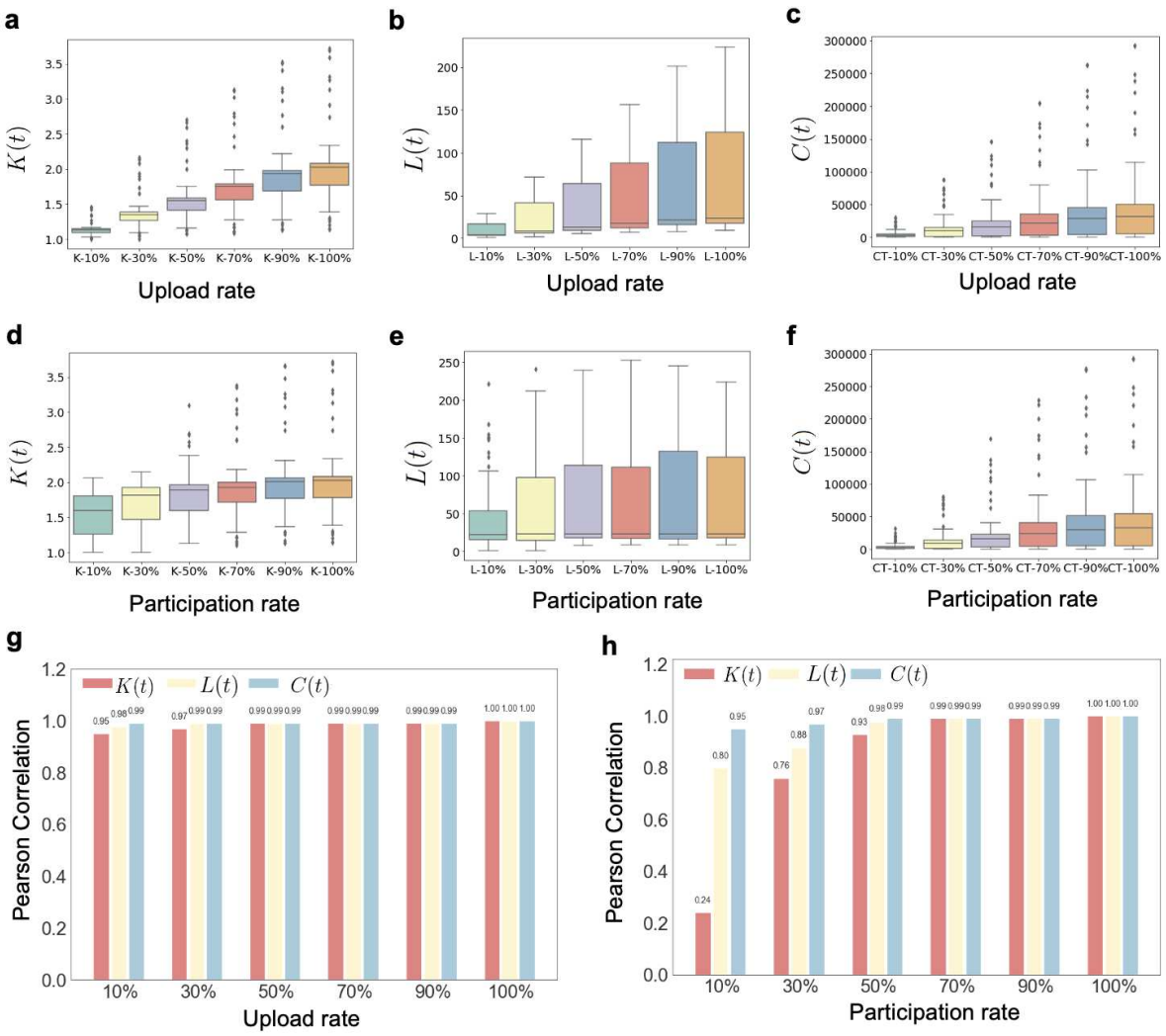

Figure 5: **Performance of different user involvement under BLC model. a-c.** Three box plots show the distribution change of $K(t)$, $L(t)$, and daily number of total contacts Vs. different user update rates under BLC model. **d-f.** Three box plots show the distribution change of $K(t)$, $L(t)$, and daily number of total contacts Vs. different user participation rates under BLC model. **g-h.** The Pearson correlation Vs. different user upload rates and user participation rates under BLC model, respectively.

strategy to allow voluntary smartphone users having a relatively low data upload rate. We have the similar analysis for RAC model (See Supplementary Fig. 10 for more details).

Discussions

Since the emergence of COVID-19, researchers has proposed many mathematical models to characterize the transmission of COVID-19[10–13,31]. These previous studies provided analytical models to understand the transmission of COVID-19. As the contact tracing APPs are advocated by many countries, it rises the pressing issue 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 obtained from tracing data. 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 approach are complementary to characterize the transmission of COVID-19.

As only few contact tracing APPs have been implemented recently, the contact tracing data is still unavailable. Their performance on COVID-19 prevention and control can not be directly evaluated. Here, we leverage a large amount of location related data contributed by 10,527,737 voluntary users to study such an issue. 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 individual-level contact estimation such as $K(t)$ and $L(t)$. Our results can provide guidelines for governments to practically deploy contact tracing APPs.

Methods

Method I: The Contact Model The crowdsourced data are contributed by 10,527,737 voluntary users in Wuhan, China and collected by crowdsourcing platforms of our industry partners. The smartphone users are voluntary to upload their located related information when they are using LBS. To protect the privacy, privacy protection mechanisms such as perturbation and pseudonymisation are adopted during data collection. There are two types of location related information: 1) building-level location-based information including POI, GPS, geomagnetic, etc., which is projected into a meshed area of about 450 m^2 , and 2) room-level anchor-based information including WiFi access point, Bluetooth, UWB anchor, etc., which indicates an indoor area of about 100 m^2. All the crowdsourced data are provided with timestamps (see Supplementary Figs. 1-4 for more details).

The confirmed cases from 18 January to 28 February, 2020, serve as the sources of the infection. Recent results indicated that transmission can be happened both before and after the symptom onset, known as pre-symptomatic transmission and symptomatic transmission. Taking into account of both types of transmission, we set the potential infectious duration of a confirmed case to be twenty days. This means that a confirmed case is regarded as infectious individual in the last twenty days upon confirmation, after which he/she is no longer considered as sources of infection since they are under quarantine for treatment.

The contact model is built on the available crowdsourced data. As smartphone users report data in a very low and irregular frequency, the contributed data are typically sparse time series. We would miss many contacts if we only count those explicit ones that two smartphone users are reporting identical information within a small time interval (e.g., less than five minutes). For the location-based information, considering the data sparsity, we relax the time criteria to two hours. Different from location-based information, the anchor-based information indicates a more intimate relationship (e.g., workmates, close friends or family members) and a longer stay period. Apparently, users reporting the same anchor-based information have a high chance to have much more close contacts though their report may not be synchronized. Therefore, we define a contact when the identical anchor-based information is reported by two users within the last twenty days.

Summarizing, the contact model consists of two types of contacts: 1) building-level location-based contact (BLC), i.e., a contact occurs when two users report identical meshed area within 2 hours interval, 2) room-level anchor-based contact (RAC), i.e., a contact exists when identical

anchor information is reported by two users within the last twenty days regardless of their report timestamps. By using the above contact model, we identify 562,280 susceptible individuals having contacts with 16,647 infectious individuals who turn to confirmed cases at last. The temporal contact graph is constructed based on the contacts between infectious and susceptible individuals, and is used in all the analysis in this article.

Method II: 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 $P(B_j, F_j|z_j = 1)$ is called the likelihood, indicating the distributions of behaviors and features for any infected individual j. Assuming the behaviors and features are independent[32], 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

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

We next try to compute the values of c and γ by maximum likelihood estimate[33]. Supposing we have N infected samples $b_1, b_2, \cdots, b_N$, we obtain the likelihood function

$$l(\gamma) = \ln P(\mathbf{B}|\gamma) = \ln \prod_{j=1}^{N} (\gamma - 1) \cdot b_j^{-\gamma} = (-\gamma) \cdot \sum_{j=1}^{N} \ln b_j + N \cdot \ln(\gamma - 1). \tag{6}$$

Then,

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

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

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

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.,

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

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

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

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

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

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

Data availability The temporal contact graphs and other key statistical information used in all the analyses will be made available upon publication.

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. 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).

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

4. Apple & Google. Privacy-preserving contact tracing. https://www.apple.com/covid19/contacttracing (2020).

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

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

7. 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).

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

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

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

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

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

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

16. 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).

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

18. 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).

19. 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).

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

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

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

23. 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).

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

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

26. 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).

27. 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).

28. 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).

29. Benesty, J., Chen, J., Huang, Y. & Cohen, I. Pearson correlation coefficient. In *Noise reduction in speech processing*, 1–4 (Springer, 2009).

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

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

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

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

Figures

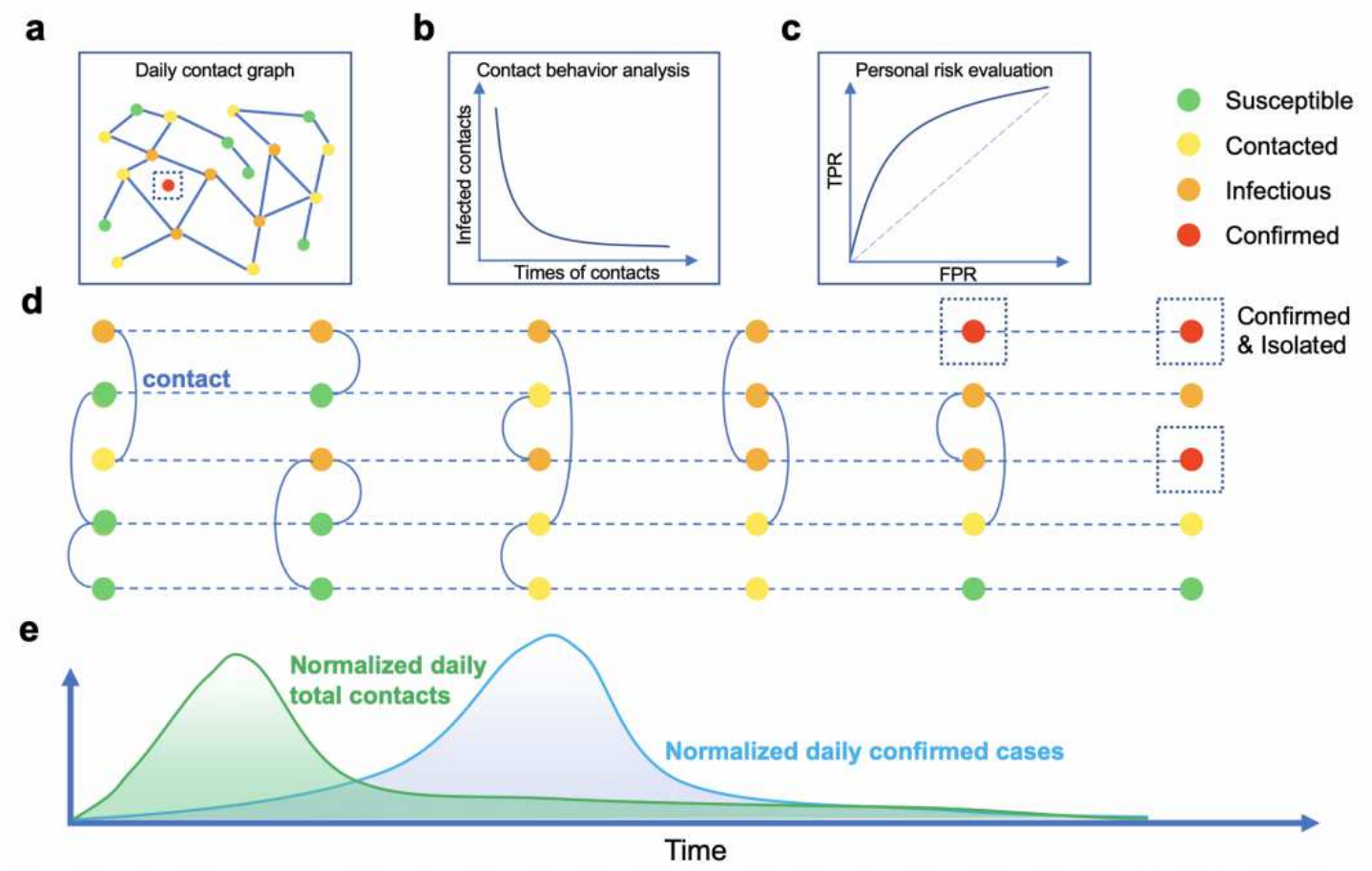

Figure 1

Temporal contact graph and schematics for its potential applications. An individual has four states: susceptible, contacted, infectious and confirmed. State 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 state 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 healthy contacted individuals. c. The personal risk evaluation based on contact behaviors. d. Contact history and state of individuals. A circle denotes an individual and different colors indicate different states.A dashed line means the state evolution of a single individual in timeline. A solid curve between two individuals means a contact. e. The correlation between normalized daily total contacts and daily confirmed cases.

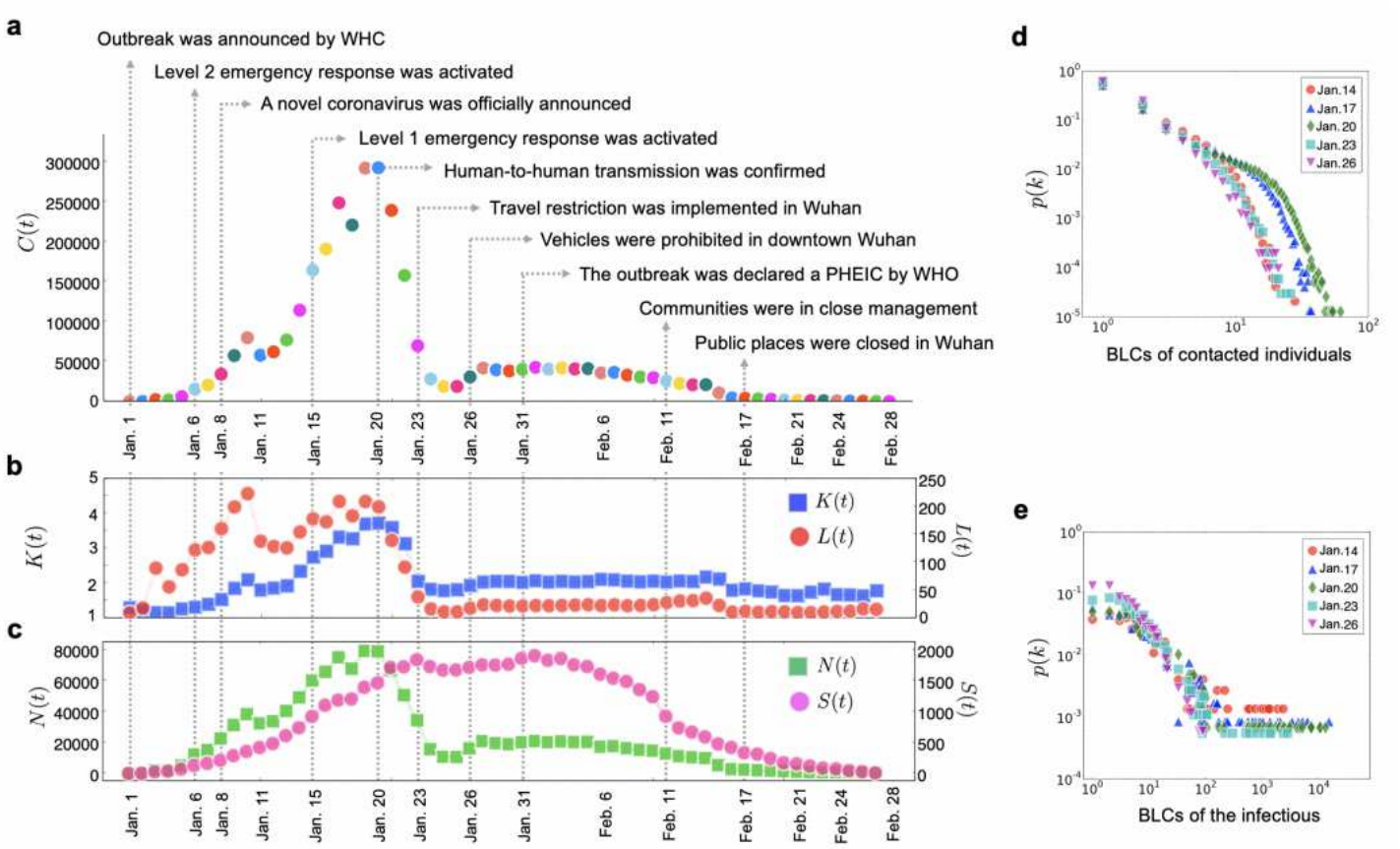

Figure 2

Daily characteristics of building-level location-based contacts (BLCs). a. C(t), the daily total times of BLCs between infectious and contacted individuals. b. K(t), the daily average times of BLCs for contacted individuals associating with infectious individuals. L(t), the daily average times of BLCs for infectious individuals associating with contacted individuals. c. 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. d. The distributions of the daily times of BLCs by all contacted individuals. e. The distributions of the daily times of BLCs by all confirmed cases.

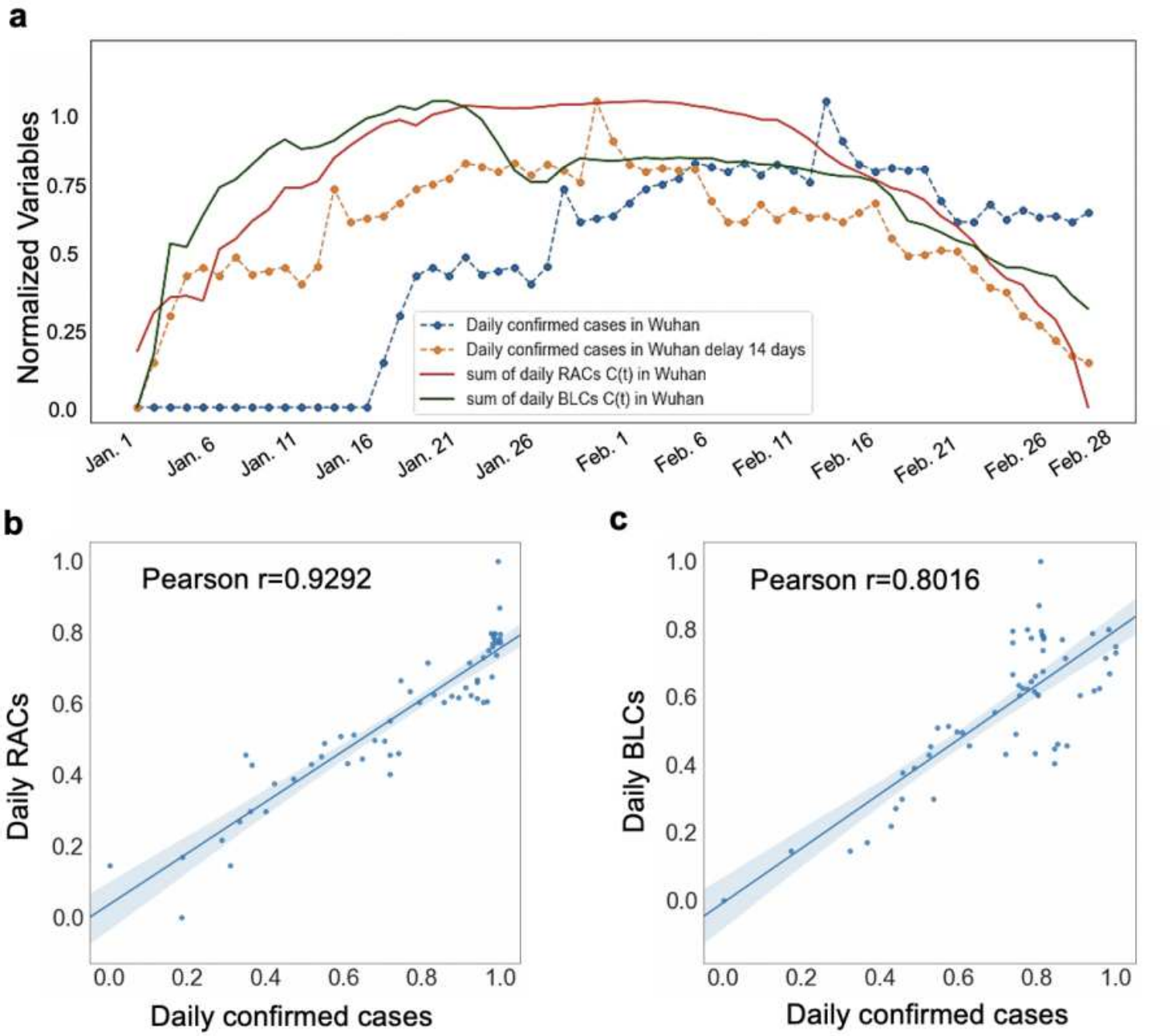

Figure 3

a. Historical time series of daily total number of contacts under RAC and BLC models, daily number of confirmed cases in Wuhan and daily number of confirmed cases with 14-day delay. b. Correlation analysis between daily number of RAC contacts and daily number of confirmed cases. c. Correlation analysis between daily number of BLC contacts and daily number of confirmed cases.

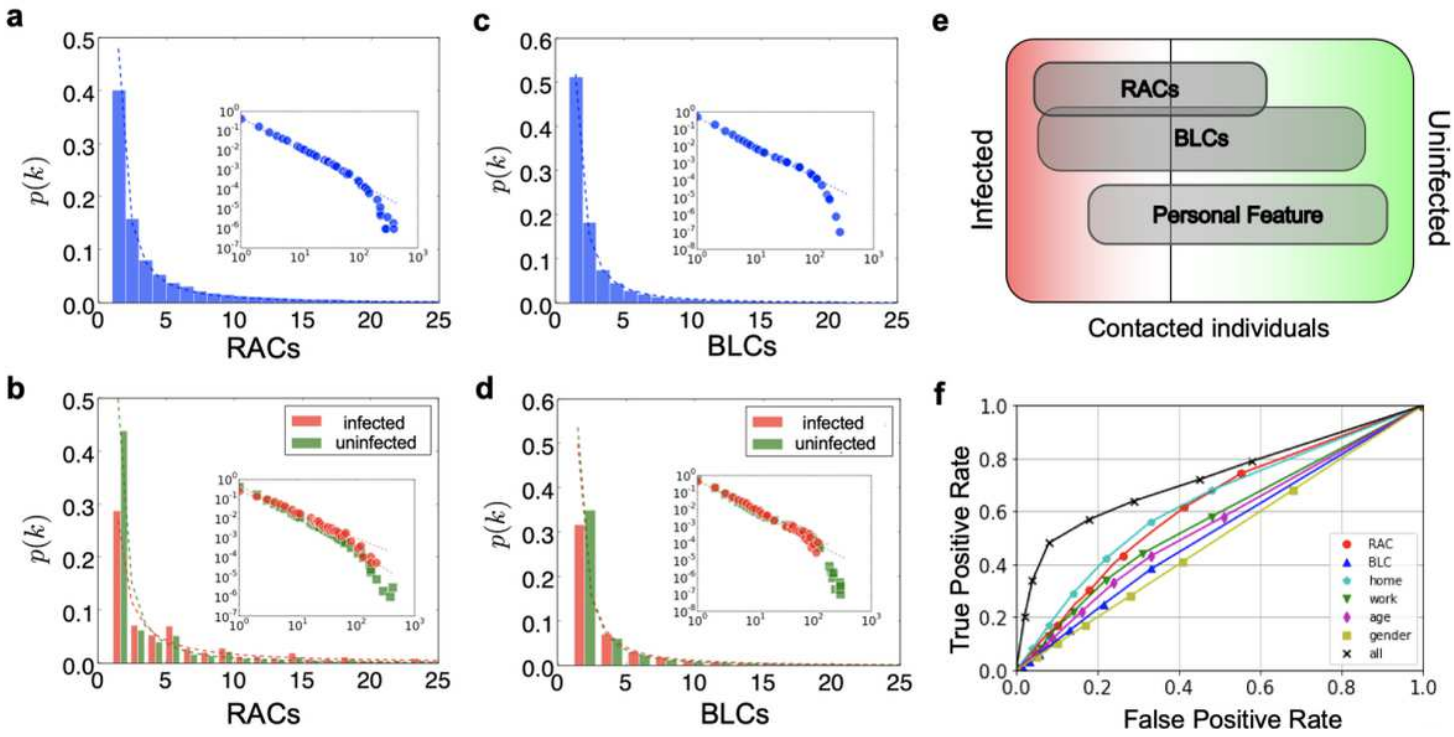

Figure 4

Infection risk evaluation based on the Bayesian framework. a. The distribution of the times of RACs with the infectious by all contacted individuals. b. The distributions of the times of RACs with the infectious by infected and uninfected contacted individuals, respectively. c. The distribution of the times of BLCs with the infectious by all contacted individuals. d. The distributions of the times of BLCs with the infectious by infected and uninfected contacted individuals, respectively. e. A Venn diagram describing the relationship between events and the posterior probability. f. 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.

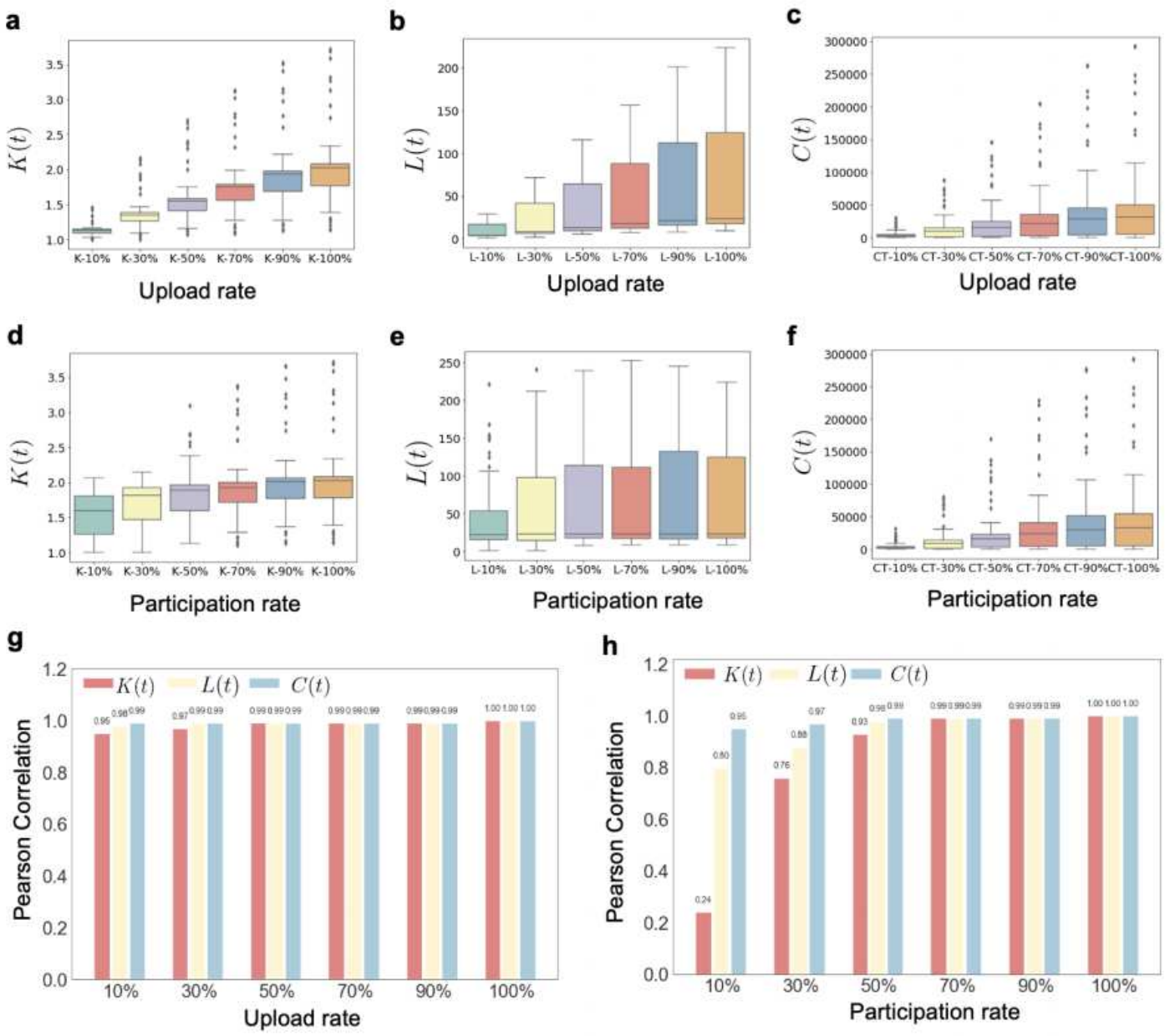

Figure 5

Performance of different user involvement under BLC model. a-c. Three box plots show the distribution change of K(t), L(t), and daily number of total contacts Vs. different user update rates under BLC model. d-f. Three box plots show the distribution change of K(t), L(t), and daily number of total contacts Vs. different user participation rates under BLC model. g-h. The Pearson correlation Vs. different user upload rates and user participation rates under BLC model, respectively.

Supplementary Files

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

- 20200508432SI.pdf