International Journal of Infectious Diseases 95 (2020) 391–398

Contents lists available at ScienceDirect

International Journal of Infectious Diseases

journal homepage: www.elsevier.com/locate/ijid

Statistical and network analysis of 1212 COVID-19 patients in Henan, China

Pei Wang[a,b,c,*], Jun-an Lu[d,*], Yanyu Jin[a], Mengfan Zhu[e], Lingling Wang[a], Shunjie Chen[a]

[a] School of Mathematics and Statistics, Henan University, Kaifeng, 475004, China
[b] Institute of Applied Mathematics, Henan University, Kaifeng, 475004, China
[c] Laboratory of Data Analysis Technology, Henan University, 475004, Kaifeng, China
[d] School of Mathematics and Statistics, Wuhan University, Wuhan, 430070, China
[e] School of Mathematics and Statistics, Zhongnan University of Economics and Law, Wuhan, 430073, China

ARTICLE INFO

Article history:
Received 18 March 2020
Received in revised form 15 April 2020
Accepted 18 April 2020

Keywords:
COVID-19
Henan province
Incubation period
Network analysis
Aggregate outbreak
Time-phased nature of epidemic

ABSTRACT

Background: COVID-19 is spreading quickly all over the world. Publicly released data for 1212 COVID-19 patients in Henan of China were analyzed in this paper.
Methods: Various statistical and network analysis methods were employed.
Results: We found that COVID-19 patients show gender (55% vs 45%) and age (81% aged between 21 and 60) preferences; possible causes were explored. The estimated average, mode and median incubation periods are 7.4, 4 and 7 days. Incubation periods of 92% of patients were no more than 14 days. The epidemic in Henan has undergone three stages and has shown high correlations with the numbers of patients recently returned from Wuhan. Network analysis revealed that 208 cases were clustering infected, and various People's Hospitals are the main force in treating COVID-19.
Conclusions: The incubation period was statistically estimated, and the proposed state transition diagram can explore the epidemic stages of emerging infectious disease. We suggest that although the quarantine measures are gradually working, strong measures still might be needed for a period of time, since ~7.45% of patients may have very long incubation periods. Migrant workers or college students are at high risk. State transition diagrams can help us to recognize the time-phased nature of the epidemic. Our investigations have implications for the prevention and control of COVID-19 in other regions of the world.
© 2020 The Author(s). Published by Elsevier Ltd on behalf of International Society for Infectious Diseases. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Background

First erupting in Wuhan (in Hubei province of China) at the end of the year 2019, novel coronavirus pneumonia (named COVID-19 by WHO on February 11, 2020) has spread all over the world. It has directly resulted in more than 74600 confirmed patients and 2121 people dead in China up to February 20, 2020 (Zhu et al., 2020; Huang et al., 2020; Zhou et al., 2020; Yang et al., 2020; Zhao et al., 2020a; Xu et al., 2020a; Yan et al., 2020; Zhang, 2020; Xu et al., 2020b; Du et al., 2020; Tang et al., 2020; Liu et al., 2019; Lu and Wang, 2020; Zhao et al., 2020b; Liu et al., 2020; Wang et al., 2020; Chen et al., 2020; Zhang et al., 2020). Various investigations reveal that COVID-19 is different from SARS in 2003 in the following aspects (Zhu et al., 2020; Huang et al., 2020; Zhou et al., 2020; Yang et al., 2020; Zhao et al., 2020a; Zhang, 2020): 1) Higher infection rate; 2) Has an incubation period; 3) Patients are infectious during the incubation period; 4) Low lethality rate; 5) High mortality among elderly patients with underlying diseases; 6) Symptomatic infection. These features make the prevention and control of COVID-19 a difficult task.

As a double-edged sword, modern traffic technologies greatly shorten the distances among people, but they also facilitate the rapid and wide spread of epidemic diseases (Lu and Wang, 2020). Since January 23, Wuhan has undertaken strict measures to close the city, and many other places have adopted various measures in succession to prevent COVID-19, such as blocking traffic and using quarantine measures. After the outbreak of COVID-19, researchers have focused on investigating it, including its control and prediction (Zhu et al., 2020; Huang et al., 2020; Zhou et al., 2020; Yang et al., 2020; Zhao et al., 2020a; Xu et al., 2020a; Yan et al., 2020; Zhang, 2020; Xu et al., 2020b; Du et al., 2020; Tang et al., 2020; Liu et al., 2019; Lu and Wang, 2020; Zhao et al., 2020b;

[*] Corresponding authors. Tel.: +86 15137867837 (Pei Wang); +86 18627883425 (Jun-an Lu); fax: +86 0378 23881696.
E-mail addresses: wp0307@126.com, wangpei@henu.edu.cn (P. Wang), jalu@whu.edu.cn (J.-a. Lu).

https://doi.org/10.1016/j.ijid.2020.04.051
1201-9712/© 2020 The Author(s). Published by Elsevier Ltd on behalf of International Society for Infectious Diseases. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Liu et al., 2020; Wang et al., 2020; Chen et al., 2020; Zhang et al., 2020). Initial investigations pointed out that *bats* are most likely the source of the virus (Zhou et al., 2020; Xu et al., 2020a), and *Manis pentadactyla* may be a potential intermediate host (Liu et al., 2019). The related investigations have great implications on prevention, control, and vaccine development (Pastor-Satorras and Vespignani, 2001; Kitsak et al., 2010; Wang et al., 2014a; Wang et al., 2016; Wang et al., 2014b; Wang et al., 2019; Lu et al., 2016; Zhang et al., 2016; Wei et al., 2018; Xu et al., 2019; Pastor-Satorras et al., 2015). Some recent publications have reported the epidemiological and clinical features of COVID-19 patients (Zhu et al., 2020; Huang et al., 2020; Zhou et al., 2020; Yang et al., 2020; Zhao et al., 2020a; Zhang, 2020). To prevent the spread of COVID-19, the Centers for Disease Control and Prevention (CDCs) of many cities have quickly released the data of confirmed cases, which enables us to explore its epidemiological characteristics and to understand the current situation and characteristics of the epidemic.

As a neighbor of the Hubei province, Henan province has a large population size, and it is also one of the hardest hit areas of the epidemic. The total number of confirmed cases in Henan is only lower than Hubei and Guangdong (Feb. 14). Statistical analysis on the patients' data can help us to understand its epidemiological and clinical features. In this paper, based on data collected from all levels of CDCs in Henan, we address the following points: 1) the current situation and characteristics of the epidemic in the 18 regions of Henan; 2) the sex and age distributions of confirmed cases; 3) estimations of the incubation period of patients and exploration of the time-phased nature of the epidemic; 4) the correlation between the number of total confirmed cases with those returning from Wuhan; 5) network analysis of aggregate outbreak phenomena.

2. Methods

2.1. Data

Various levels of CDCs in Henan province have publicly released patients' data, which can be freely obtained from their official websites.

2.2. Modeling the incubation period

We suppose the incubation period τ follows the logarithm normal distribution (Armenian and Lilienfeld, 1983):

$$\ln\tau \sim N(\mu, \sigma^2). \tag{1}$$

Thus, the probability density function (PDF) of τ can be written as

$$p(t) = \frac{1}{\sigma t\sqrt{2\pi}} e^{-(\ln t-\mu)^2/2\sigma^2}, t > 0. \tag{2}$$

Here, μ, σ^2 are two parameters to be estimated from data. We consider three approaches to estimate the parameters: the moment estimation (ME), the ordinary least square (OLS) estimation and the maximum likelihood estimation (MLE) (Buhlmann and Geer, 2011). As to the ME, based on the PDF of the incubation period, we can easily obtain

$$\begin{cases} E(\tau) = e^{\mu+\sigma^2/2}, \\ D(\tau) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}. \end{cases} \tag{3}$$

Setting $E(\tau)$, $D(\tau)$ with the estimated values from 483 patients (shown in Table 1), we obtain

$$\{ e^{\hat{\mu}_m} + \hat{\sigma}_m^2/2 = 7.4286, (e^{\hat{\sigma}}_m 2 - 1)e^{2\hat{\mu}_m} + \hat{\sigma}_m^2 = 24.1458. \tag{4}$$

Using the fsolve function in Matlab, we obtain $\hat{\mu}_m = 1.8239, \hat{\sigma}_m^2 = 0.3629$.

Similarly, for the OLS estimation, the two parameters can be estimated from the following optimization problem:

$$(\hat{\mu}_o, \hat{\sigma}_o^2) = arg \ \min \sum\nolimits_{i=1}^{m} (p_i - f(t_i))^2. \tag{5}$$

Here, $p_i = n_i/n$ denotes the ratio of patients with incubation period $i (i = 1, 2, \ldots, m)$, which can be computed from the $n = 483$ patients. $f(t_i)$ corresponds to the PDF with incubation period t_i, which is a function of the unknown parameters μ, σ^2. Since $\tau > 0$ in Eq. (2), we omit the case $\tau = 0$ (29 patients). By solving Eq. (5), we obtain $\hat{\mu}_o = 2.0723, \hat{\sigma}_o^2 = 0.5929$.

For the MLE, suppose the incubation periods for the n patients are independent, the estimation of μ, σ^2 can be obtained from minimizing the following negative logarithmic likelihood function.

$$(\hat{\mu}_l, \hat{\sigma}_l^2) = arg \ \min\left\{ - \sum\nolimits_{j=1}^{n} \inf(t_j)\right\}. \tag{6}$$

Here, $t_j (j = 1, 2, \ldots, n)$ denotes the incubation period for the $j'th$ patient. Similarly to OLS, the case with $\tau = 0$ will not be considered. Finally, we obtain

$$\hat{\mu}_l = \frac{\sum_{j=1}^{n} \ int_j}{n} = 1.8560, \hat{\sigma}_l^2 = \frac{\sum_{j=1}^{n} \ (int_j - \hat{\mu}_l)^2}{n} = 0.7174.$$

3. Results

3.1. The epidemic situations and characteristics in Henan

From January 21 to February 14, 2020, data from a total of 1212 confirmed cases have been released in Henan. The epidemic situation, time evolution of daily increase, and cumulatively confirmed patients are shown in Fig. 1.

The daily increment of patients reached a peak value (109) on February 3, 2020. After that, the daily increment of patients significantly decreased. Among the 18 regions of Henan province, Xinyang, Nanyang, Zhumadian, Zhengzhou, Shangqiu and Zhoukou encompass more confirmed cases than the other regions. Until February 14, Xinyang had the highest daily increment as compared with the other regions. In fact, Xinyang, Nanyang and Zhumadian are all near Hubei. It was reported that about five million people have left Wuhan since January 23. According to the investigation from Xu et al. (Xu et al., 2020b; Du et al., 2020), before Wuhan was closed on January 23, Xinyang, Zhengzhou, Nanyang, Zhumadian, Zhoukou, and Shangqiu were ranked among the top 50 cities in China that received a huge amount of Wuhan personnel. The

Table 1
Statistical results of estimated as well as fitted incubation periods (Fig. 3A) from 483 confirmed patients in Henan.

Source	Mean	Mean 95% CI	Median	Mode	Variance	Interquartile range
Data	7.4286	(2.0000,20.0000)	7.0000	4.0000	24.1458	[4.0000,11.0000]
ME	7.4287	(1.9025,20.1787)	6.1960	4.3103	24.1432	[4.1495,9.3523]
OLS	10.6840	(1.7561,35.9274)	7.9431	4.3903	92.3723	[4.7254,13.3519]
MLE	9.1463	(1.2202,33.5482)	6.3981	3.1308	87.3000	[3.6175,11.3160]

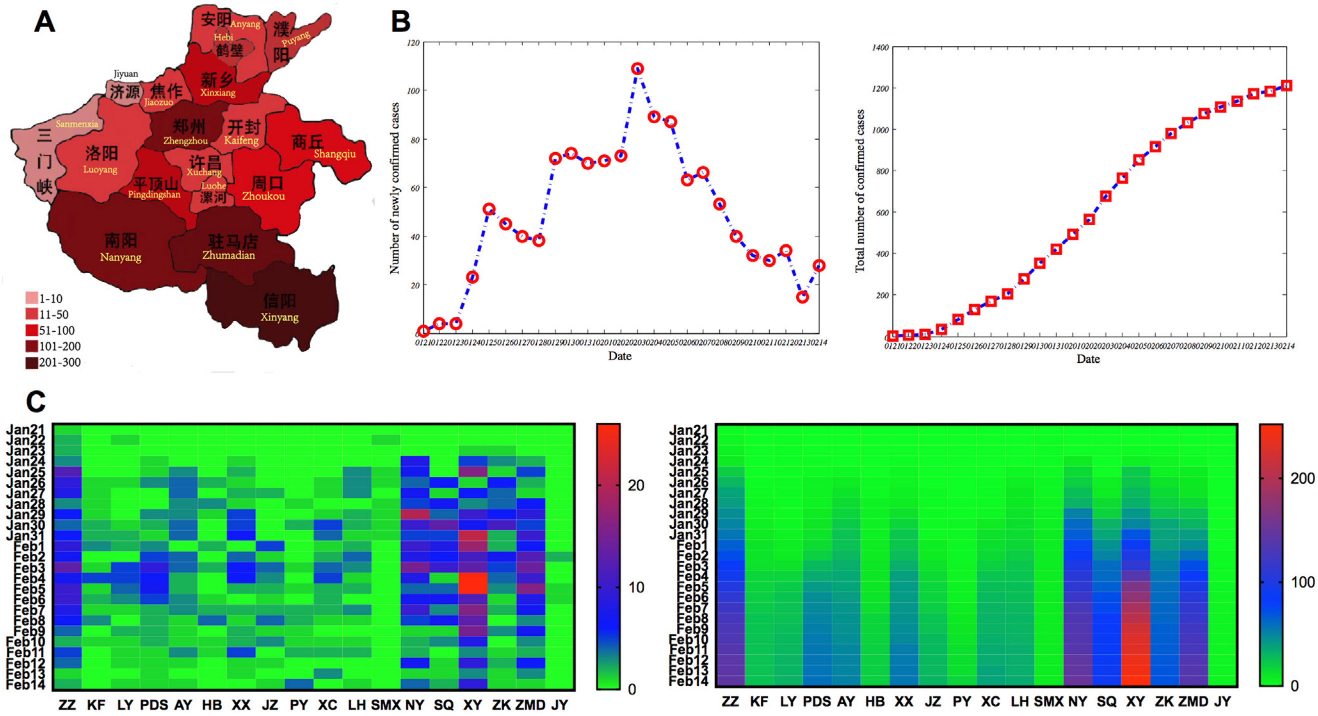
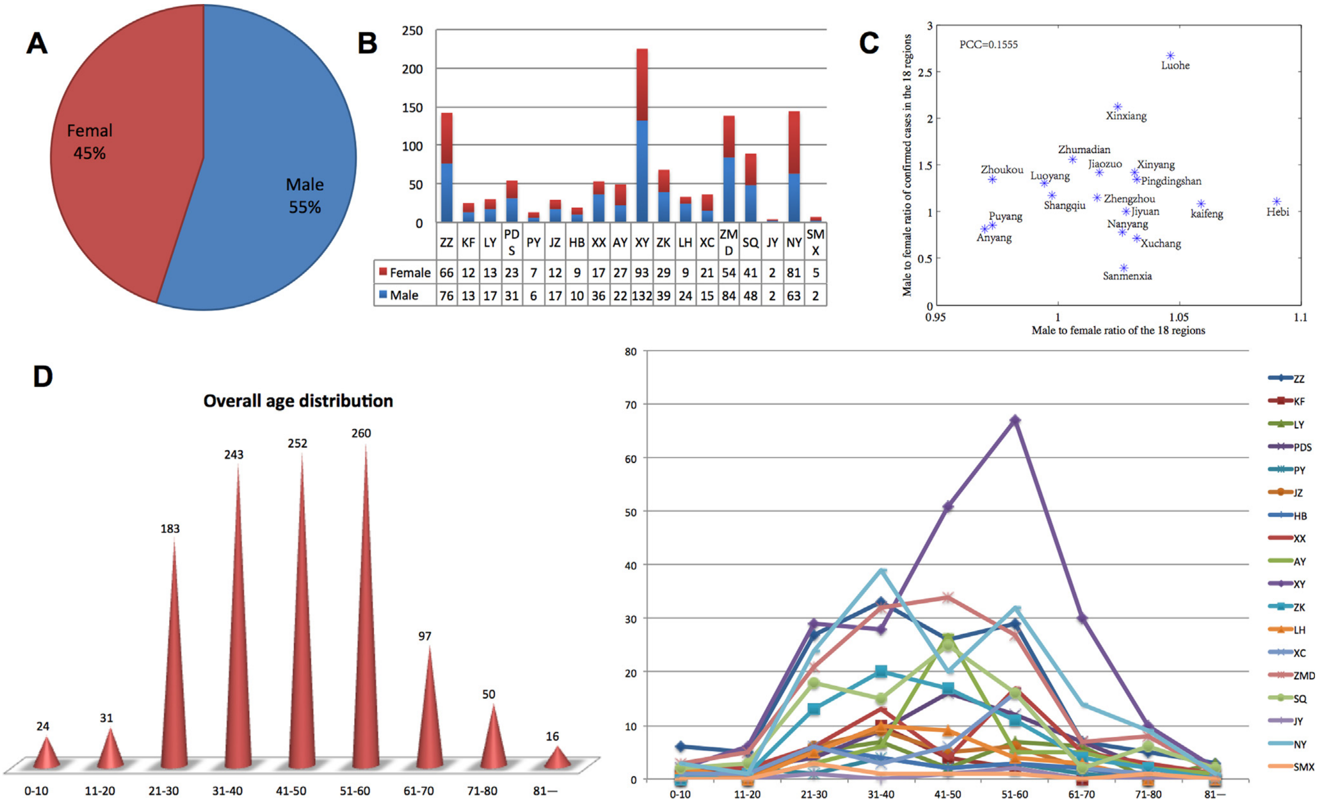


Fig. 1. Epidemic situation in Henan through February 14, 2020. A. Epidemic situation on February 14, 2020; B. The evolution of provincial daily increase in confirmed cases and provincial cumulatively confirmed cases. C. The heatmaps of daily increase in confirmed cases and cumulatively confirmed cases in 18 regions. Here, ZZ: Zhengzhou; KF: Kaifeng; LY: Luoyang; PDS: Pingdingshan; AY: Anyang; HB: Hebi; XX: Xinxiang; JZ: Jiaozuo; PY: Puyang; XC: Xuchang; LH: Luohe; SMX: Sanmenxia; NY: Nanyang; SQ: Shangqiu; XY: Xinyang; ZK: Zhoukou; ZMD: Zhumadian; JY: Jiyuan. Similarly hereinafter.



	ZZ	KF	LY	PDS	PY	JZ	HB	XX	AY	XY	ZK	LH	XC	ZMD	SQ	JY	NY	SMX
🟥 Female	66	12	13	23	7	12	9	17	27	93	29	9	21	54	41	2	81	5
🟦 Male	76	13	17	31	6	17	10	36	22	132	39	24	15	84	48	2	63	2

Fig. 2. Summary statistics on the gender and age distributions of confirmed cases in Henan. A. MFR in the whole province; B. MFRs in the 18 regions. C. Scatter plot of MFR versus MFPR in the 18 regions; D. Age distribution for confirmed cases in the whole province and its 18 regions.

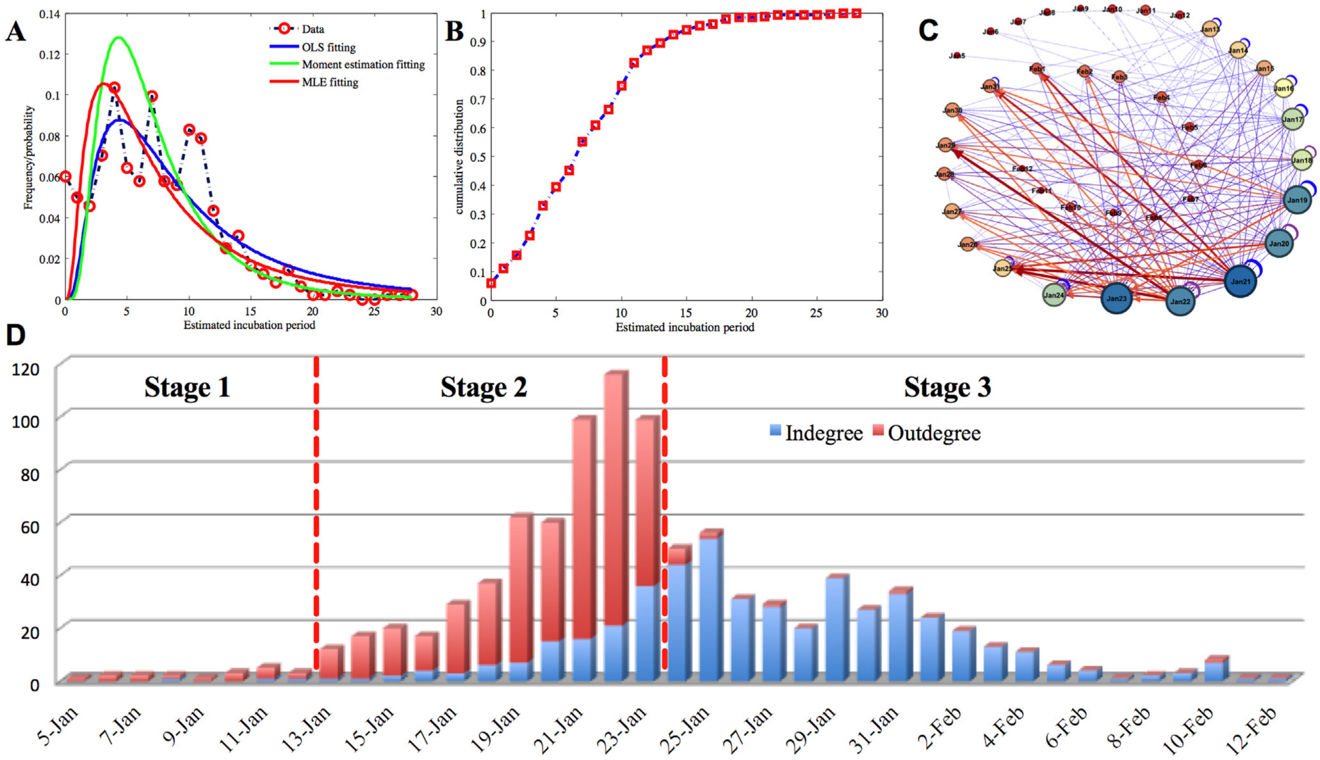


Fig. 3. Statistical results of estimated incubation periods for 483 confirmed patients in Henan. A. Frequency distribution of estimated incubation periods and the probability density curve for fitted incubation period. B. The cumulative frequency distribution of estimated incubation periods. C. Transfer diagram from exposure to infection for the 483 patients. Nodes at the two ends of an edge correspond to the exposed date and the date with clinical symptoms or diagnosis. Edges represent the transfer from the two dates, and thicknesses of edges are proportional to the numbers of patients. Self-loops mean that the dates for exposure and appearance of clinical symptoms or diagnosis were the same. D. Weighted indegree and outdegree distributions of the directed graph as shown in C. Weighted indegree denotes the total number of patients with clinical symptoms or diagnosis; while weighted outdegree represents the total number of exposed persons that will be confirmed to be infected later.

number of virus carriers is supposed to be proportional to the received amount of Wuhan personnel, which may be one of the main reasons why the mentioned regions in Henan are so severely affected.

3.2. Gender and age distribution

Among the 1212 patients, we extracted the gender information of 1158 patients (95.54%) and the age information of 1156 patients (95.38%).

Statistical results reveal that 637 out of the 1158 patients (55%) are male, and apparently higher than the number of females (Fig. 2A), although there are slight differences among different regions (Fig. 2B). We guess that three possible reasons may result in such a gender difference. Firstly, men may be more active and have wider social activities than women, which increases their risk of COVID-19. Secondly, an existing medical investigation reported that the expression and distribution of ACE2 was wider in male patients than in females (Zhao et al., 2020a). Similarly to SARS, it is reported that COVID-19 invades the human body through ACE2 receptors. Based on single cell RNA sequencing technology, researchers have investigated the expression profiles of ACE2 in two male and six female COVID-19 patients from single cell resolution, and they found that the expression of ACE2 was correlated with gender. The proportion of ACE2 expressed cells is higher in men than in women (1.66% vs. 0.41%)(Zhao et al., 2020a). Additionally, the distribution of ACE2 was also wider in men. Zhao et al. (Zhao et al., 2020a) reported that there were at least five different types of cells in the lungs of men with expression of the ACE2 receptors, while the number was about 2 to 4 in female patients. This may be one of the reasons why male patients were higher in number than female patients. Thirdly, one may doubt that the distribution difference may correlate with population structure. To verify whether the male to female ratio (MFR) affects the finding, based on the statistical yearbook report of Henan province in 2019, the MFR in the whole province was 1.0126:1, which is very different from the male to female patient ratio (MFPR) (637:521 = 1.2226:1). Moreover, we performed a correlation analysis for the two ratios in the 18 regions; the two ratios show very weak correlation, with Pearson correlation coefficient PCC = 0.1555 (Fig. 2C). Kolmogorov-Smirnov (KS) test (Hodges, 1957) on the MFR and MFPR in the 18 regions also shows that the two have a significant difference ($p = 0.0018 < 0.05$). Thus, we speculate that the MFR has no apparent correlation with the MFPR. In conclusion, the former two reasons may be the driving force for gender difference.

Ages of patients in the whole province or each region all roughly follow normal distributions (Fig. 2D). Patients between ages 21 and 60 years old make up more than 81% (938 out of 1156). There are many migrant workers and college students in this age interval. Moreover, those people may have wider social circles than the others. As a summary, people between ages 21 and 60 are at high risk, and we should give priority to migrant workers and college students in the prevention and control of COVID-19.

3.3. Statistical estimation of incubation period and state transition diagram

3.3.1. Definition of incubation period and statistical results

Traditionally, the incubation period is defined as the period between the infection of an individual by a pathogen and the manifestation of the illness or disease it causes (Incubation period,

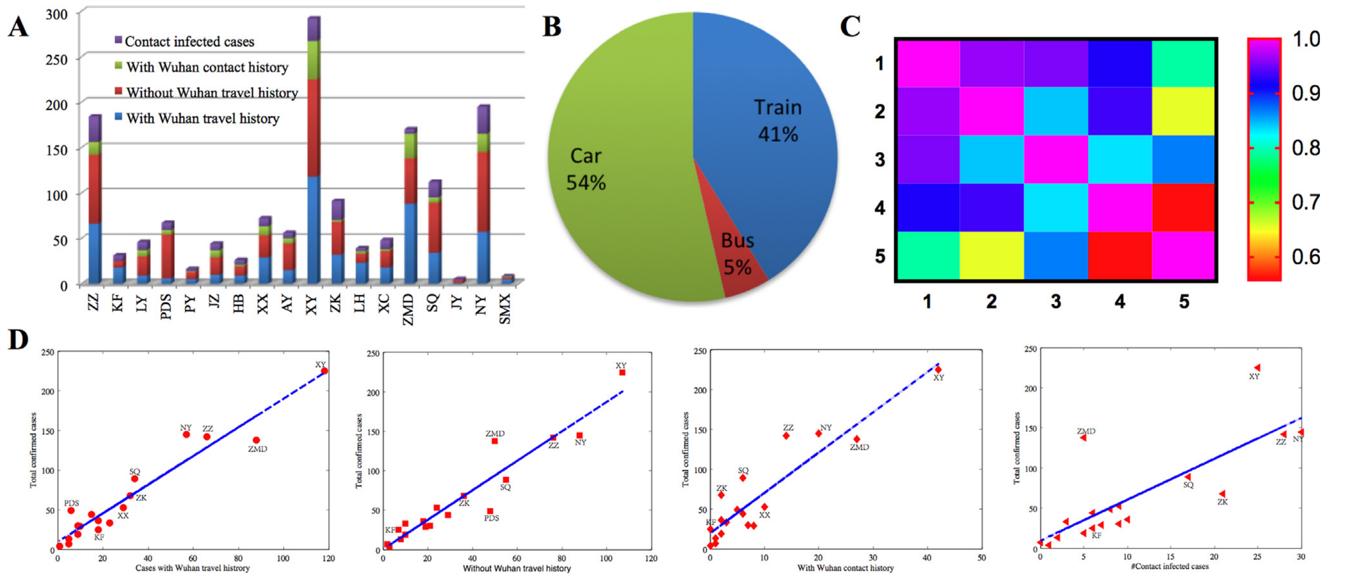


Fig. 4. COVID-19 infection and Wuhan travel histories. A. Bar-plots for the numbers of four categories of patients in the 18 regions of Henan. B. Transportation manners of 250 patients that came from Wuhan. C. The heat-map of the correlation matrix among $Y, X_i (i = 1, 2, 3, 4)$ (corresponding to 1-5). D. The scatter plots of Y versus X_1-X_4 respectively and the corresponding fitted linear regression lines.

2020; Armenian and Lilienfeld, 1983). Different infectious diseases have different incubation periods. However, for a certain infectious disease, its incubation period is relatively fixed. Since the numbers of pathogens entering the body, virulence and reproductive capacity, as well as power of resistance are varied for different people, the incubation periods for patients with a certain disease may follow logarithmic normal distribution (Armenian and Lilienfeld, 1983). Generally, incubation period can be measured by physiological observations and biological experiments (Armenian and Lilienfeld, 1983). The determination of incubation period has great implications for disease control and policy making.

The estimation of the incubation period of COVID-19 is a very difficult task. The main difficulty is that it is difficult to determine the infected time. Based on patients' information, Yang et al. (Yang et al., 2020) reported that the median incubation period of COVID-19 is 4.75 days; the interquartile range is 3.0-7.2 days. Hereinafter, we try to estimate the incubation period of patients in Henan. The collected data are of varied quality, many descriptions of patients only told us when the person left Wuhan, or when she/he had suspicious contacts with persons from Wuhan or suspicious persons (hereinafter, we name these as exposed). To facilitate the estimation of the incubation period, we define estimated incubation period as the period from the date of exposure to the date of appearance of clinical symptoms or making a definite diagnosis. Apparently, such a definition overestimates the actual ones. For example, the information from a patient (Mr Zhou, No.2 patient) in Nanyang said that his son returned from Wuhan on January 6; he and his son were disease-free for a long period. On February 3, he became sick and was confirmed with infection. His son has been disease-free until now. From our definition, his incubation period is 28 days, which is surprisingly high. We note that our definition may be a little higher than the actual case, since we cannot exclude the possibility that the person has contacted some other infected persons. That is, we cannot know the intermediate spreader (unknown person B), and when she/he transmitted the virus from person A to C.

Among the collected data, incubation periods of 483 confirmed patients could be estimated. Statistical results are shown in Fig. 3A, B and Table 1. The estimated incubation periods roughly follow logarithmic normal distribution. Specifically, the average estimated incubation period was 7.4 days, the mode was 4 days (50 cases) and the median was 7 days (48 cases). About 55% and 92% of patients had incubation periods no more than 7 and 14 days, respectively. About 7.45% patients were overestimated with longer than 14-day incubation periods. 285 of the 483 patients were men and 198 were women, their average incubation periods were 7.8 and 7.2 days. KS test revealed that men and women have no significant difference ($p > 0.05$) on incubation periods, which was different from existing findings (Xiong et al., 2020).

The fitted PDF curves based on three statistical methods can be found in Fig. 3A, and some theoretical results are summarized in Table 1. Comparing among the three methods, the fitted curve from the ME method has the smallest variance; the variances of the OLS and MLE approaches are both very high. High variance indicates higher proportions of patients with long incubation periods. The ME method may well mimic the actual long right tail of the incubation periods, while the OLS estimation could well describe the cases with middle incubation periods (2-14). The estimated interquartile range from the ME method is [4.1495, 9.3523], which is close to the result from the 483 patients ((Yang et al., 2020; Tang et al., 2020)).

The theoretical modeling of incubation periods further illustrates that they may follow the logarithmic normal distribution. Moreover, it can be used to estimate the actual statics of incubation periods, and provide valuable references for decision-making. Though the estimated incubation periods may be longer than the actual cases, the distribution has some implications. It indicates that a few patients may have very long incubation period, which increases the difficulty of controlling and preventing COVID-19, and strong quarantine measures still may be needed for a period of time after the last confirmed case is reported.

3.3.2. State transition diagram reveals the time-phased nature of epidemic

The epidemic of COVID-19 in Henan has undergone three stages (Fig. 3C, D). For the first stage before January 13, the numbers of both exposed and confirmed patients were not high, and the number of exposed (outdegree) patients was higher than confirmed patients (indegree). For the second stage, between January 13 and 23, the disease was in outbreak period, numbers of both exposed and

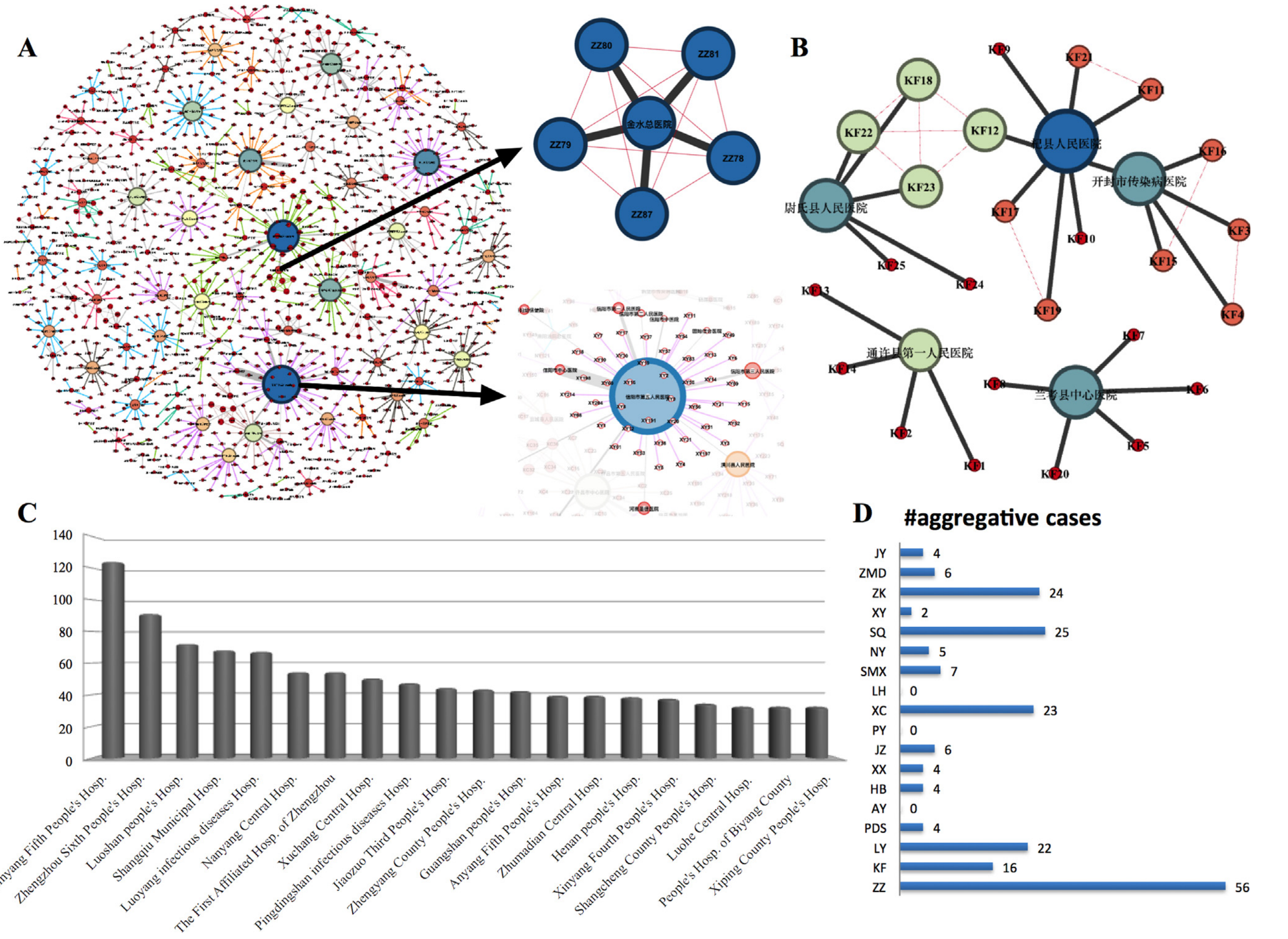


Fig. 5. Network analysis among patients, hospitals and their relationships. A. A heterogeneous network that contains 1105 patients, 248 hospitals, 123 inter-hospital transfer relationships that involved 206 patients and 208 clustering infected patients. Nodes include patients and hospitals. Edges among patients represent relationship between relatives, friends or colleagues; Edges between patients and hospitals denote the patients were treated in the related hospitals; while edges among hospitals indicate the inter-hospital transfer treatments. Sizes of nodes were proportional to nodes' degree. The thickness of edges was proportional to the numbers of patients. B. The heterogeneous network for patients in Kaifeng. C. The top-20 hospitals with the most patients in treatments. D. Distributions of the 208 aggregate outbreak patients in the 18 regions.

confirmed cases were considerable, and the number of exposed cases was far more than the confirmed ones. During this period, many college students and migrant workers in Wuhan returned to their hometowns; the large number of people flowing from Wuhan increased the risks of exposure and infection. After January 23, the epidemic entered the third stage, daily-confirmed cases were higher than exposed ones, and the confirmed cases roughly decreased with time. This may be attributed to the adoption of prevention and control measures. With the closing of Wuhan city on January 23, many regions of Henan quickly adopted very strict measures to prevent the spread of COVID-19. Thus, after January 23, the number of exposed cases became fewer and fewer (Fig. 3D), which is a good phenomenon, indicating that the prevention and control measures were gradually working.

3.4. Correlation with Wuhan travel histories

For the 18 regions of Henan province, we also summarized the numbers of currently confirmed patients with Wuhan travel histories (noted as vector X_1), the numbers of patients without Wuhan travel histories (X_2), the numbers of patients that had contacts with persons from Wuhan (X_3), as well as the numbers of patients that were without Wuhan travel histories, but who had contacts with suspicious persons (X_4). The bar-plots for the four categories of patients in the 18 regions are shown in Fig. 4A (In total of 1149 patients). Among patients with Wuhan travel histories, the released information of 250 patients contained transportation information. We found that 46% of the 250 patients had recently travelled by train or bus; these patients may transmit COVID-19 to a lot of people in the same train or bus.

We investigated the correlation between the current total numbers of confirmed patients (Y) with X_1-X_4. The heatmap of the correlation matrix is shown in Fig. 4C. It reveals that Y has the highest correlation with X_1, the PCC between the two is 0.9643, the fitted regression equation (Samuels et al., 2016) is

$$Y = 1.8146X_1 + 9.0920. \tag{6}$$

Such linear relationship statistically holds ($R^2 = 0.9299$, $p < 0.001$). However, although there are certain correlations between Y and X_2-X_4, their strength of correlations are lower than that between Y and X_1. This indicates that persons with Wuhan travel histories have high risks of infection, and regions with more of such persons tend to encompass more COVID-19 patients.

3.5. Network analysis on aggregate outbreak phenomena

From the 1212 patients, we constructed a heterogeneous network and performed some analysis on the network (Fig. 5). The network contains 1105 patients that have been treated in 248 hospitals, 123 inter-hospital transfer relationships that involved 206 patients, and 208 patients that were clustering infected.

A few hospitals encompass a large number of patients in treatments, and the aggregate outbreak phenomena were ubiquitous (Fig. 5). For example, five patients that were treated in Zhengzhou Jinshui General Hospital are relatives of each other, who were successively infected. Similarly, 12 patients in Kaifeng are clustering infected. In fact, a total of 208 patients were clustering infected, and Zhengzhou, Shangqiu, Zhoukou, Xuchang and Luoyang encompassed the most aggregative outbreak patients. Among the designated hospitals in Henan province, people's hospitals in different cities encompass a great number of patients. The Fifth People's Hospital of Xinyang received the most patients in Henan, followed by the The Sixth People's Hospital of Zhengzhou and the Luoshan County People's Hospital. Network analysis can provide us important information about patients, hospitals and their relationships; it can also provide valuable guidance for the distribution of epidemic prevention materials.

4. Discussion

Based on publicly released patient data by various CDCs in Henan province during January 21 to February 14 of 2020, we performed epidemiological analysis on COVID-19 patients. Newly confirmed cases reached a peak at February 3, and started to fall after that day. Among the 18 regions of Henan, Xinyang has the heaviest epidemic. We found that 55% of patients in Henan were male, which is higher than females. The gender difference of patients in Henan is 1.2226:1, while Yang et al.(Yang et al., 2020) reported that the gender difference for 4021 patients from over 30 provinces is 1.1481:1. We reported that the MFRs of the whole province and its 18 regions are not the main factor that results in the gender difference in patients. Two possible factors include 1) the expression and distribution of ACE2s were wider in male than in female patients, and 2) men may be more active and have wider social activities than females in many regions of Henan. The ages of patients generally follow normal distribution, and more than 81% of patients are ages 21-60 years old. Our analysis suggested that migrant workers or college students might be the key crowds during the prevention and control of COVID-19 in Henan.

Incubation period is very important but very difficult to be estimated. We define the period from first contact with suspicious persons or recent return from Wuhan to first appearance of clinical symptoms (fever, cough etc) as incubation period, and we estimate its distribution. Statistical analysis on 483 patients reveals that the average estimated incubation period was 7.4 days; the mode was 4 days. About 55% and 92% of patients had incubation periods no more than 7 and 14 days. Incubation periods of about 7.45% of patients were overestimated with more than 14 days. Based on ME, OLS and MLE approaches, we modelled the incubation period as logarithmic normal distributions, and we found the ME method can closely mimic the long right tails of the PDF for the incubation periods. Based on the ME approach, interquartile range of theoretically estimated incubation period is between 4.1 and 9.4 days. Different from existing findings (Buhlmann and Geer, 2011), among the 483 patients in Henan, we found there was no significant difference on incubation periods between men and women. Due to our definition, the estimated values are unavoidably higher than the actual cases; however, the distribution and statistics from the estimated ones are undoubtedly meaningful for real-world decision-making.

State transition diagram reveals that the epidemic of COVID-19 in Henan has undergone three stages, which could be explained by the strategies that we adopt at each stage. We also found that the numbers of patients with recent Wuhan travel histories are highly correlated with the total confirmed cases in the 18 regions of Henan. During our analysis, we considered the linear relationship between Y with each X_i, mainly for the purpose of avoiding multicollinearity (Buhlmann and Geer, 2011) among $X_i(i = 1,2,3,4)$. Furthermore, we performed network analysis on aggregate outbreak phenomena; we reported that 208 cases were clustering infected ones. Various people's Hospitals in Henan are the main forces in treating patients.

It is noted that, since the data quality from different regions was varied, only some of the 1212 patients were used for certain questions. We also noted that the 1212 patients' data were collected before February 14, 2020; after that day, only tens of patients were confirmed, thus, the considered data are close to full sample in Henan. Furthermore, although we only considered patients in Henan, the obtained conclusions are suitable for other places in the world. The related investigations may provide valuable suggestions and guidance for the prevention and control of COVID-19 in other regions of the world.

Conflict of Interests

The authors declare that they have no conflicts of interest to this work.

Funding

This work was supported by the National Natural Science Foundation of China [Grant numbers 61773153, 61773175, 61573262]; the Supporting Plan for Scientific and Technological Innovative Talents in Universities of Henan Province [Grant number 20HASTIT025], and the Training Plan of Young Key Teachers in Colleges and Universities of Henan Province [Grant number 2018GGJS021]; partly supported by the supporting grant of Bioinformatics Center of Henan University [Grant nubmer 2018YLJC03]. We thank associate professor Shaoli Wang and Aimin Chen at Henan University for valuable discussion. We also want to express our grattitude for the valuable comments and feedback from the readers of the first version of the paper, which was released on the WeChat official accounts of the Swarma Club at February 13, 2020.

Ethical Approval

This study did not involve laboratory animals or private matter of patients; all data are publicly available.

References

Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. N. Engl. J. Med. 2020;382:727–33.

Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020;, doi:http://dx.doi.org/10.1016/S0140-6736(20)30183-5.

Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020;579:270–3.

Yang Y, Lu Q, Liu M, Wang Y, Zhang A, Jalali N, et al. Epidemiological and clinical features of the 2019 novel coronavirus outbreak in China. medRxiv 2020;, doi:http://dx.doi.org/10.1101/2020.02.10.20021675 20021675 [Preprint]; February 11, 2020 [cited 2020 Feb 28].

Zhao Y, Zhao Z, Wang Y, Zhou Y, Ma Y, Zuo W. Single-cell RNA expression profiling of ACE2, the putative receptor of Wuhan 2019-nCov. bioRxiv 919985; [Preprint]. , doi:http://dx.doi.org/10.1101/2020.01.26.2020.919985 January 26, [cited 2020 Feb 28].

Xu X, Chen P, Wang J, Feng J, Zhou H, Li X, Zhong W, Hao P. Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. Sci China Life Sci 2020a;63:457–60.

Yan Y, Chen Y, Liu K, Luo X, Xu B, Jiang Y, Cheng J. Modeling and prediction for the trend of outbreak of NCP based on a time-delay dynamic system. Scientia Sinica Mathematica 2020;50:385–92 (In chinese).

Zhang YP. Analysis of Epidemiological characteristics of new coronavirus pneumonia. Chin. J. Epidemiol. 2020;41(2):1–7.

Xu X, Wen C, Zhang G, Sun H, Liu B, Wang X. The geographical destination distribution and effect of outflow population of Wuhan when the outbreak of the 2019-nCoV Pneumonia. J. Univ. Electron. Sci. Tech. China. 2020b;, doi:http://dx.doi.org/10.12178/1001-0548.2020033 (In chinese).

Du Z, Wang L, Chauchemez S, Xu X, Wang X, Cowling BJ, et al. Risk for transportation of 2019 novel coronavirus disease from Wuhan to other cities in China. Emerg. Infect. Dis. 2020;, doi:http://dx.doi.org/10.3201/eid2605.200146.

Tang B, Wang X, Li Q, et al. Estimation of the transmission risk of 2019-nCov and its implication for public health interventions. SSRN 2020;, doi:http://dx.doi.org/10.2139/ssrn.3525558 [Preprint]. January 27. [cited 2020 Feb 28].

Liu P, Chen W, Chen J. Viral metagenomics revealed sendai virus and coronavirus infection of *Malayan Pangolins* (*Manis javanica*). Viruses 2019;11(11):979.

Lu JA, Wang P. Understanding novel coronavirus pneumonia from the small world, scale-free and high clustering properties of complex networks. Swarma Club 2020; Feb. 6, 2020; https://swarma.org/?p=18300. Accessed on Feb. 25.

Zhao S, Zhuang Z, Ran J, et al. The association between domestic train transportation and novel coronavirus (2019-nCoV) outbreak in China from 2019 to 2020: A data-driven correlational report. Travel Med. Infect. Dis. 2020b;33:101568.

Liu Z, Magal P, Seydi O, Webb G. Understanding unreported cases in the 2019-nCov epidemic outbreak in Wuhan, China, and the importance of major public health interventions. SSRN, 3530969; [Preprint]. February 4. , doi:http://dx.doi.org/10.2139/ssrn.3530969 [cited 2020 Feb 28].

Wang P, Lu J, Jin Y, Zhu M, Wang L, Chen S. Epidemiological characteristics of 1212 COVID-19 patients in Henan, China. medRxiv 20026112; [Preprint]. February 25. , doi:http://dx.doi.org/10.1101/2020.02.21.20026112 [cited 2020 Feb. 28].

Chen D, Xu W, Lei Z, Huang Z, Liu J, Gao Z, Peng L. Recurrence of positive SARS-CoV-2 RNA in COVID-19: A case report. Int. J. Infect. Dis. 2020;93:297–9.

Zhang S, Diao M, Yu W, Pei L, Lin Z, Chen D. Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven analysis. Int. J. Infect. Dis. 2020;93:201–4.

Pastor-Satorras R, Vespignani A. Epidemic spreading in scale free networks. Phys. Rev. Lett. 2001;86:3200–3.

Kitsak M, Gallos LK, Havlin S, et al. Identification of influential spreaders in complex networks. Nat. Phys. 2010;6:888–93.

Wang P, Tian C, Lu J. Identifying influential spreaders in artificial complex networks. J. Syst. Sci. & Complex. 2014a;27:650–65.

Wang P, Chen Y, Lu J, et al. Graphical features of functional genes in human protein interaction network. IEEE Trans. Biomed. Circuits Syst. 2016;10(3):707–20.

Wang P, Lu J, Yu X. Identification of important nodes in directed biological networks: a network motif approach. PLoS One 2014b;9(8):e106132.

Wang P, Wang D, Lu J. Controllability analysis of a gene network for Arabidopsis thaliana reveals characteristics of functional gene families. IEEE-ACM Trans. Comput. Biol. Bioinformat. 2019;16(3):912–24.

Lu LY, Chen D, Ren X, et al. Vital nodes identification in complex networks. Phys. Rep. 2016;650:1–63.

Zhang ZK, Liu C, Zhan XX, et al. Dynamics of information diffusion and its applications on complex networks. Phys. Rep. 2016;651:1–34.

Wei X, Wu X, Chen S, Lu J, Chen G. Cooperative epidemic spreading on a two-layered interconnected network. SIAM J. Applied Dyn. Syst. 2018;17(2):1503–20.

Xu S, Wang P, Zhang CX, Lu J. Spectral learning algorithm reveals propagation capability of complex network. IEEE Trans. Cyber. 2019;49(12):4253–61.

Pastor-Satorras R, Castellano C, Van Mieghem P, et al. Epidemic processes in complex networks. Rev. Mod. Phys. 2015;87:925–46.

Hodges Jr. JL. The significance probability of the Smirnov two-sample test. Arkiv for Matematik 1957;3:469–86.

Incubation period. Merriam-Webster.com dictionary, Merriam-Webster. https://www.merriam-webster.com/dictionary/incubation%20period. Accessed 12 Feb. 2020.

Armenian HK, Lilienfeld AM. Incubation period of disease. Epidemiologic Rev. 1983;5(1):1–15.

Buhlmann P, Geer Svan de. Statistics for high-dimensional data: methods, theory and applications. Berlin Heidelberg: Springer; 2011.

Samuels ML, Witmer JA, Schaffner AA. Statistics for the life sciences. Pearson; 2016.

Xiong Q, Xu M, Zhang J, et al. Women may play a more important role in the transmission of the corona virus disease (COVID-19) than men. Preprints with The Lancet 2020; [Preprint]. March 3, 2020 [cited 2020 March 3].