


---

## Research and Applications

# Mining relationships between transmission clusters from contact tracing data: An application for investigating COVID-19 outbreak

Tsz Ho Kwan <sup>1</sup>, Ngai Sze Wong<sup>1,2</sup>, Eng-Kiong Yeoh<sup>3</sup>, and Shui Shan Lee<sup>1</sup>

<sup>1</sup>Stanley Ho Centre for Emerging Infectious Diseases, The Chinese University of Hong Kong, Shatin, Hong Kong, <sup>2</sup>Jockey Club School of Public Health and Primary Care, The Chinese University of Hong Kong, Shatin, Hong Kong, and <sup>3</sup>Centre for Health Systems and Policy Research, The Chinese University of Hong Kong, Shatin, Hong Kong

Corresponding Author: Shui Shan Lee, FRCP, Room 207, Postgraduate Education Centre, Prince of Wales Hospital, Shatin, Hong Kong; sslee@cuhk.edu.hk

Received 5 May 2021; Revised 12 July 2021; Editorial Decision 2 August 2021; Accepted 4 August 2021

### ABSTRACT

**Objective:** Contact tracing of reported infections could enable close contacts to be identified, tested, and quarantined for controlling further spread. This strategy has been well demonstrated in the surveillance and control of COVID-19 (coronavirus disease 2019) epidemics. This study aims to leverage contact tracing data to investigate the degree of spread and the formation of transmission cascades composing of multiple clusters.

**Materials and Methods:** An algorithm on mining relationships between clusters for network analysis is proposed with 3 steps: horizontal edge creation, vertical edge consolidation, and graph reduction. The constructed network was then analyzed with information diffusion metrics and exponential-family random graph modeling. With categorization of clusters by exposure setting, the metrics were compared among cascades to identify associations between exposure settings and their network positions within the cascade using Mann-Whitney *U* test.

**Results:** Experimental results illustrated that transmission cascades containing or seeded by daily activity clusters spread faster while those containing social activity clusters propagated farther. Cascades involving work or study environments consisted of more clusters, which had a higher transmission range and scale. Social activity clusters were more likely to be connected, whereas both residence and healthcare clusters did not preferentially link to clusters belonging to the same exposure setting.

**Conclusions:** The proposed algorithm could contribute to in-depth epidemiologic investigation of infectious disease transmission to support targeted nonpharmaceutical intervention policies for COVID-19 epidemic control.

**Key words:** COVID-19, social network analysis, data mining, algorithms, medical informatics

---

## INTRODUCTION

The role of contact tracing is paramount when an infectious agent has been newly introduced into a jurisdiction. Immediate identification of close contacts of the infected persons followed by subsequent quarantine and testing could break the chain of transmission, if these can be implemented during the short but highly infectious pe-

riod as demonstrated in the coronavirus disease 2019 (COVID-19) epidemics caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).<sup>1</sup> An investigation of the chains of transmission further reveals the degree of spread and the presence of multiple transmission clusters and their interrelationships.<sup>2</sup> The process of contact tracing uncovers a broad range of data covering index cases,

their contacts, and the characteristics of clusters formed in the course of time. With such big data, further exploration in mining would generate useful information for enhancing epidemiologic understanding of the COVID-19 outbreaks.

Contact tracing data constituted one unique form of big data of public health importance. Relationships from large dataset could be mined from large datasets using different approaches, such as coauthorship,<sup>3</sup> spatiotemporal co-occurrence,<sup>4</sup> and mutual interests.<sup>5</sup> Mined networks could be used for identifying advisory relationships between academics,<sup>6</sup> uncovering potential drugs for diseases,<sup>7</sup> and inferring social support for improving adherence to medical treatment.<sup>8</sup> Through systematic analyses, it is possible to further quantify and characterize the importance of transmission role of patients in the networks as has been applied in COVID-19<sup>9</sup> and human immunodeficiency virus (HIV)<sup>10</sup> epidemics. One potentially impactful outcome for such research could be the design of a useful tool for predicting future transmission such that targeted intervention could be developed and adopted for disease control.<sup>11,12</sup>

So far, most of the published works that explored transmission networks using contact tracing data have focused on the relationships between individuals to identify superspreaders.<sup>9,13,14</sup> Limited studies were concerned with the macroscopic linkages among multiple transmission clusters, the understanding of which could be important in guiding public health interventions.<sup>15</sup> This article describes an algorithm for constructing a network of transmission clusters from contact tracing data by capturing both temporal and network topological features for social network analysis to identify SARS-CoV-2 clusters of public health importance.

## MATERIALS AND METHODS

### Data

#### Contact tracing data

For each laboratory confirmed SARS-CoV-2 infected patient identified in Hong Kong, the Centre for Health Protection of the Department of Health conducted contact tracing interviews to identify persons potentially exposed to the virus for quarantine and testing. If 2 or more persons had been to the same place, or had their contact relationship confirmed, they were assigned to a transmission cluster. The index patient of each cluster would have been identified. Together with demographics, clinical data including symptom onset date and reporting date were retrieved from the surveillance data. Each epidemiologically unlinked cluster or “cascade” containing multiple linked clusters was given a unique identifier to facilitate processing. Data access approval was obtained from the Department of Health. This study was approved by the Survey and Behavioural Research Ethics Committee of The Chinese University of Hong Kong (Ref. no. SBRE-19-595).

#### Exposure setting

Following the prior definition,<sup>15</sup> each transmission cluster was classified into 1 of the following 5 categories of exposure settings: residences, daily activities, social activities, work/study, and healthcare. These categories were differentiated by population size, relationship type, environment, and exposure duration and frequency, and they could be subdivided into specific subsetting type. Residences included co-living and non-co-living households, dormitories, hotels and neighborhood. Daily activities included eateries, shopping, transportation, and other often one-to-one personalized services. Mass events, parties, religious gatherings, and entertainment were

classified as social activities. Workplaces, schools, and training sessions were in the work/study category. Last, healthcare setting included long-term-care facilities and public and private health services.

## Algorithm design

### Purpose

The objective of the algorithm was to construct a directed network with minimum edges from contact tracing data to evaluate relationships between transmission clusters. The contact tracing data were used to construct affiliation network composed of SARS-CoV-2 infected patient belonging to 1 or more clusters. As the assessment of cluster-cluster relationships required a one-mode network, it was necessary to transform the two-mode network into a one-mode network.

### Data structure

Prerequisites include 2 column vectors and a matrix. A bit vector  $I$  of a size of number of patients indicated the index patient of each cluster by 1, otherwise 0. There can only be 1 index per cluster. The second column vector  $D$  contains positive integers denoting unique cascade identifier. The matrix  $M$  coded the relationship between patients and clusters with rows denoting patients and columns denoting generation of clusters. Each cluster id can only appear in one column, and the placement of column is assigned by generation. For example, the first column stores all primary generation clusters and the second column marks secondary generation clusters. Therefore,  $M_{ij}$  is the cluster identifier, in contrast to an adjacency matrix in which it could only be a binary value or weight. It is instead akin to adjacency list in which each list describes a node's relationship, with a notable variation here that the order matters. Empty cells are replaced by 0. No subclusters shall be included in the matrix or they have to be removed before processing.

### Algorithm

The algorithm could be described in 3 steps: horizontal edge creation, which forms the backbone of the network; vertical edge consolidation; and last, graph reduction (Figure 1).

#### Horizontal edge creation

The principle is to make an out-star or a directed path from each patient horizontally across columns, then merge them all to become a network. Let  $S$  be the index clusters vector,  $T$  be the processed clusters vector, and  $E$  be an edge list. Given the 3 prerequisites,  $M$  will first be subset per cascade as  $K$ , then the index patient  $i$  will be identified and the primary cluster  $j$  will be defined. All nonzero values in row  $M_i$  will be put into  $S$ , which then has a higher precedence over  $T$  in the next step. If there are more than 1 nonzero values, the primary cluster will link to each and every other clusters, otherwise it will link to a value -1. All links are put into  $E$  as an edge list. After creating edges, the entire row will be removed from  $K$ . While the size of matrix  $K$  is not 0, each row will be processed for edge creation before removing from  $K$ . Edge creation for nonindex patients started with checking if a patient's affiliating cluster has been processed, first evaluating the values in  $S$  then in  $T$ . If not, the row will be ignored and check the next until there is one. When found, the primary cluster will be defined as the one that linked with any in  $S$  or  $T$  according to the sequence within the vector. This design preserves the order of occurrence of cluster since the index cluster. Similarly, the primary cluster will link to all other clusters, or -1 if there is only 1. All processed clusters will then be appended to  $T$ . After

**Algorithm 1: Horizontal edge creation (M, I, D)**

Input: Patient-cluster matrix M, Index vector I, Cascade identifier vector D

Output: Edge list E

1. For each cascade  $d \in D$  do
2.    $K = M_{D=d}$
3.    $\hat{I} = I_{D=d}$
4.    $\hat{i} = i \in \hat{I}: \hat{i}_i = 1$
5.    $\hat{j} = \min(j \in \mathbb{Z}^+: K_{ij} \neq 0)$
6.   Put  $K_{ij}$  into S
7.   If the patient is associated with only 1 cluster then
8.     Put  $\langle K_{ij}, -1 \rangle$  into E
9.   Else
10.     Put  $\langle K_{ij}, K_{ij} \rangle$  into E where  $j \neq \hat{j} \cap K_{ij} \neq 0$
11.     Put  $K_{ij}$  into S where  $j \neq \hat{j} \cap K_{ij} \neq 0$
12.   Remove  $K_{i^*}$  from K
13.   While  $|K| > 0$  do
14.     For each row  $i$  in K do
15.        $\hat{j} = \min(j \in \mathbb{Z}^+: K_{ij} = 1 \cap K_{ij} \in S)$
16.       If  $\hat{j}$  then
17.          $\min(j \in \mathbb{Z}^+: K_{ij} = 1 \cap K_{ij} \in T)$
18.         If  $\hat{j}$  then
19.         Next  $i$
20.     For each  $j$  in  $1..|K_{i^*}|$  where  $K_{ij} = 1 \cap j \neq \hat{j}$  do
21.       Put  $\langle K_{ij}, K_{ij} \rangle$  into E
22.       Put  $K_{ij}$  into T
23.     If no edges were newly created then
24.       Put  $\langle K_{i^*}, -1 \rangle$  into E
25.     Remove  $K_{i^*}$  from K

**Algorithm 2: Vertical edge consolidation (E, N)**

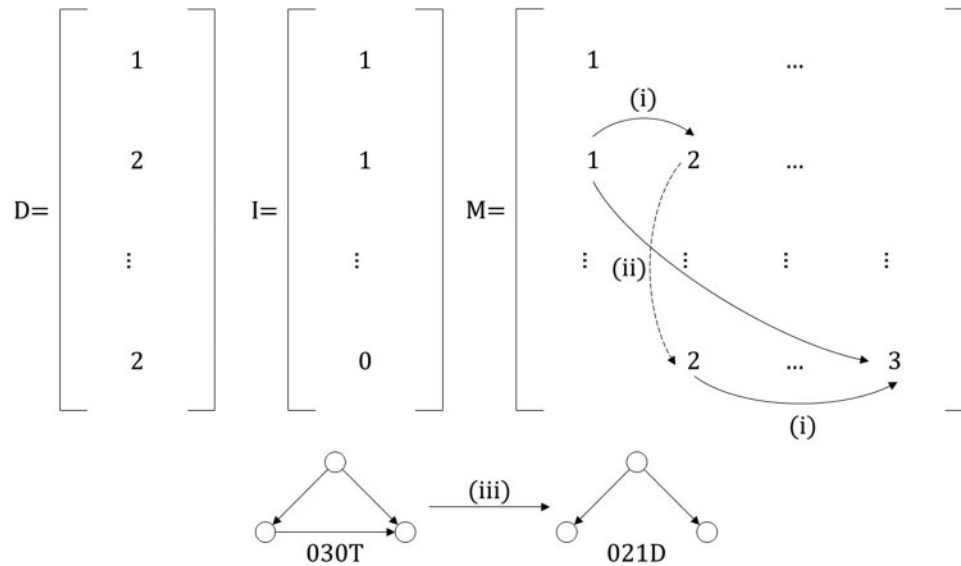
Input: Edge list E, Simple graph N

Output: Simple graph N

1. Put all clusters without a target node of -1 in E into Q
2. Find all shortest paths from the index to all leaf nodes in N
3. For each cluster  $q \in Q$  do
4.   If  $q$  does not present in any shortest paths then
5.     Remove node  $q$  from N
6.     Remove  $q$  from Q
7. While  $|Q| > 0$  do
8.   Find all shortest paths from the index to all leaf nodes in N
9.   For each shortest path P in descending order of path length  $> 2$  do
10.    For each node  $p \in P$  in reverse sequence do
11.     If node  $p-1 \in Q$  then
12.       If node  $p-2 \notin Q$  then
13.         For each node  $r$  being pointed by node  $p-1$  do
14.         Add an edge from node  $p-2$  to  $r$  in N
15.         Remove node  $p-1$  from N
16.         Remove node  $p-1$  from Q
17.     If an edge was drawn then
18.       Break For

processing all cascades, edges in E can then be amalgamated to construct a directed multigraph  $N'$  with each connected component as a cascade and number of edges between 2 nodes as number of patients

affiliating with both clusters.  $N'$  is subsequently simplified to become a simple graph N, after removing the node -1 and all associated edges.



**Figure 1.** Illustration of data structure and algorithm. Vectors  $D$  and  $I$  and matrix  $M$  are inputs in Algorithm 1. Step (i) illustrated horizontal edge creation, while step (ii) denoted consolidation of 2 horizontal edges into a vertical one. Step (iii) graphically showed the graph reduction process by removing the transitive edge from 030T to become a 021D triad.

### Vertical edge consolidation

Some clusters did not contain any patients uniquely belonging to it but served the bridging role between 2 other clusters. For example, patients A and B belonged to cluster X, patients C and D belonged to cluster Z, and patients B and C belonged to cluster Y. Clusters X and Z were not directly linked but were related through Y. The intermediate cluster Y had a feature in  $E$  that they do not have a target node of  $-1$ . This step aims at consolidating these relationships vertically across patients.

Let  $Q$  be a vector of intermediate clusters. The topology of a transmission cascade was likened to that of a tree. All shortest paths from root to leaf were first identified. Any intermediate clusters which did not exist in any shortest paths were removed. It is noted that intermediate clusters can only be present in paths of length 3 or above by definition. Starting from the longest shortest path from root to leaf, traverse back from the leaf node to check if the upper node is an intermediate node and the one further up being nonintermediate, remove the intermediate node by linking the upstream node and downstream ones if so. If there are 2 adjacent intermediate nodes, the one closer to the leaf would retain, then continue the traversal. After processing the path, the list of shortest paths will be regenerated and the whole process of traversal would be repeated until all intermediate clusters are dealt with.

### Graph reduction

The network so far consisted of 021D (out-star), 021C (directed line), and 030T (transitive semicycle) triads. With the goal of minimizing the number of edges to achieve an arborescence, or a rooted directed tree with all out-going edges, a reduction step was required for each cascade. Our design minimized transmission cascade's range but maximize its scale by removing the edge between 2 child nodes in a 030T triad to become a 021D one. It was the inverse of transitive reduction where the reachability matrices were meant to be retained. The rationale here was to minimize the number of hops from one cluster to another, as there has been known transmission history between the two. In such a scenario, it was unnecessary for a mediating cluster to present, therefore a shortest path approach was

adopted. The implication on network structure was that out-stars, rather than directed lines, were selected.

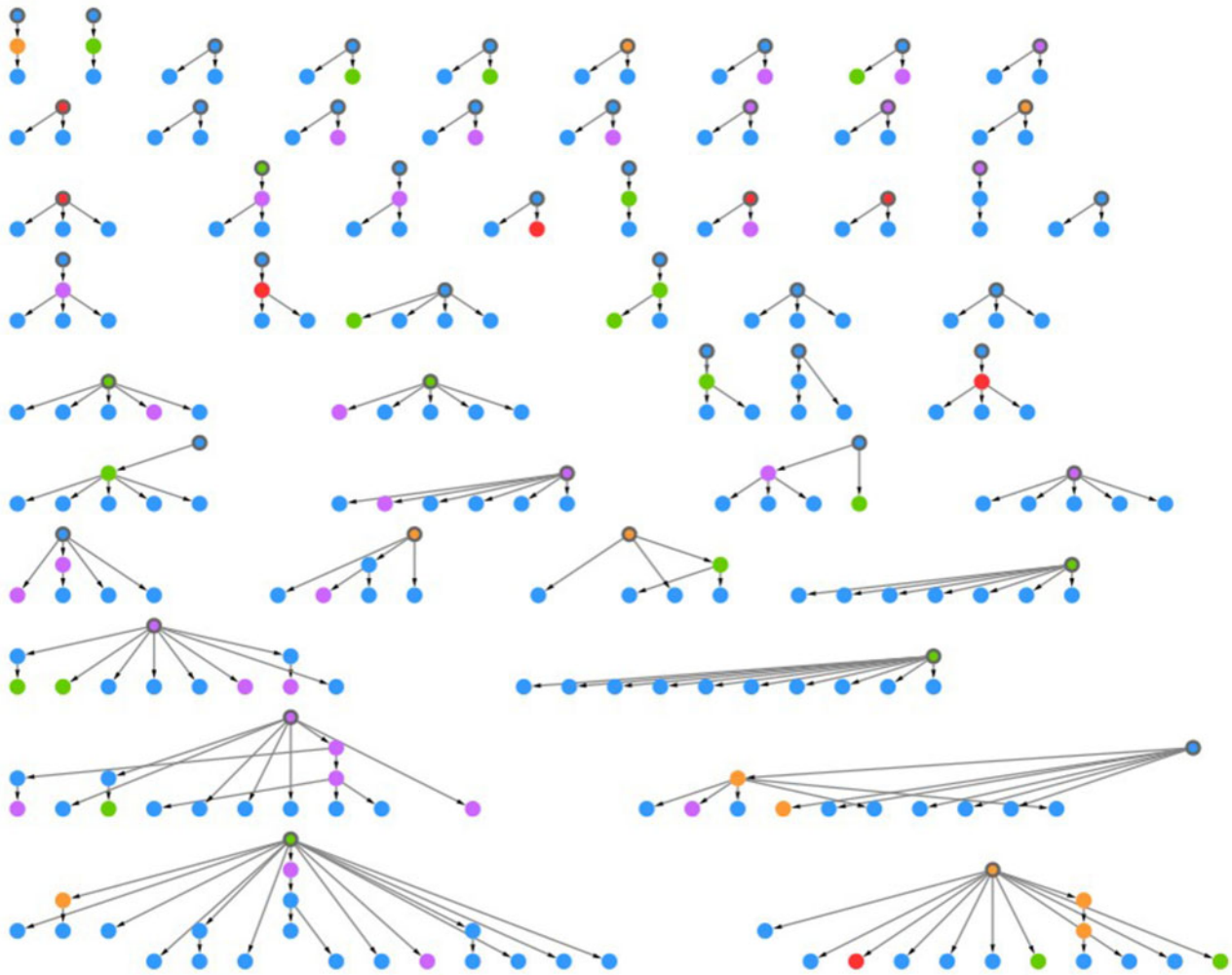
## Analysis

### Information diffusion metrics

Derived from graph theory and with reference to a previous study,<sup>16</sup> 3 information diffusion metrics were measured: scale, range, and speed. Scale refers to the number of persons one can transmit to, range refers to the number of generations can an infection be passed on, and speed refers to the average time for the infection to traverse from root to tip along the transmission path. In our study, we defined the 3 metrics per cascade as the maximum out-degree of a cluster within a cascade, the length of the longest directed path, and the number of edges divided by the difference between the first symptom onset date and the last reporting date. To describe the full extent of transmission of a cascade, we calculated its order, which is the number of nodes within the cascade. Triad census was performed to evaluate the distribution of out-stars (021D) and directed lines (021C). The metrics were compared among cascades containing and seeded by different exposure settings using Mann-Whitney  $U$  test. R was used to perform these analyses with *igraph* package. The network graph was visualised by Cytoscape 3.8.2.

### Exponential-family random graph modeling

To understand the roles of network structure and dyadic relationship between nodes in network formation, exponential-family random graph modeling (ERGM) was applied. As described previously, the network was founded by out-stars and directed lines. One of them (directed lines, also known as two-paths in the package) was selected in the model. To assess the relationship between exposure settings, mixing pattern of settings of clusters was incorporated at 2 levels in separate models: overall mixing, and individual setting mixing. As a positive log-odds could compute a higher probability, a negative value for two-paths meant it is less likely for a new edge to be created to form a two-path and vice versa. Positive and negative conditional log-odds with a significant  $P$  value of lower than .05 represented a homo-



**Figure 2.** Network of transmission cascades composing of at least 3 epidemiologically linked clusters during the second wave of the COVID-19 (coronavirus disease 2019) epidemic in Hong Kong. Cascades are laid out hierarchically. Length of edges is for illustration only. Nodes with bold borders are index clusters. Node color represents categories of exposure settings (blue: residence, purple: daily activities, red: healthcare, orange: social activities, green: work/studies).

geneous and heterogeneous mixing pattern, respectively. R package *ergm* was used for the construction of 2 models.

## RESULTS

During the second wave of COVID-19 outbreak between June 20 and October 23, 2020, in Hong Kong, a total of 138 connected components (cascades) consisting of 441 nodes (clusters) and 303 edges were identified. The number of clusters within each cascade ranged from 2 to 20, of which 87 were dyads consisting of 2 nodes. The remaining 51 cascades with a size of at least 3 were formed by 397 out-stars (021D) and 54 directed lines (021C) were included for further analyses (Figure 2). In term of index cluster's exposure setting, the most popular setting was residence ( $n = 28$ ), followed by 8 and 6 cascades seeded by daily activity and work/study clusters, respectively. Six cascades had their initiation identified in a social activity setting, while the remaining 4 were in healthcare settings.

All cascades contained residence clusters therefore no comparisons could be made to distinguish cascades without involving residence settings (Table 1). Cascades containing daily activity clusters had a higher transmission speed (median 0.25 vs 0.16; interquartile

range [IQR], 0.17-0.28 vs 0.12-0.25;  $P = .017$ ). Social activity cluster-containing cascades had a longer range of transmission (median 2.00 vs 1.00; IQR, 1.25-2.75 vs 1.00-2.00;  $P = .021$ ). Cascades involving work/study environment had a greater number of nodes (median 6.00 vs 3.00; IQR, 3.25-10.25 vs 3.00-4.00;  $P = .0031$ ), a longer range (median 2 vs 1; IQR, 1-2 vs 1-2;  $P = .026$ ), and a wider scale (median 3.50 vs 2.00; IQR, 2.00-7.00 vs 2.00-3.00;  $P = .026$ ). If the cascade was seeded by a daily activity cluster, the speed of propagation between clusters appeared to be faster (median 0.27 vs 0.18; IQR, 0.20-0.40 vs 0.13-0.25;  $P = .0096$ ), whereas those seeded by healthcare clusters transmitted more slowly (median 0.11 vs 0.19; IQR, 0.07-0.17 vs 0.13-0.25;  $P = .045$ ). Residence-initiated cascades had a lower scale of transmission (median 2 vs 3; IQR, 2-3 vs 2-7;  $P = .017$ ), but it was higher in work/study-seeded cascades (median 6.00 vs 2.00; IQR, 4.25-10.75 vs 2.00-3.00;  $P = .0047$ ). The latter also had a higher number of clusters within the cascade (median 7 vs 3; IQR, 5.50-13.25 vs 3.00-5.50;  $P = .0044$ ).

Results from an ERGM showed that, it was less likely for an edge to be added to form a directed line ( $P < .0001$ ) or to connect 2 clusters of the same setting category ( $P < .0001$ ) (Table 2). The probability of adding an edge forming an out-star with another category was 1.42%, while the probability of adding a directed line with



**Table 1.** Association between information diffusion metrics and exposure settings during the second wave of COVID-19 outbreak in Hong Kong

	Daily	Healthcare	Social	Residence	Work/study
Contained setting					
Order				<sup>a</sup>	.0031
Range			.021	<sup>a</sup>	.026
Scale				<sup>a</sup>	.026
Speed	.017			<sup>a</sup>	
Index setting					
Order					.0044
Range					
Scale				.017 <sup>b</sup>	.0047
Speed	.0096	.045 <sup>b</sup>			

Only significant *P* values of positive association were shown.

COVID-19: coronavirus disease 2019.

<sup>a</sup>All clusters contained residence settings; therefore, no comparisons could be made.

<sup>b</sup>Significant negative associations.

the same category and with a different category was 0.10% and 0.55%, respectively. To examine the mixing pattern of exposure categories, another ERGM was performed. A homogeneous mixing pattern among social activities clusters ( $P = .013$ ) and heterogeneous mixing patterns among residence ( $P < .0001$ ) and healthcare clusters ( $P < .0001$ ) was found. It was noted that daily activities ( $P = .74$ ) and work/study ( $P = .45$ ) clusters did not mix preferably with either the same or a different category of cluster.

## DISCUSSIONS

In our study, the actor focused in the transmission cascades is the cluster/event, rather than an individual in conventional network analyses. This approach highlighted the important roles of clusters and their associated exposure setting in the propagation of SARS-CoV-2 in the community. Epidemiologically, the severity of an outbreak is dependent on the sizes and extents of transmission clusters instead of the linear relationship between an infected person and their contact.<sup>17</sup> The main data fields for the network analyses in the present study was the history of gathering events, which would likely be more reliable than one's self report symptom onset or the later report or diagnosis date. The algorithm presented herein therefore defined direction of edges by the order of cluster co-occurrence

**Table 2.** Exponential random-graph models

	Conditional log-odds	Standard error	<i>P</i> value
Model 1			
Edge	-4.24	0.10	<.0001
Two-path	-0.97	0.11	<.0001
Same category	-1.70	0.17	<.0001
Model 2			
Edge	-4.24	0.10	<.0001
Two-path	-0.95	0.11	<.0001
Same category: daily	0.16	0.48	.74
Same category: healthcare	<sup>a</sup>	<sup>a</sup>	<.0001
Same category: residence	-1.90	0.19	<.0001
Same category: social	1.69	0.68	.013
Same category: work/study	-0.74	0.98	.45

<sup>a</sup>The small sample size caused negative infinity conditional log-odds.

instead of one's symptom onset date or case report date. One may also need to assert assumptions on the network structure, such as one-to-many relationships or a clique among all cases within the same cluster, which by design greatly prejudices the analysis result.

The directed networks constructed by this algorithm consist of out-stars and directed lines only, which is suitable for computing information diffusion metrics.<sup>16</sup> For the same number of nodes, an out-star would have a higher scale, while a directed line would have a higher range. The feature of different components could be easily contrasted in describing the epidemiological spread of SARS-CoV-2. As per the algorithm design, out-stars and directed lines were fundamental building blocks of the transmission network but out-stars were more prevalent in the second wave of COVID-19 outbreak in Hong Kong. It was observed that mixing pattern of categories was also crucial in the formation of transmission cascades. Notably, social activities setting clusters tended to connect with related clusters, whereas residence settings were more likely to be connected with a cluster of another exposure setting. It showed that people were attending more than 1 social activities during the period and were able to pass on the virus to another social activity after getting it from a social function.<sup>13</sup> On the other hand, transmission between residences was rare, which could be related to fomite transmission or through defective wastewater plumbing system<sup>18</sup> within a neighborhood or in the same building, but that was apparently not manifested in Hong Kong during the second wave, even in the hospital setting.<sup>19</sup>

The information diffusion metrics could be used in comparing the clustered transmission of SARS-CoV-2 between exposure settings. In our result, we have demonstrated that social and work/study setting-containing cascades had a higher range of transmission, while the latter one, particularly being the index cluster, also had a higher transmission scale and a higher order. Transmission occurring in work or study environments in which people from different households came together was evidenced to have connections with multiple clusters, which gave a higher measure of scale.<sup>20</sup> These subsequent clusters were primarily household transmissions with high secondary attack rates.<sup>21</sup> The engagement of family members in other activities could further propagate the virus through different indoor settings resulting in cascades with long range.<sup>21</sup> The product of wide scale and long range gave a high order. Among cascades having longer propagation history, social activities setting clusters were more likely to be involved. This signified the important role of social events in bridging transmission between clusters. Cascades involving daily activity clusters, particularly as the index cluster, had a higher transmission speed. This observation highlighted the high social connectivity of certain environments facilitating virus transmission within a short time, and therefore these settings should be specifically dealt with in controlling the spread of epidemic.<sup>22</sup> In practice, nonpharmaceutical interventions such as limiting the operating hours and maximum capacity of business are relevant policies to contain transmissions through daily activity clusters.<sup>23</sup> The low speed of healthcare setting-initiated cascade indicated these transmission was limited.<sup>24</sup> The transmission scales of residence showed that it did not play an important role in population spread.

The analyses in this study were founded on the syntheses of contact tracing data. Importantly, effective contact tracing relied on individual's voluntary disclosure of the location and time of visit prior to diagnosis. Noncooperation could lead to delay or failure to identify exposed individuals and events, and would affect any subsequent analyses. While the emergence of the COVID-19 pandemic has shown the importance of superspreader, their identification as

individuals could result in moral blame.<sup>25</sup> Despite the public health importance of the contact tracing mechanism, there was concern of stigma that individuals may withhold information to protect the confidentiality of one's identity. Similar limitations on recall bias and social desirability bias exist, particularly if someone has been to a place that they wanted to conceal from the public or family members. Although the network was primarily constructed from the sequence of cluster co-occurrence, in the presence of multiple clusters, the order of the generation of clusters would be taken into account. With a cluster-oriented perspective, the actual case load of each cluster was not taken into account when performing the analysis. Therefore, cluster size did not affect network construction. Missed contacts, however, could affect network structure if and only if they were the only ones who bridged 2 clusters. It could be represented as the edge weight in future implementation of this algorithm. Incorporating edge weights for measuring relationship between clusters could enrich the analysis. Likewise, the reachability matrix was not the same after graph reduction, performance of further path analysis should be cautioned as this might affect the result. Although ERGM could be useful for predicting future growth of network, we did not perform simulation, as incidentally the outbreak wave in Hong Kong had reached its end. Definition of exposure settings could vary among different cultures or societies. The method of classification could be adopted in other places but the actual categorization has to be tailored to the unique social norms and habits.

## CONCLUSION

The study results proved that the algorithm was useful in mining relationships between disease transmission clusters, identifying important clusters and assessing cluster relationships by applying it on the second wave of COVID-19 outbreak in Hong Kong. In controlling an epidemic, population intervention policies should rely on scientific evidence. Homogeneous mixing among social setting clusters highlighted control on social events has to come in place at a right time with a proportionate force. Transmission resulting in household clusters are inevitable, and therefore testing and quarantining close contact was useful and has to be continued before further transmission from the household. Daily and work or study environments posed significant transmission risk when people got together. Depending on local epidemic situation and cultural context, the metrics adapted in this study could be used as an objective scale for serving policy implementation purposes. By identifying setting-specific clusters of public health importance, targeted nonpharmaceutical interventions could be in place to strike the balance between outbreak control and daily lives.

## FUNDING

This work was supported by the Health and Medical Research Fund of Food and Health Bureau, Hong Kong Special Administrative Region Government under Grants COVID190105, INF-CUHK-1, and 19181132.

## AUTHOR CONTRIBUTIONS

All authors contributed to the conception of the study. THK designed the algorithm and implemented the study. NSW and E-KY contributed to data acquisition. All authors contributed to the interpretation of the data. The first draft of the manuscript was written by THK and SSL. All authors reviewed the manuscript and contributed to revisions.

## ACKNOWLEDGMENTS

Ms Carrie Yam, Ms Mandy Li, and Ms Sharon Chung are thanked for their assistance in data collection and collation. The authors thank Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong for providing technical support.

## COMPETING INTERESTS

The authors have no competing interests to declare.

## DATA AVAILABILITY STATEMENT

The data underlying this article were provided by the Department of Health by access data approval. Data will be shared on request to the corresponding author with permission of the Department and ethics committee.

## REFERENCES

- Cheng H, Jian S, Liu D, Ng T, Huang W, Lin H; Taiwan COVID-19 Outbreak Investigation Team. Contact tracing assessment of COVID-19 transmission dynamics in Taiwan and risk at different exposure periods before and after symptom onset. *JAMA Intern Med* 2020; 180 (9): 1156–63.
- Kwan TH, Wong NS, Lui GCY, *et al.* Incorporation of information diffusion model for enhancing analyses in HIV molecular surveillance. *Emerg Microbes Infect* 2020; 9 (1): 256–62.
- Wang W, Liu J, Tang T, *et al.* Attributed collaboration network embedding for academic relationship mining. *ACM Trans Web* 2020; 15 (1): 4.
- Atluri G, Karpatne A, Kumar V. Spatio-temporal data mining: a survey of problems and methods. *ACM Comput Surv* 2018; 51 (4): 1.
- Nasution MKM, Noah SAM, Saad S. Social network extraction: superficial method and information retrieval. In: Proceedings of International Conference on Informatics for Development; November 26, 2011; c2-110-5; Yogyakarta, Indonesia.
- Liu J, Xia F, Wang L, *et al.* Shifu2: A network representation learning based model for advisor-advisee relationship mining. *IEEE Trans Knowl Data Eng* 2021; 33 (4): 1763–77.
- Zhang L, Hu J, Xu Q, Li F, Rao G, Tao C. A semantic relationship mining method among disorders, genes, and drugs from different biomedical datasets. *BMC Med Inform Decis Mak* 2020; 20 (Suppl 4): 283.
- Kwan TH, Wong NS, Lee SS. Participation pattern of methadone users and its association with social connection and HIV status: analyses of electronic health records data. *PLoS One* 2019; 14 (5): e0216727.
- Nagarajan K, Muniyandi M, Palani B, Sellappan S. Social network analysis methods for exploring SARS-CoV-2 contact tracing data. *BMC Med Res Methodol* 2020; 20 (1): 233.
- Zheng M, Yu M, Cheng S, *et al.* Characteristics of HIV-1 molecular transmission networks and drug resistance among men who have sex with men in Tianjin, China (2014-2018). *Virol J* 2020; 17 (1): 169.
- Zhong L, Zhang Q, Li X. Modeling the intervention of HIV transmission across intertwined key populations. *Sci Rep* 2018; 8 (1): 2432.
- Block P, Hoffman M, Raabe IJ, *et al.* Social network-based distancing strategies to flatten the COVID-19 curve in a post-lockdown world. *Nat Hum Behav* 2020; 4 (6): 588–96.
- Adam DC, Wu P, Wong JY, *et al.* Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nat Med* 2020; 26 (11): 1714–9.
- Saraswathi S, Mukhopadhyay A, Shah H, Ranganath TS. Social network analysis of COVID-19 transmission in Karnataka. *Epidemiol Infect* 2020; 148: e230.
- Wong NS, Lee SS, Kwan TH, Yeoh EK. Settings of virus exposure and their implications in the propagation of transmission networks in a COVID-19 outbreak. *Lancet Regi Health West Pac* 2020; 4: 100052.
- Yang J, Counts S. Predicting the speed, scale, and range of information diffusion in Twitter. In: *Proceedings of the Fourth International Conference*

- on Weblogs and Social Media; May 23–26, 2010: 355–8; Washington, DC. <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1468/1896>. Accessed March 29, 2021.
17. Benincà E, Hagenaars T, Boender GJ, van de Kasstele J, van Boven M. Trade-off between local transmission and long-range dispersal drives infectious disease outbreak size in spatially structured populations. *PLoS Comput Biol* 2020; 16 (7): e1008009.
  18. Gormley M, Aspray TJ, Kelly DA. COVID-19: mitigating transmission via wastewater plumbing systems. *Lancet Glob Health* 2020; 8 (5): E643.
  19. Cheng VC, Wong SC, Chan VW, So SY, et al. Air and environmental sampling for SARS-CoV-2 around hospitalized patients with coronavirus disease 2019 (COVID-19). *Infect Control Hosp Epidemiol* 2020; 41 (11): 1258–65.
  20. Lan FY, Wei CF, Hsu YT, Christiani DC, Kales SN. Work-related COVID-19 transmission in six Asian countries/areas: a follow-up study. *PLoS One* 2020; 15 (5): e0233588.
  21. Leclerc QJ, Fuller NM, Knight LE, Funk S, Knight GM; CMMID COVID-19 Working Group. What settings have been linked to SARS-CoV-2 transmission clusters? *Wellcome Open Res* 2020; 5: 83.
  22. The Royal Society. SARS-CoV-2: Where do people acquire infection and ‘who infects whom’? 2020. <https://royalsociety.org/topics-policy/projects/set-c-science-in-emergencies-tasking-covid/>. Accessed March 29, 2021.
  23. Jarvis CI, Gimma A, van Zandvoort K, Wong KLM, Edmunds WJ; CMMID COVID-19 working group. The impact of local and national restrictions in response to COVID-19 on social contacts in England: a longitudinal natural experiment. *BMC Med* 2021; 19 (1): 52.
  24. Baker MA, Fiumara K, Rhee C, et al.; CDC Prevention Epicenters Program. Low risk of COVID-19 among patients exposed to infected health-care workers. *Clin Infect Dis* 2020 Aug 28 [Online ahead of print].
  25. Cave E. COVID-19 super-spreaders: definitional quandaries and implications. *Asian Bioeth Rev* 2020; 12 (2): 235–42.