Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
A knowledge graph-based method for epidemic contact tracing in public transportation

Tian Chen\textsuperscript{a,1}, Yimu Zhang\textsuperscript{b,1}, Xinwu Qian\textsuperscript{c}, Jian Li\textsuperscript{a,*}

\textsuperscript{a} Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, 4800 Cao’an Road, Shanghai 201804, China  
\textsuperscript{b} Urban Mobility Institute, Tongji University, 4800 Cao’an Road, Shanghai 201804, China  
\textsuperscript{c} The University of Alabama, Tuscaloosa, AL 35487, United States

ARTICLE INFO

Keywords:  
Digital Contact Tracing  
Contact Network  
Knowledge Graph  
Public Transportation  
Epidemic Control

ABSTRACT

Contact tracing is an effective measure by which to prevent further infections in public transportation systems. Considering the large number of people infected during the COVID-19 pandemic, digital contact tracing is expected to be quicker and more effective than traditional manual contact tracing, which is slow and labor-intensive. In this study, we introduce a knowledge graph-based framework for fusing multi-source data from public transportation systems to construct contact networks, design algorithms to model epidemic spread, and verify the validity of an effective digital contact tracing method. In particular, we take advantage of the trip chaining model to integrate multi-source public transportation data to construct a knowledge graph. A contact network is then extracted from the constructed knowledge graph, and a breadth-first search algorithm is developed to efficiently trace infected passengers in the contact network. The proposed framework and algorithms are validated by a case study using smart card transaction data from transit systems in Xiamen, China. We show that the knowledge graph provides an efficient framework for contact tracing with the reconstructed contact network, and the average positive tracing rate is over 96%.

1. Introduction

As of 31 May 2021, the novel coronavirus disease 2019 (COVID-19) has resulted in more than 170 million confirmed cases and 3.7 million deaths worldwide (WHO, 2021). Existing studies have revealed a positive correlation between human mobility and the number of COVID-19 infections (Jia et al., 2020; Kraemer et al., 2020; Xiong et al., 2020), and have further suggested that enforced travel restrictions can effectively slow the spread of the pandemic (Chinazzi et al., 2020; Kraemer et al., 2020). Among various travel modes, urban public transportation is found to play an important role in facilitating the spread of COVID-19 due to its essential mobility functionality in densely populated areas (Liu et al., 2020; Qian and Ukkusuri, 2021), and public transportation vehicles and stations served as key modes of virus transmission, given their relatively enclosed spaces and high passenger volume (Sun et al., 2013; Zhen et al., 2020). Therefore, many countries implemented various control measures over public transportation during the COVID-19 pandemic, including regular disinfection, mask enforcement, reduced capacity for social distancing, altered operation schedules,
and even complete service shutdowns (Zhen et al., 2020). All of these measures can be considered general approaches for operating public transportation systems during a pandemic. They represent intermediate strategies that often lag behind the pandemic outbreak and aim to reduce secondary infections at the aggregate level. However, aggregate control strategies are usually not as effective as target control strategies that focus on the timely spotting and isolating of infectious individuals. The recent advances in pervasive computing and the availability of large-scale trip data present urgent opportunities that allow for the deployment of more precise control at the individual level in public transportation systems. However, challenges remain in determining how accurate knowledge can be efficiently mined from large-scale unstructured travel data for disease mitigation purposes.

In the field of epidemiology, manual contact tracing has been used as a target control strategy in which confirmed infections are surveyed to determine the historical activities sequence of the infected individual (e.g., 14 days for COVID-19 (Shen et al., 2020)); however, manual tracing is inefficient when applied to large-scale investigations. Since the outbreak of COVID-19, the deployment of digital contact tracing has received broad attention from public health administrators and epidemiology researchers, as it is expected to produce high resolution information about contact between strangers (Anglemeyer et al., 2020) based on big data collection about people’s daily activity. In terms of available data access, mobile phone data may not be effective for defining meaningful close contact for several reasons: (1) GPS error may be too great (e.g., COVID-19 may spread effectively within 3 m (Sun and Zhai, 2020)); (2) a high penetration rate is required (Robert et al., 2020); and (3) activities may be unclear in some complex scenarios (e.g., underground spaces and certain movements), especially for transportation systems. In this regard, smart card data with fixed routes and operation schedules are considered helpful for capturing contacts and trace infections in public transportation systems. The spread of infectious disease through contact among members of the population naturally defines a contact network (Keeling et al., 2020), which is ideally suited to contact tracing, as it can represent the process of an epidemic spreading through potential dynamics rather than identifying a detailed etiology. Sun et al. (2013) used smart card data to construct a contact network, which was followed by later studies that proposed various contact network structures and modeled epidemic spread among passengers (Bota et al., 2017; Mo et al., 2021), and such contact networks can be transferred to private transportation and other social scenarios. However, using a relational database to represent large-scale smart card data as a contact network is redundant, because contacts are stored between pairs of passengers; therefore, such relational data cannot directly represent the actual network structure, and may result in poor performance in the execution of multiple recursive connections and queries for contact tracing.

Models and algorithms used to analyze public transportation data have been well developed in terms of operation, management, and decision-making under daily conditions, but they have not been introduced into contact networks to perform contact tracing under pandemic control scenarios. The knowledge graph is a technology that has been broadly considered in recent years and is distinct from the traditional relational database. A knowledge graph is generally considered a graphic network that stores data in the form of nodes and edges, where nodes represent entities or concepts, and edges represent relationships between nodes, and generally supports a network in scale of tens of billions of nodes and edges. In this way, a knowledge graph provides a consistent and compatible resource description framework that can intuitively represent anything in the real world and theoretically construct a semantic-rich network. The knowledge graph has been successfully applied in the field of Web Searching to solve semantic analysis problems. Despite the popularity of the knowledge graph, to the best of our knowledge, it has received little attention for organizing large-scale transportation data, and it has significant implications for the efficient construction of high-resolution contact networks. Therefore, we explore the application of the knowledge graph for use in digital contact tracing. We propose a knowledge graph-based framework that enables the general reuse of existing data, models, and algorithms to construct contact networks, thus supporting epidemic spread modeling, digital contact tracing, and other epidemiological studies.

In this paper, we propose a knowledge graph-based method to reconstruct a semantic-rich contact network for public transportation systems, supporting the effective modeling of epidemic spread and enabling efficient digital contact tracing. First, by utilizing a trip chaining model, we integrate multi-source data collected from smart cards, automatic vehicle location (AVL) devices, shift records, and route sheets of buses, bus rapid transit (BRT), and metro systems. We combine “top-down” and “bottom-up” methods to construct a public transportation knowledge graph, which can be considered a semantic-rich network characterizing the city-wide spatial–temporal movements of passengers. Second, a targeted simplified contact network is extracted from the constructed knowledge graph. We use an infectious risk prediction model based on individual contact features to simulate epidemic spread in the contact network, and we propose a breadth-first algorithm for digital contact tracing, which aims to effectively locate secondarily infected individuals in the transportation system based on already detected cases. Finally, the proposed framework and algorithms are verified by a case study of Xiamen, China, and the effectiveness and real-world implications are discussed. We highlight the main contributions of our study as follows:

- We propose a knowledge graphed-based public transport trip-chain model, which is more computational efficiency and has good scalability compared with the method based on relational database.
- We propose a lightweight time-continuous contact network structure, in which node denotes either a passenger or a vehicle, and an edge denotes the act of riding. Such a network structure reduces data redundancy, and is conducive to being extended.
- We propose a breadth-first search algorithm to enable efficient and effective digital contact tracing in our proposed large-scale contact network.
The rest of this paper is organized as follows. Section 2 reviews recent research efforts that are relevant to the context of this study. Section 3 provides a brief description of public transportation multi-source data and a classical trip chaining model. Section 4 describes our modeling methods and algorithms. Section 5 presents the experimental results. Section 6 includes the discussion, and section 7 concludes this paper.

2. Related work

2.1. Digital contact tracing

The outbreak of COVID-19 has raised concerns regarding digital contact tracing technology, and the verification of the effectiveness and practicability of such technology in a real-word pandemic is a persistent challenge. Many studies have designed functions for portable devices and apps to obtain personal activity data and protect privacy (Scassa, 2021; Shubina et al., 2020). Yasaka et al. (2020) proposed a location-hiding smart phone app to build an anonymous interpersonal graph for contact tracing, while similar studies were conducted that recorded proximity events of mobile phones using encrypted IDs (Abeler et al., 2020; Ferretti et al., 2020). The essential point of these studies is in line with the representation of graphic networks. Most applications of digital contact tracing focus on identifying close contacts through detected cases, and then isolating them to reduce further transmission (Abeler et al., 2020; Keeling et al., 2020; Robert et al., 2020). Asabere et al. (2020) applied betweenness centralities and tie strengths in a social network to recommend probable infectors. Keeling et al. (2020) predicted that approximately 36 persons must be traced per individual infected case to reduce further infections by more than the 5/6. Kucharski et al. (2020) simulated different scenarios of medical testing, isolation, tracing, and physical distancing, and suggested a combination of isolation and tracing strategies. Further, to identify infected persons and transmission routes in a population, some mathematical algorithms for networks have been adopted for digital contact tracing based on small medical datasets (Spada et al., 2004). However, the sufficient uptake and usage of devices or apps in a large population, which is fundamentally required to guarantee the effectiveness of digital contact tracing (Anglemyer et al., 2020; Braithwaite et al., 2020), are a major practical challenge.

We compare some studies on contact tracing with our algorithm in Table 1. It can be seen that this paper achieves certain results that are meaningful for contact tracing: (1) we construct a large-scale contact network in public transportation systems to verify digital contact tracing; (2) we conduct data fusion, network construction, epidemic spread, and contact tracing in an efficient framework based on a knowledge graph; (3) we strengthen the contact tracing and obtain excellent results that find more than 96% of infected persons.

2.2. Contact network

Due to the advantage of the potential representation of transmission routes, the contact network among populations is generally applied to model epidemic spread and to evaluate intervention methods (Keeling et al., 2020; Li et al., 2021). Abundant public transportation data has been widely used for contact network construction. Sun et al. (2013) found that repeated encounters of individuals in a public transportation network will result in a large-scale contact network, an undirected graph in which the nodes denote

| Studies               | Data source                          | Data scale                      | Method                                                                 | Effectiveness                                                                 |
|-----------------------|--------------------------------------|---------------------------------|------------------------------------------------------------------------|-------------------------------------------------------------------------------|
| Abeler et al. (2020)  | An app uses bluetooth technology      | Not mentioned                   | Doctor report the app ID of confirmed case, and other contacted apps will be informed | Not mentioned                                                                |
| Ferretti et al. (2020)| A mobile phone app to record proximity events between individuals | Not mentioned                   | Isolate symptomatic individuals; trace the contacts of symptomatic cases to quarantine them | Reduce R0 to less than 1                                                     |
| Keeling et al. (2020) | A survey asked participants to report social encounter features with other persons | More than 50,000 encounters from over 5800 respondents | All reported contacts of 15 min or more as close contacts               | Less than 1 in 6 cases will generate subsequent infections, averagely 36 individuals traced per case |
| Kojaku et al. (2021)  | Body contact information estimated by Bluetooth signal strength | More than 700 students in a university | “Backward” tracing: tracing close contacts who will be infected         | Contact tracing lowers the peak of infections by ~50%                          |
| Kucharski et al. (2020)| BBC Pandemic data                    | 40,162 participants            | Estimate the reduction in transmission under different control measures | Reduce 29%-66% further transmission; 20,000 new cases require over 500,000 contacts to be quarantined |
| Robert et al. (2020)  | A smart phone app collecting contacts of the past 7 days | Simulated on an urban population of 1 million individuals | From cases self-reporting of symptoms to trace their contacts and inform self-isolation | Epidemic can be suppressed with 80% of smartphone users or 56% of the population use the app |
| Yasaka et al. (2020)  | Smartphone app to construct an anonymized graph of interpersonal interactions | Not mentioned                   | Individuals will self-isolate if contact infected person                | Simulate the infection curves with different app adoption rate with a period |
| This study            | Multi-source public transportation data | 14.8 million edges and 1.76 million nodes | Contact tracing in a contact network to find out all infected persons   | Find more than 96% infected persons; need to test 35–41 % of passengers      |
individuals and the edges denote contact, and Sun’s work has been followed by later studies that model epidemic spread (Bota et al., 2017; Keeling et al., 2020). Mo et al. (2021) proposed a lightweight theoretical framework with a low computational cost to describe a time-varying contact network. Transportation models and other networks can also be introduced to construct contact networks. Liu et al. (2020) developed a method of matching passengers with trains to construct a contact network of metro travelers. Qian et al. (2020) developed a generational model to construct metro contact networks and analyzed the associations between the network structures and the risk level of infectious diseases. Qian et al. (2020) introduced an observation network into a contact network of metro travelers to capture the interactions between epidemic spread and information transmission. However, such contact networks are rarely combined with contact tracing algorithms in the existing studies.

2.3. Knowledge graph construction

A knowledge graph derives from the challenge of extracting semantic information from Web text with a machine. Its purpose is to provide a general data fusion framework by mapping data into entities and relationships, thus achieving powerful data representation and search performance by means of its graphic network. In 2012, Google announced its Knowledge Graph project to take advantage of semantic knowledge for the enhancement of search systems and the improvement of query result relevance. Due to its advantages in the representation of semantic knowledge, the enhancement of graphic searches, and the improvement of knowledge reasoning, the knowledge graph has been widely applied in various fields, such as machine translation (Moussallem et al., 2018), question-answering systems (Sun et al., 2018), natural language inference (Annervaz et al., 2018), and recommendation systems (Cao et al., 2019). A knowledge graph can be divided into a general knowledge graph and a domain knowledge graph (Yu et al., 2020), according to a broad division of commonsense knowledge and domain knowledge; the latter generally requires more comprehensive background knowledge and specific datasets than the former (Lin et al., 2021). Among the domain-dependent applications of a knowledge graph, the top two are the health domain and the mobility domain (Elvira et al., 2019). Studies in the health domain mainly focus on providing intelligent healthcare solutions but have rarely involved epidemiology (Elvira et al., 2019), while studies in the mobility domain mainly focus on urban transportation, e.g., the Km4City ontology adopted by the Sii-Mobility project (Bellini and Nesi, 2018), which supports a series of research studies on different urban traffic issues ((Badii et al., 2019; Badii et al., 2018; Bellini et al., 2014)). However, in the specific domain of public transportation, some studies have applied the functional components of the knowledge graph independently, such as knowledge reasoning (Santofimia et al., 2017) and knowledge representation (Liang et al., 2018), but an efficient framework that integrates data and models of public transportation for pandemic contact tracing using knowledge graphs has not yet been developed.

3. Preliminaries

3.1. Data

The multi-source public transportation data used in this study covers smart card data, AVL data, shift records, and the route schedules of bus transit; BRT and metro systems include only smart card data and route schedules. Our study area, Xiamen City in China, has over 467,000 daily passengers that use smart card transactions, contributing to an average of 1.293 million daily trip records as of July 2020. The smart card can be universally used across the bus, BRT, and metro systems. For busses, the automatic fare collection devices are mounted at the front doors, and the smart card transaction data can only capture the boarding records. The automatic fare collection equipment of the BRT and metro is available at both station entrances and exits; hence each transaction record has complete entrance and exit information. Nevertheless, the records are not explicitly associated with a particular vehicle or

| Data type            | Attribute              | Sample                  |
|----------------------|------------------------|-------------------------|
| Smart card data      | Card ID                | 8012013032724xxx        |
|                      | Time                   | 2018-07-15 15:37:57     |
|                      | Plate number (for bus only) | 089xx                |
| AVL data             | Plate number           | 089xx                   |
|                      | Time                   | 2018-07-15 15:38:10     |
|                      | Longitude              | 118.17xxxx              |
|                      | Latitude               | 24.50xxxx               |
| Shift record         | Plate number           | 089xx                   |
|                      | Departure time         | 2018-07-15 14:50:02     |
|                      | Arrival time           | 2018-07-15 16:41:23     |
|                      | Operation line         | 46                      |
|                      | Direction              | Up                      |
| Route schedule       | Line name              | 46                      |
|                      | Station name           | Jinshan station         |
|                      | Station serial number  | 21                      |
|                      | Direction              | Up                      |
train. AVL data records the spatial–temporal sampling nodes of bus vehicles. Shift records are operational schedules that arrange the arrival and departure of bus vehicles. Route schedules summarize the transit lines, direction, and station information for every month (Table 2).

3.2. Trip chaining model

The smart card data of busses lacks alighting information, so the trip chaining model aims to infer the alighting stations and alighting times of bus trips. Unlike previous studies in which alighting information was inferred solely from bus trip records, we are able to resolve this issue by leveraging the trip chaining information with the availability of comprehensive city-wide transit data. Specifically, the trip chains of passengers reflect continuous travel trajectories, thus there are spatial–temporal correlations between consecutive trips. In this paper, a trip chain is defined as the sequence of trip segments (trips completed by a single public transportation mode) that was made by a passenger to travel from the origination to the destination in the public transportation system. Four assumptions were adopted to facilitate the construction of trip chains based on multi-source data, which were proposed and verified in previous studies. Barry et al. (2002) proposed a method by which to estimate travelers’ stops in the New York subway system based on “next trip” and “last trip” assumptions. Trépanier et al. (2007) adopted the assumptions of Barry et al. (2002), further incorporated the “similar trip” assumption to reveal the likelihood of trips taken the next day, and also observed weekly travel patterns to complete missing bus boarding information. Munizaga and Palma (2012) proposed a method by which to estimate the alighting time and location by adding the “return trip” assumption and generalized time constraints in a multi-modal public transport system to estimate the alighting points. Munizaga and Palma (2012) reported an 80% success rate, while Trépanier et al. (2007) achieved a 66% success rate for bus trips.

1. “Next Trip”: the alighting station of a trip segment is assumed to be close to the boarding station of the next trip segment.
2. “Last Trip”: the alighting station of the last trip is assumed to be close to the boarding station of the first trip taken on the same day.
3. “Return trip”: if a passenger takes the same transit line on two consecutive trips, and the line direction (up and down) is opposite, then the alighting station of the first trip is the boarding station of the second trip, and the boarding station of the first trip is the alighting station of the second trip.
4. “Similar trip”: if a passenger only produces one single trip in a day, the alighting station of the trip may be similar to trips that start from the same boarding station.

Following these assumptions, we illustrate the concept of trip chaining using the example shown in Fig. 1. In the example, a passenger has three trips on one day, and the first trip transfers to the second trip. The alighting station of the first trip is estimated by the boarding station of the second trip according to nearest distance rule, and the alighting station of the third trip is estimated by the boarding station of the first trip. We use multi-source data collected from the bus, BRT, and metro systems, which covers the main transit modes within an urban area, to improve the accuracy of estimation results.

4. Methods

Based on the public transportation multi-source data and the trip chaining model, this paper constructs a public transportation knowledge graph, extracts a contact network to model the epidemic spread process, and conducts efficient digital contact tracing.

4.1. Public transportation knowledge graph construction

A knowledge graph consists of two parts: a schema layer and a data layer (Xu et al., 2016). The schema layer generally adopts an ontology, a formal specification of an agreed upon conceptualization of a domain in the context of knowledge description (Gruber,
1993)—for example: author, writes, book. The data layer generally adopts a graph database to store data in triple formats—for example: Shakespeare, writes, Hamlet; this can be considered as an instantiation of an ontology. Therefore, there are three ways to construct a knowledge graph: (1) top-down: first defining an ontology as a schema layer, then mapping data into entities and relationships and importing them into a graph database in the form of triples as a data layer; (2) bottom-up: first extracting entities and relationships from the origin data and importing them into a graph database as a data layer, then summarizing, condensing, abstracting, and reasoning on the data layer to construct an ontology; (3) combining the two: defining an ontology to guard data mapping and reasoning. This study combines the first two methods and adopts a trip chaining model to construct a public transportation knowledge graph, as shown in Fig. 2.

The ontology of a public transportation knowledge graph contains 5 types of entities and 10 types of relationships, and these nodes and associations have certain property characteristics. Table 3 exhibits some instantiations of the ontology.

4.2. Knowledge representation and reasoning

Because a knowledge graph represents semantics, it can also be considered an information network composed of nodes and edges, represented as a directed graph \( G = (V, E) \), where \( V \) is the set of nodes in the information network, and \( E \) is the set of edges between the nodes. If there is more than one node type and edge type in a contact network constructed from the knowledge graph, then the contact network is a heterogeneous information network. Given a heterogeneous information network, the association path between two nodes \( V_1 \) and \( V_n \) is denoted in the form of \( V_1 \xrightarrow{E_1} V_2 \xrightarrow{E_2} \cdots \xrightarrow{E_{n-1}} V_n \), which defines a composite relationship \( E_c = E_1 \circ E_2 \circ \cdots \circ E_{n-1} \) between nodes \( V_1 \) and \( V_n \), where \( \circ \) denotes a composition operator. Such path definition can also be implemented in a relational database, which is called a recursive search. An infectious disease can be regarded as a kind of information that flows along the edges to activate the infection status of the connected nodes.

An association path contains information on how two nodes are connected and how they interact. Compared with the data stored in a relational database in the form of structured tables with columns and rows, more comprehensive relationships (i.e., paths) can be mined by network representation methods, which can extract association features and generate new relationships between entities. This method is called knowledge reasoning in the context of a knowledge graph, which can be realized simply by adding new edges to the nodes if their relationships conform to self-defined rules or other principles.

Considering the public transportation knowledge graph, several representations are illustrated to describe basic relationships, which are essential to building a trip chain model and analyzing the contact characteristics. The representations include the travel order relationship, common ride and common wait relationships, and direct and indirect contact relationships. Among them, the travel order is the identification of passenger travel divided according to the trip chain and sorted according to travel stages, and the travel of public transport passengers can build directional relationships, including interchange based on the travel order. This leads to the knowledge graph representation of passenger behavior based on the passenger trip chain, which is the key to the passenger travel knowledge graph network. Co-riding and co-waiting are the definitions of the possible contact scenarios of transit passengers, which
are conducive to the realization of contact relationship judgment. Direct and indirect contact are contact behaviors of public transport passengers in the bus environment where epidemic transmission may occur, and contact behaviors are the core concepts of public transportation contact networks. Fig. 3 presents the representation of these relationships.

(1) The order of trips

Considering a travel plan of a passenger including three trips in one day, in which two successive trips alongside a transfer are taken before the third trip, the entity of trip and the relationship of the transfer are shown as below:

\[ T_p = V_{n=1}^{1} E_{T=1} V_{n=2}^{2} E_{T=0} V_{n=3}^{3} \]

where \( T_p \) denotes the chained trips of passenger \( p \) in one day, \( V_{n=trip}^{n} \) denotes the \( n \)th trip, \( E_T \) denotes next trip, and \( (E_T = 1) \) denotes a transfer between two trips.

We use a composite relationship to describe the three kinds of trips: the first trip, the last trip, and the trip in which the transfer ends.
on that day, as represented below:
\[
T^\text{trav}_p = V^1_{\text{trip}}, \text{ as } \exists V^1_{\text{trip}} \rightarrow V_{\text{vehicle}} \text{ and } \not\exists V_{\text{vehicle}} \rightarrow V^1_{\text{trip}}
\]
\[
T^\text{tran}_p = V^3_{\text{trip}}, \text{ as } \exists V^3_{\text{trip}} \rightarrow V^1_{\text{trip}} \text{ and } \not\exists V^1_{\text{trip}} \rightarrow V^3_{\text{trip}}
\]
where \( \exists \) denotes “exist”, and \( \not\exists \) denotes “does not exist”.

For the transfer trips at the two endpoints of a successive trip chain (i.e., continuous transfers), the composite relationship is represented by the context, as follows:
\[
V^1_{\text{trip}} \not\exists \text{transfer} \rightarrow V_{\text{vehicle}} \rightarrow V^1_{\text{trip}} \not\exists \text{transfer} \rightarrow V^1_{\text{trip}} \not\exists \text{transfer} \rightarrow V^1_{\text{trip}}
\]
where \((\text{transfer} = 1)^* n (n \geq 1)\) denotes \(n\) successive transfers.

(2) Co-riding and co-waiting

The co-riding and co-waiting in this study specifically refer to public transportation scenarios, such as vehicles and stations. Consider the contact between two passengers if they take a same bus or wait at the same station:
\[
V^1_{\text{trip}} \rightarrow E^1_{\text{vehicle}} \rightarrow V^1_{\text{trip}} \rightarrow E^1_{\text{vehicle}} \rightarrow V^1_{\text{trip}}
\]
\[
V^1_{\text{trip}} \rightarrow E^1_{\text{vehicle}} \rightarrow V^1_{\text{vehicle}} \rightarrow E^1_{\text{vehicle}} \rightarrow V^1_{\text{trip}}
\]
where \(V_{\text{trip}}\) denotes the passenger \(p\), \(E_{\text{vehicle}}\) denotes having a trip, \(V^p_{\text{trip}}\) denotes a particular trip of passenger \(p\), \(E_\text{boarding}\) denotes an action boarding at a station, and \(V_{\text{vehicle}}\) denotes the scenario of a vehicle while \(V_{\text{station}}\) denotes the same station.

A direct co-riding relationship is defined if two passengers took the same bus vehicle:
\[
V^1_{\text{trip}} \leftrightarrow E^1_{\text{vehicle}} V^2_{\text{trip}} \text{ if } |V^1_{\text{trip}}(\text{boarding time}) + V^2_{\text{trip}}(\text{boarding time})| - |T| \leq T_{\text{threshold}}
\]
where \(E_{\text{co-riding}}\) denotes co-riding, and \(V^i_{\text{trip}}(\text{boarding time})\) denotes the boarding time of the trip.

A co-waiting relationship is defined if two passengers boarded at the same station within a short time interval \(T_{\text{threshold}}\) (e.g., 15 min), although the accuracy needs to be checked:
\[
V^1_{\text{trip}} \leftrightarrow E^1_{\text{vehicle}} V^2_{\text{trip}} \text{ if } |V^1_{\text{trip}}(\text{boarding time}) - V^2_{\text{trip}}(\text{boarding time})| \leq T_{\text{threshold}}
\]

Where \(E_{\text{co-waiting}}\) denotes co-waiting.

(3) Direct contact and indirect contact

Considering the travel connections of three passengers, we have:
\[
V^1_{\text{trip}} \leftrightarrow E^1_{\text{vehicle}} V^2_{\text{trip}} \leftrightarrow E^1_{\text{vehicle}} V^2_{\text{trip}} \leftrightarrow E^1_{\text{vehicle}} V^2_{\text{trip}} \leftrightarrow E^1_{\text{vehicle}} V^2_{\text{trip}} \leftrightarrow E^1_{\text{vehicle}} V^2_{\text{trip}}
\]
where \(E_\text{t} \in \{E_{\text{DC}}, E_{\text{CW}}\}, V^p_{\text{trip}}\) denotes the \(n\)th trip of passenger \(p\), and the new relationship between \(V_{\text{trip}}\) and \(V_{\text{vehicle}}\) is defined as direct contact, while the new relationship between \(V_{\text{trip}}\) and \(V_{\text{vehicle}}\) is defined as indirect contact:
\[
V^1_{\text{trip}} \leftrightarrow E^1_{\text{vehicle}} V^2_{\text{trip}} \leftrightarrow E^1_{\text{vehicle}} V^2_{\text{trip}} \leftrightarrow E^1_{\text{vehicle}} V^2_{\text{trip}}
\]

\(E_{\text{DC}}\) denotes direct contact, and \(E_{\text{IC}}\) denotes indirect contact.

4.3. Digital contact tracing algorithm based on knowledge graph

In this section, we propose a digital contact tracing algorithm based on the knowledge graph. To investigate contact tracing, previous studies have focused mostly on identifying the close contacts of detected cases cooperating with quarantine or self-isolation measures, thus reducing further infections. However, the tracing of persons infected earlier receives less attention, and therefore undetected cases may still continue to transmit the virus, and the pandemic control process may be persistent and recurrent. Considering the highly infectious nature of COVID-19, an ideal and reliable approach is to identify all infected persons in the network within a short period of time to prevent recurrent outbreaks and control the pandemic. Like inferring the whole picture from a local view, we strengthen digital contact tracing to identify the remaining infected persons (including those already infected and those with secondary infections) in a contact network based on detected cases (i.e., index cases). Because the virus is transmitted through certain
kinds of routes and produces secondary infections, we hope to reverse the process by tracing back possible routes to find the source of infections based on index cases.

We can study the task of contact tracing in a simplified scenario: a contact network in which a node denotes either an individual or a place, and an edge links only one individual with one place to represent a kind of activity. The edge contains properties of entrance times and exit times, thus representing the interactions and duration of individuals in a place on the timeline. The virus is likely to be transmitted to a healthy person when he or she is close to an infected person in a place for a period of time. Such transmission happens in different places until some infected persons are detected. However, due to the limitations of medical resources, it may be not possible to test everyone in the network. Based on these index cases, digital contact tracing is expected to generate a reference population containing the rest of the undetected infectors, thus assisting administrative or medical decision-making that is implemented to control the pandemic. Based on the constructed public transportation knowledge graph, we first extract a simplified subgraph (i.e., the contact network) with only two entities denoting the passenger and the bus and one relationship denoting the action of taking a bus, since a lightweight network structure is more conducive to simulating epidemic spread and implementing a contact

![Diagram](image)

**Fig. 4.** Extracting a sub-graph from public transportation knowledge graph for contact tracing algorithm.
tracing algorithm. The ontology of the extracted contact network is shown in Fig. 4. By comparison, the traditional contact network structure is an undirected graph, in which the node denotes the individual and the edge denotes the potential contact intensity.

We propose a breadth-first algorithm for digital contact tracing, the main idea of which is to trace back all possible transmission routes of detected cases to add directed infection relationships until traversing them. Our infection dataset comes from the simulation results of the epidemic spread. Initially, all infected persons are labeled “infected”, but only a sub-set is considered to be index cases that are labeled “index”. We give different marks to passenger nodes that have not been labeled “index”: “checked”, “found”, and “unchecked”. A “checked” mark is given to a node if it is searched by our algorithm, representing the potential of infection, as it is a close contact or potential transmission medium of the index cases. If an “index” node is marked “checked”, then it is “found”. Other nodes are marked “unchecked”. The algorithm starts from “index” nodes to search potential routes of increasing lengths to other “index” nodes until all “index” nodes are marked “found” and no new potential vectors are “checked”, meaning that we find all possible infection routes for index cases; therefore, undetected cases are probably in the medium nodes and connected nodes of the routes. The “infected” nodes marked “found” are used to evaluate the effectiveness of the algorithm.

Table 4 shows the concrete steps of the proposed algorithm. Step 0: set the initial path length as 2, which is the minimum length needed to link two passengers. Search all paths between “index” nodes where the path length is 2, and search “index” nodes that are marked “checked” as well as “found”, which means we find all possible transmission routes with the length of 2 between these nodes.

Step 1: add 2 to the path length; if all “index” nodes are “found”, go to step 4, or else go to step 2 and step 3 to continue searching paths.

Step 2: search all paths between “checked” nodes and “index” nodes that meet our path length, and the medium nodes should be “unchecked”; then, mark the medium nodes as “checked”, and update the infected dates of the medium nodes. Step 3: if there are new paths searched in step 2, then return to step 2, or else go back to step 1. Step 4: search all paths between “unchecked” nodes and “checked” nodes with the length of 2, and then mark them as “checked”. The “checked” nodes searched in step 2 are considered primary contacts, as they are linked to no less than two nodes and might be potential transmission media, while the “checked” nodes searched in step 4 are regarded as close contacts, as they are linked to only one possible infected node. Generally, the number of close contacts is far greater than the number of primary contacts.

We mainly focus on the detection rate of infected persons (i.e., those who have tested positive). The validity of the digital contact tracing algorithm can be evaluated by using the false positive rate (FPR, rate of negative cases predicted to be positive) and the true positive rate (TPR, the rate of positive cases predicted to be positive) and the true positive rate (TPR, the rate of positive cases predicted to be positive). Further, the primary positive rate (PPR) represents the detected positive rate using primary contact cases, and the maximum TPR(MTPR) represents the most ideal detection rate using all primary contacts and close contacts.

\[
\text{FPR} = \frac{FP}{(FP + TN)} \quad (1)
\]

\[
\text{TPR} = \frac{TP}{(TP + FN)} \quad (2)
\]

\[
\text{PPR} = \frac{(IP + TP)}{C} \quad (3)
\]

| Algorithm description. |
|------------------------|
| **Input:** the sub-graph of public transportation knowledge graph and index nodes |
| **Output:** checked nodes and found nodes |

**Step 0: initialization**
- set path_length = 2
- search: path1 = V_{node}(index) \rightarrow E, V_{node}(index) where length(path1) = path_length
- mark nodes in path1 as “checked” and “found”;

**Step 1: loop judgment**
- set path_length = path_length + 2;
- if “index” nodes are all marked “found”:
  - go to step 4;
- else:
  - go to step 2;

**Step 2: path searching**
- search: path2 = V_{node}(index) \rightarrow E, V_{node}(index) where length(path2) = path_length, and medium nodes are “unchecked”;
- mark the medium nodes in path2 as “checked” (if “index” then “found”);
- update “infected date” of the medium nodes;

**Step 3: iterative judgment**
- if count(path2) > 0:
  - back to step 2;
- else:
  - back to step 1;

**Step 4: further tracing method**
- search: path3 = V_{node}(checked) \rightarrow E, V_{node}(checked) where length(path3) = 2

**End**
\[ MTPR = \frac{MTP}{MTP + FN} \]  

where FP denotes false positive cases (negative but predicted to be positive), TN denotes true negative cases (negative but predicted as negative), FN denotes false negative cases (positive but predicted to be negative), and TP denotes true positive cases (positive but predicted to be positive). These indexes are used for primary contacts, while MTP is specified for maximum true positive cases, including primary contacts and close contacts. IP is the number of index cases that are detected as positive. C is the total number of persons in the network.

According to the epidemic control requirements, it is more important to identify TP cases than to check FP cases, so TPR and MTPR are the most important indicators.

4.4. Benefits of transportation knowledge graph for contact tracing

To represent a graph, the relational database retrieves the connectivity between pairs of nodes by validating the pairwise adjacency of two pieces of records in the database, and a recursive query with a longer path length (more than 1) requires a multi-table join. However, a table join is very expensive in a relational database due to the large number of I/O operations and memory consumption that results in low performance of the data queries. We first illustrate the benefits of the transportation knowledge graph in terms of the time complexity for contact network construction in a relational database and a native graph database (the database adopted by the knowledge graph). Consider the case in which we need to track the contacts of passenger A based on the smart card transaction data stored in a relational database versus a graph database following the transportation knowledge graph. In the relational database, the time complexity may be \( O(N) \) (\( N \) is the number of rows) and be iterative over all records. In addition, it may be proportional to \( O(\log(N)) \) but may require the sorting of the records based on the departure time; however, the sorting itself is an expensive operation in a large relational database. As such, the total time complexity for constructing a contact network from the transaction records stored in the relational database is \( O(N^2) \) or at least \( O(N \log N) \) plus the sorting cost for each passenger. On the other hand, the cost of contact tracing in the graph database is proportional to the number of vehicles \( V \) that is adjacent to the passenger, which can directly map to all other passengers (denoted by \( K_v, \forall v \in V \)) that are in the same vehicle based on the knowledge graph. In this regard, the cost for constructing the complete contact network would only require \( O(NV_kv) \) with \( VK_v \ll N \) in any real-world transportation systems. This indicates a significant performance gap for large-scale applications.

Alternatively, as one of the key objectives of our study, the benefits of the transportation knowledge graph can also be understood from our contact tracing algorithm with index cases. One notable difference is that our contact tracing algorithm does not require the explicit knowledge of a complete contact network and can be directly performed on the constructed database. On the contrary, applications with a relational database need to explicitly build the contact network first and then track the contacts starting from the index cases. Denoting \( S \) as the number of index cases, the complexity for contact tracking in the contact network will be \( O(SN) \).

Fig. 5. BRT and metro networks of Xiamen.
Following our transportation knowledge graph, the time complexity will simply be \( O(SVk) \). This will again result in a performance gap in terms of tracing susceptible contacts in addition to the cost of constructing the contact network. As such, we conclude that the transportation knowledge graph will provide a significant benefit for analyzing the contact dynamics in transportation systems and for contact tracing in applications associated with detecting infectious diseases.

5. Numerical experiments

The proposed method was applied based on empirical public transportation data collected in Xiamen, China. Due to the difficulty of obtaining empirical pandemic investigation data, an infection risk prediction model was used to obtain the underlying data. Two sets of comparison experiments were designed to demonstrate the advantages of the proposed method. The test platform system was Windows 10, the processor environment was an Intel® Xeon® Gold 5118 processor, the CPU frequency was 2.30–2.29 GHz, and the running memory was 64.0 GB.

The first set of experiments was designed to illustrate the advantages of the proposed knowledge graph-based method by comparing it with the relational database-based contact tracing method. The public transportation knowledge graph-based model was constructed using the construction method proposed in Section 4, and the relational model was constructed based on a fused data table containing passenger card numbers, travel vehicle numbers, travel dates, infection dates (initial value of 0), and infection days (initial value of 0). The effectiveness of the two methods was then compared.

The second set of experiments was designed to illustrate the robustness of the proposed knowledge graph-based contact tracking considering the change of constraints. Due to possible recall bias and individual differences in viral pathology, two new cases with varying constraints were supplemented for the second set of experiments. One case was given a relaxed restriction on viral latency, and the other was given a moderate error-tolerant interval for medical data.

5.1. Experiments setting

5.1.1. Knowledge graph construction

This study takes Xiamen, China as an example. By July 2018, there were 340 bus lines, 3 BRT lines, and 1 Metro line in Xiamen. The data is collected as smart card data, AVL data, shift records, and route sheets from the bus, BRT, and metro systems from July 1 to 14, 2018. The BRT and metro networks are shown in Fig. 5, and the bus lines are mainly distributed on the island. The three transit modes cover most of the daily travels of the residents of Xiamen, hence avoiding the insufficient coverage of passenger groups and improving the modeling accuracy. Considering the acceptable walking distance, travel flexibility, and preferential policies of transfers, a threshold of 30 min is set to judge the transfers.

We construct our public transportation knowledge graph in a popular knowledge graph platform called Neo4j. After the ontology construction, data mapping, knowledge fusion, and knowledge reasoning, an estimation accuracy of 89% for public transport alighting transactions was achieved using data from multiple public transport modes, including buses, BRT, and subways. Moreover, the effectiveness of the estimation results can be confirmed by travel survey data, as reported in the authors’ previous research (Zhang et al., 2021). The data loss mainly comes from single bus trips, which may be related to the fact that Xiamen city is a tourist city, and tourists usually take random bus trips. Due to the lack of necessary shift departure data for the BRT and metro systems, these passengers cannot be matched to BRT vehicles and trains. Therefore, we extract a contact network with passenger nodes and vehicle nodes, and one relationship denotes a trip based only on bus data, including 14.8 million trips, 1.3 million passengers, and 0.46 million bus shifts.

5.1.2. Infection simulation data

Due to the lack of actual large-scale pandemic investigation data, an infection risk prediction model was applied to simulate the epidemic on an individual level in the contact network to obtain the underlying data. Similar epidemic spread simulations have been analyzed via classical epidemiological models, such as mathematical models and statistical models (Li et al., 2021). The Wells-Riley model is one of the most classical models used to predict the infection risk in a certain situation based on the contact of individuals. Referring to the research of Sun and Zhai (2020) and Zhang et al. (2020), the Wells-Riley model can be modified to predict the infection risk in public transportation scenarios:

\[
P_I = \frac{C}{S} = 1 - \exp\left(-\frac{\beta I q pt}{\alpha Q}\right)
\]

where \( P_I \) is the probability of infection (infection risk), \( C \) is the number of infected persons, \( S \) is the number of susceptible persons, \( I \) is the initial number of infected persons, \( p \) is the pulmonary ventilation rate of one person (m³/s), \( q \) is the quantum generation rate produced by one infected person (quantum/s), \( t \) is the exposure time (s), and \( Q \) is the ventilation rate (m³/s). Moreover, \( \beta \) is the load rate of a transit vehicle, which is defined as the ratio of passengers to the rated passenger capacity (which ranges from 0 to 1), and implies the change of social distance in the vehicle. Additionally, \( \theta \) is the coefficient of protection provided by wearing masks, and is set as 1 if no passengers wear masks (if no protection is provided). Finally, \( \alpha \) is the ventilation efficiency, the value of which ranges from 0.8 to 1.

According to recent research of COVID-19 transmission in bus vehicles (Sun and Zhai, 2020), the parameters in our modified Wells-Riley model are specified as follows: \( q = 0.238 \) (quantum/s), \( p = 0.3 \) (m³/h) when people sit or conduct light activities in vehicles, \( t \) is
calculated by the co-riding duration of the infected person, and \( Q = 5000 \text{(m}^3/\text{s}) \) according to the bus design standards. Due to the fixed trips, the epidemic spread will not affect passenger travel the study period, which may be similar to the initial outbreak when few passengers were aware of the infectious disease and had not taken prevention measures. We set \( \theta = 1 \) and \( \alpha = 0.8 \). The incubation period of COVID-19 is set as 3 days, which means a passenger will be infectious three days after infected.

We adopt three assumptions to simplify the model: (i) the contact infection only occurs in public transportation vehicles, (ii) the infection will not lead to changes in passenger travel behavior, and (iii) there is no individual difference in the transmission characteristics of the virus, and its infectivity remains constant. It needs to be emphasized that this study does not aim to accurately simulate or predict the transmission of COVID-19, and the parameters used here need further discussion in epidemiological applications. Simplified assumptions are helpful to run the algorithm intuitively. After calculating the risk level for passengers co-riding with infected persons, their status change (from healthy to infected) is updated randomly corresponding to the infection risk. We set a 1% random probability for passengers to be infected who take the same vehicles with infected persons to simulate the matching error of the trip chaining model.

According to the modified model, the actual infection risk of taking the bus is roughly 1–6%, which is close to the 2.5–4.2% reported by the survey data (Sun and Zhai, 2020). Based on the bus network in our constructed knowledge graph, random samples are selected as initial infected passengers in different fractions on July 1, 2018, which are 0.0003% (one passenger), 0.01%, 0.05% (171 passengers), and 0.1%. The transmission will spread in the next 11 days after the three days incubation period.

As shown in Fig. 6, the weekends are marked by a sky-blue background, and the new infected cases are counted by increments of 0.5 h 48 times every day. The new infected cases fluctuate during bus operation times and show two peaks in line with the commuting peaks in the morning and evening, while decreasing on weekends as compared with the previous weekdays. According to the Wells-Riley model, the infection risk has a positive correlation with the co-riding duration and load factor; hence the number of new infections should synchronously change with the passenger flow. This proves the sensitivity and effectiveness of our model for capturing time-varying contacts in network.

5.2. Results of digital contact tracing

The simulation results of the 0.05% fraction were used for subsequent digital contact tracing, and a total of 9,288 passengers were found to be infected by 171 passengers after 14 days. Although digital contact tracing is expected to perform better than manual contact tracing, manual epidemiological investigations are essential, especially for index cases, to obtain infection dates and other features. In performing infected person tracking for both models separately, it is believed that digital tracking is characterized by the same reliable medical data and strict time constraints. The reliability of medical data mainly refers to the accuracy of the date on which a patient was infected, and the time constraint is mainly focused on the infection interval, which should correspond to the incubation period of the virus.

Based on the index case search process, it was found that as the sampling fraction increased, more transmission routes were solved with path length = 2. Almost all index cases were “found” after the search when path length = 4, and only very few index cases required a further search with path length = 6. This indicates that when a relatively quick and reliable search for infected persons is required, setting the maximum path length = 4 may be a sufficient compromise, as more than 99% of index cases can be “found.”

We found that the graph model-based digital contact tracking algorithm exhibited substantial superiority in terms of time efficiency. Fig. 7 reveals that the graphical model ran for less than 1000 s per sample fraction, while relational model cost 700% more computational time in completing the same task. The overall running efficiency of the digital tracking algorithm based on the graphical model was nine times higher than that based on the relational model. This means that the algorithm based on the graphical database is
able to respond faster in the detection and tracking of infected persons, thereby making it more likely that management will have more adequate time to contemplate decisions.

The evaluation of the proposed algorithm is presented in Fig. 8. Fig. 8(a–c) reflect on the primary contacts searched in step 2 of the algorithm, while Fig. 8(d) reflects on the primary contacts and close contacts searched by all the steps of the algorithm. As primary contacts indicate a greater potential for infection (associated with at least two cases of infection), the results can be referenced for the prioritization of resource allocation in the case of limited hospital capacity.

According to the results of each evaluation indicator, the average TPR level was about 47% when the sampling score was 0.1 in Fig. 8(a), which means that 47% of undetected infections could be detected. As the sampling fraction increases, up to 80% of undetected infections can be detected at a sampling fraction of 0.9.

The FPR in Fig. 8(b) implies the potential waste of medical resources. With an average FPR of 96% across all sampled cases, this implies that only 1 infection can be detected for every 25 people tested. However, it should be noted that the proposed digital contact tracing algorithm uses only simple medical data (i.e., infection dates) to set time constraints. With more adequate investigation data,

Fig. 7. Digital contact tracking operational efficiency assessment.

Fig. 8. Digital contact tracing evaluation under reliable medical data and strict time constraints.
and with the addition of epidemiological criteria and constraints into the contact network and the algorithm, the PPR can be improved.

The PPR indicates the total infection detection rate in the entire population, including infections detected as index cases and positive cases found by the algorithm. As shown in Fig. 8(c), the PPR increased approximately linearly and eventually reached 98%, when the sampling fraction increased. This means that the index cases gradually constituted the highest proportion of all detected infection cases, and the utility of the algorithm gradually decreased. To identify as many infected persons as possible in the contact network with limited index cases, the search scope of suspected infected people must be expanded to close contacts, as explained in step 4 of the algorithm.

Fig. 8(d) demonstrates that the average MTPR reached more than 94%, and reached a maximum of 96% with a sampling fraction of over 0.1, by searching both primary and close contacts; this means that almost all the infected persons in the contact network were traced. And in general, almost all infected individuals were found to be tracked in the contact network, although the results were affected by the random selection of samples.

The red line in Fig. 8(a–d) represents the variation in the results of the 10 experiments due to random sampling. When the sampling fraction was small, the TPR and PPR produced relatively large fluctuations in the results, but the value was still only 7%, and the fluctuations produced by the MTTPR were the next-largest, namely 2%. With the increase of the sampling fraction, the variation in the results became insignificant. There is substantial variability in sample selection when the sampling fraction is small, and the algorithm search results are more influenced by the frequency of transit trips for each sample, e.g., if there happens to be a very active transit user in the indexed sample, more infection paths can be found through that user.

In addition, in the expression of the algorithm, the digital contact tracing algorithm based on the graphical model can also be realized via concise algorithm statements and timely expansion and update. On the one hand, the graphical model is networked, and there is a natural correlation between nodes; this makes it possible to search and contact passengers directly from candidate passenger nodes along the connection relation of intermediate nodes, which is very difficult to achieve via a relational model. This is because languages that require complex associations of multiple “joins” are difficult to write, and traversal computations are expensive. On the other hand, each node and relationship in the graph database has disposable descriptive attributes; thus, the passenger infection situation reflected by the attribute characteristics can be conveniently updated at any time, which is conducive to the expansion and update of the digital contact network. The state of the relational model is complex because the data structure must be redesigned, and the complexity of the algorithm is further increased if external constraints are considered.

To further illustrate the advantages of the proposed method, we compared the performance of the knowledge graph-based method with the backward tracing method, which was recently proposed by Kojaku et al. (2021). This method aims to trace the subsequently infected person according to the identified infected cases and performs backward tracing based on the contact probability in the contact network. Compared with the proposed method, backward tracking is more like a breadth-first brute force search with less consideration of the infection paths.

We compared the two methods based on the same experimental contact network with the number of the confirmed population and the same sampled seed infected individuals. The contact tracing in subsequent bus travel activities was performed based on the identified infected individuals. Fig. 9 shows the comparison results between the two methods. In general, the performance of the
The proposed method is better than the backward tracing method. The backward tracing method is only better than the proposed method in the FPR results at high sampling rates. The reason is that the backward tracing method does not consider the infection pathway based on index cases compared with the proposed method. When the sample size is relatively large (e.g., a sampling proportion of 70% or higher), as demonstrated by MTPR, most of the infection cases can be obtained through the backward tracing method by performing close contact tracing only, resulting in a higher TPR. In reality, however, we may not be able to know a considerable proportion of infected persons. In this case, we believe that the knowledge graph-based method is more practical and performs better than the backward tracing method.

5.3. Effects of longer infection processes on digital contact tracing

This section extended the length of the disease outbreak to 17 days. The total number of infected people was 11,879, 53,541 and 86,231 after 15, 16, and 17 days, respectively. The average length of the infection transmission chains was 2.94, 3.42, and 3.46 after 15, 16, and 17 days, respectively. Fig. 10 shows that there were more than 40,000 new infections in a day when the simulated infection process lasted up to 17 days without any intervention, and the overall cumulative increase in infection cases showed an exponential trend.

10 sets of random seeds were used to carry out the experiments for different transmission processes. Fig. 11 shows the results of the computation time, including the time spent for each sampling ratio in Fig. 11(a) and the total time spent for all samples in Fig. 11(b). For the 14 days tracing scenario, it took 1.2 h to complete the sampling ratio of 0.1–0.9, and it took about 1.4 h for 15 days, 4 h for 16 days, and 7 h for 17 days. The experimental results show that the proposed tracing algorithm searches by the path, and more infections and longer infection paths will result in a higher computation time. Total time consumption is found to resemble an exponential growth in the number of total infections.

From the perspective of algorithm effectiveness indicators in Fig. 12, our algorithm also produces a good performance in identifying and tracing infection cases under a longer infection process. When searching with a small sample, the TPR is found to even increase by 200% with the FPR decreased. This is due to the fact that FPR is equal to 1 minus the difference between infected cases divided by exposed cases, when the infection process grows, the increased risk of positive cases and vehicle exposure to infection makes the exposure pathway to infection increase, so the FPR value decreases.

5.4. Robustness of digital contact tracing

Due to the individual differences in the pathological characteristics of the virus and recall bias, the medical data obtained by interviewing the patients may not be reliable. Considering the modest fault tolerance of time constraints based on the investigated medical data, the proposed digital contact tracing algorithm was applied to three cases:

- Case 1: reliable medical data, strict time constraints;
- Case 2: reliable medical data, slightly relaxed time constraints;
- Case 3: not very reliable medical data, relaxed time constraints.

The reliability of medical data mainly refers to the accuracy of the date on which a patient was infected, and an unreliable date indicates that the patient may have been infected a few days before or a few days after that date. The time constraint is mainly focused on the infection interval that should correspond to the incubation period of the virus. The infection dates in case 1 with reliable medical data were set as absolutely accurate, and the incubation period met the requirement of 3 days. Case 2 had the same reliable medical data as case 1; however, because the incubation period of the infectious disease may vary due to individual differences (He et al., 2020), the incubation period was shortened to 2 days considering the worst-case scenario. The infection date of the index in case 3 was set to have a 1-day deviation due to possible memory bias in the manual survey. Different time constraints denote different fault
Fig. 11. Computation time of knowledge graph-based method with different lengths infection process.

Fig. 12. Results of knowledge graph-based method with different lengths infection process.

Fig. 13. The process to mark index cases in the digital contact tracing algorithm. (a), (b), (c) respectively represents case 1, case 2, case 3, and (d) represents the proportion of index cases marked “found” when path length = 2.
tolerances of the searching conditions, i.e., the parameter $E_c$ in the digital contact tracing algorithm. A strict time constraint reflects high confidence in the medical data and epidemiological model, while a relaxed time constraint may cover more possible transmission routes.

The time consumed by the algorithm to complete a sampling fraction from 0.1 to 0.9 in cases 1 to 3 ranged from 1.5 to 3.8 h, and the stricter the constraint, the more quickly the algorithm converged. Moreover, the shortest consumption of one sampling fraction was 100 s, while the longest consumption was 1500 s, and more than 99% of the time was consumed in the steps with path length $= 4$. Fig. 13(a–c) demonstrate that the compromise solution with path length $= 4$ “found” more than 99% of the indexed cases for different scenarios. The proportions exhibited in Fig. 13(d) indicate that as the time constraint was gradually relaxed in the three cases, more index cases were “found” when path length $= 2$. The excess parts in cases 2 and 3 as compared with case 1 represent incorrectly identified transmission routes, as the infection relationships searched with path length $= 2$ in case 1 were closer to the actual situation.

Fig. 14 exhibits the rates to evaluate the effectiveness of the proposed algorithm in the case of the changed constraints. Fig. 14(a) reveals that the best TPR performance was achieved in case 2, followed by that in case 3. When the sampling score was 0.1, the TPR reached a high level of more than 60%. As shown in Fig. 14(b), the FPR was higher in cases 2 and 3 than in case 1 because more healthy individuals were identified as suspected cases of infection when the medical data and time constraints were slightly relaxed. In contrast, the FPR for case 3 was slightly higher than that for case 2 because the relaxed medical data constraint in case 3 identified more false routes due to more incompatible model data. Again, due to the relaxed time constraints and the greater number of identified routes, the PPR was better for cases 2 and 3 than for case 1.

Fig. 14(d) demonstrates that a high MTPR with a sampling fraction of over 96% was achieved in all cases by searching both primary and close contacts. The MTPR cannot reach 100% because most infected persons will eventually infect another person. Excluding hospital capacity and detection costs, digital contact tracing may be effective; however, the actual conditions should be comprehensively considered to balance its effectiveness.

Fig. 15 presents the primary and close contacts found by digital contact tracing. Fig. 15(a) can be reasonably explained by referring to Fig. 14(b). In case 3, more cases of infection were found, as the search yielded more false paths. In addition, as the sampling fraction increased, the number of primary contacts obtained by the search gradually increased because the number of possible infection paths obtained by the search increased. As shown in Fig. 15(b), the levels of primary and close contacts in the sampling fractions of 0.1 and 0.9 differed considerably less. Referring to Fig. 14(d), this may be because the MTPR remained more consistently high from sampling fractions 0.1 to 0.9 in case 1 to case 3, and there was little difference in the number of primary and close contacts obtained from the search. Fig. 15(b) reveals that case 2 yielded more contacts at smaller sampling fractions. This is because, at smaller sampling fractions, the sample distribution was more dispersed and benefited from the relaxed date information in case 2, and more contact nodes were likely to be yielded.

There were 1.3 million passenger nodes in the contact network in 14 days, and the ratio of passengers to be tested was approximately 38% when the sampling fraction was 0.1. In other words, to find most of the 9,288 infected persons, only 35%-41% of the passengers may need to be tested. Although this is still a large number that significantly challenges hospital capacity, the result is encouraging as compared with the prospect of testing the entire population.

6. Discussion

This paper uses the knowledge graph to realize the digital contact tracing in a public transportation network. A knowledge graph uses a graph database to store data in the form of triples, which is provided as a generic way to map and integrate structured data, semi-structured data, and unstructured data, to thus construct a semantic-rich network. A knowledge graph provides a more efficient tool for data storage and network analysis, and existing models and algorithms can be reused. Therefore, a knowledge graph has the potential to be integrated into a normalized data management system. It can be used as a powerful tool for data integration and queries in daily operation, and it can also provide customized network structures to support epidemic investigation and contact tracing in emergencies.

We have to point out that it is an information-limited experiment to measure the effectiveness of the digital contact tracing algorithm under such simple time constraints, but we can still find more than 96% of infected persons by testing 35%-41% of passengers. We use the experiment to test the upper bound of the digital contact tracing, and the results are still encouraging. Moreover, we can consider applying detailed constraints for more refined searching. The contacts in our algorithm are represented by taking the same vehicle should not be infected, and passengers who get on after the infected person alights the vehicle should be at a lower risk of infection. In this way, the number of primary contacts and close contacts can be further reduced, as can the time consumption. Due to the error of the trip chaining model and other possible factors, the sensitivity of the algorithm should be comprehensively researched when adding new constraints. A more detailed understanding of virus characteristics and activity data will also be of great use.

However, we use a binary mark to identify the possible close contacts, which cannot provide more information about the infection risk of different groups. Under the condition of limited hospital capacity, especially when the results include a large number of healthy persons, it is not productive to take the necessary graded measures according to the degree of contact. In our infection risk prediction model, each person is calculated with an infection probability according to the degree of contact. Similar to this idea, we can apply fuzzy theory to the process of digital contact tracing; the time constraint provides the search and filter functions, and it also can be used to calculate the flexible infection indexes, which can be defined as (0, 1) to represent the confidence of infection for the identified contacts, thus strengthening the fault tolerance. The calculation and use of the confidence indexes should be carried out under the guidance of professional epidemiological knowledge. Combined with the confidence and sensitivity, we can suggest close contacts with lower risk to adopt measures such as self-isolation, so as to reduce the pressure on medical resources.
From the perspective of practical applications, some parts of our experiment need to be improved. The proposed algorithm is currently verified for use only in buses. However, the infection situation can be potentially extended to metro and BRT systems if the correlation data between passengers and carriages/vehicles are available. Moreover, other data in addition to transaction card data (e.g., public transportation e-travel passes) should be integrated to achieve higher contagion tracing accuracy. Moreover, the effectiveness and robustness of the proposed algorithm when applied to an actual comprehensive social network must be verified. Other routes of transmission (e.g., infection via contaminated surfaces) except contact transmission are not considered. The Wells-Riley model does not consider the distribution of passenger locations in the vehicle. Some parameters of passengers should be differentiated, and the dynamic changes in virus infectivity are not considered. Generally, the sampling fraction of the index cases may be small, because it may be obtained by personal reports, medical diagnoses, or community testing, considering limited hospital capacity. When the number of detected cases is too small, it may be better to use traditional manual contact tracing than digital contact tracing, and the selection criteria require more discussion based on the contact network structure. Due to the lack of behavior changes and control measures as the epidemic spreads, the experiment may be more similar to the early stage of the outbreak. The application limitations of digital contact tracing should be comprehensively understood, because they are closely related to pandemic situations and restraint factors, including hospital capacity, administrative executive ability, social acceptance, resource allocation, and localization (Anglemyer et al., 2020).

Fig. 14. Evaluation of digital contact tracing in case1 to case3.

Fig. 15. Primary contacts and close contacts of digital contact tracing.
7. Conclusions

In this study, we propose a knowledge graph-based framework to integrate public transportation multi-source data for contact network construction and realize effective digital contact tracing on a large-scale contact network. Compared with previous studies, this paper aims to trace back the contact relationships that are potentially related to infections in the contact network, hence to actively discover infected persons and keep the pandemic from spreading. As a potential high-risk place of virus transmission, public transportation systems are selected as the research background. We construct a public transportation knowledge graph by applying a trip chaining model and extract a simplified contact network consisting of passengers and vehicles with millions of nodes. Then, a modified infection risk prediction model is used to simulate epidemic spread using contact features on the individual level. Based on the simulation results, we evaluate the proposed digital contact tracing algorithm in the contact network to verify the effectiveness. The MTPR index shows that the proposed contact tracing algorithm can find more than 96% of infected persons based on small index samples.

In this study, we use two-weeks data for simulation and contact tracing. Experiments over a larger time span are helpful for studying the dynamic changes and time complexity of virus transmission. Moreover, the effects of pandemic prevention and control policies (e.g., travel restrictions) on the efficiency of the algorithm during the ongoing phase of the pandemic are worthy of further exploration. In addition, it is the belief of the authors that the knowledge graph-based model has significant potential applications, such as in spatial–temporal correlation analysis, traveler behavior portraits, etc. Further, the proposed knowledge graph can be combined with other knowledge graphs (e.g., built environment, political strategy, or geographic attributes) to investigate the influence factors on pandemic spread. For instance, the proposed knowledge graph can be related to an additional knowledge graph on built-environment to mine connections between travel sequences in transportation system with the underlying land use types. The tracing algorithms can then be implemented to reveal how sequence of land use types are associated with the transmission of the pandemic and identify major sequences of activity locations for future prevention purposes. Continued efforts will be devoted to the expansion and mining of the knowledge graph-based transport model. The public transportation vehicles and the contact between passengers can be regarded as a representative social scene and social activity, respectively. With the expansion of digital devices and apps, we hope to obtain detailed data to express complex scenes and activities, and to verify the effectiveness of digital tracking in the real world.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is support by the National Key R&D Program of China (Grant No. 2018YFB1601100). The authors also thank the anonymous reviewers’ constructive comments and suggestions.

Author contributions

The authors confirm contribution to the paper as follows: study conception and design: J. Li; data collection: T. Chen, Y. Zhang, J. Li; analysis and interpretation of results: T. Chen, Y. Zhang, X. Qian; draft manuscript preparation: T. Chen, Y. Zhang, manuscript revision: Y. Zhang, J. Li, X. Qian. All authors reviewed the results and approved the final version of the manuscript.

References

Abeler, J., Baecker, M., Buermeyer, U., Zillesen, H., 2020. COVID-19 Contact Tracing and Data Protection Can Go Together. Jmir Mhealth Uhealth 8 (4). https://doi.org/10.2196/preprints.19359.
Anglemeyer, A., Moore, T.H.M., Parker, L., Chambers, T., Grady, A., Chiu, K., Parry, M., Wilczynska, M., Flemyng, E., Bero, L., 2020. Digital contact tracing technologies in epidemics: a rapid review. Cochrane Database System. Rev. 8 https://doi.org/10.1002/14651858.cd013699.
Annervaz, K.M., Chowdhury, S., Dukkipati, A., 2018. Learning beyond datasets: Knowledge Graph Augmented Neural Networks for Natural language Processing. Proc. NAACL-HLT 1, 313–322. https://www.aclweb.org/anthology/N18-1029/.
Asabere, N.Y., Acakpovi, A., Ofori, E.K., Torgby, W., Kuuboore, M., Lawson, G., Adjaloko, E., Pinheiro, P.R., 2020. SARPPIC: Exploiting COVID-19 Contact Tracing Recommendation through Social Awareness. Comput. Math. Methods Med. 2020, 1–14. https://doi.org/10.1155/2020/3460130.
Badli, C., Bellini, P., Difino, A., Nesi, P., 2019. Sii-Mobility: An IoT/IoE Architecture to Enhance Smart City Mobility and Transportation Services. Sensors 19 (1), 1. https://doi.org/10.3390/s19010001.
Badli, C., Nesi, P., Paoli, I., 2018. Predicting Available Parking Slots on Critical and Regular Services by Exploiting a Range of Open Data. IEEE Access 6, 44059–44071. https://doi.org/10.1109/ACCESS.2018.2864157.
Barry, J.J., Newhouzer, R., Rahbee, A., Sayeda, S., 2002. Origin and destination estimation in New York City with automated fare system data. Transp. Res. Rec.: J. Transp. Res. Board 1817 (1), 183–187. https://doi.org/10.3141/1817-24.
Bellini, P., Nesi, P., 2018. Performance assessment of RDF graph databases for smart city services. J. Visual Lang. Comput. 45, 24–38. https://doi.org/10.1016/j.jvlc.2018.03.002.
Bellini, P., Nesi, P., Venturi, A., 2014. Linked open graph: Browsing multiple SPARQL entry points to build your own LOD views. J. Visual Lang. Comput. 25 (6), 703–716. https://doi.org/10.1016/j.jvlc.2014.10.003.
Bota, A., Gardner, L.M., Khani, A., 2017. Identifying Critical Components of a Public Transit System for Outbreak Control. Netw. Spatial Econ. 17 (4), 1137–1159. https://doi.org/10.1007/s11067-017-9361-2.
Zhang, Y., Chen, T., Wang, J., Li, Z., Li, J., 2021. Multi-Mode Public Transportation Travel Chain Construction Based on Multi-Source Data. Urban Transp. China 19 (4), 120–126. https://doi.org/10.13813/j.cn11-5141/u.2021.0024.

Zhang, Y., Wang, X., Bi, Q., 2020. Travel-Infected Susceptibility Based on Transmission Mechanism of COVID-19. Transp. Res. 6 (01), 73–80. https://doi.org/10.19721/j.cnki.1001-7372.2020.11.004.

Zhen, J., Chan, C., Schoonees, A., Apatu, E., Thabane, L., Young, T., 2020. Transmission of respiratory viruses when using public ground transport: A rapid review to inform public health recommendations during the COVID-19 pandemic. Samj S. Afr. Med. J. 110 (6), 478–483. https://doi.org/10.7196/SAMJ.2020.v110i6.14751.