**A Novel Graph Indexing Approach for Uncovering Potential**

**COVID-19 Transmission Clusters**

XULIANG ZHU, Hong Kong Baptist University, China XIN HUANG, Hong Kong Baptist University, China LONGXU SUN, Hong Kong Baptist University, China JIMING LIU, Hong Kong Baptist University, China

The COVID-19 pandemic has caused the society lockdowns and a large number of deaths in many countries. Potential transmission cluster discovery is to find all suspected users with infections, which is greatly needed to fast discover virus transmission chains so as to prevent an outbreak of COVID-19 as early as possible. In this paper, we study the problem of potential transmission cluster discovery based on the spatio-temporal logs. Given a query of patient user

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2 | 14 | **Query:** (Amy, 'May 17 00:00') | | | | | Xuliang Zhu, Xin Huang, Longxu Sun, and Jiming Liu | | | |
| Gq, tq | | | |
| ID | | User | Location | Time |
| 1 | Amy | | Starbucks | May 5 18:00 | May 5  18:05 | | May 9  15:00 | |
| 2 | Bob | | Starbucks | May 5 18:05 |
| Bob | | | Ella |
| days | 3 | Cora | | Starbucks | May 5 18:10 |
| May 16 | | | |
| 4 | Bob | | Ikea | May 9 15:00 |
| 10:10 | | | |
| 5 | Ella | | Ikea | May 9 15:00 | Amy | | | |
| May 5 | | | |
| 6 | Amy | | Starbucks | May 16 10:00 |
| 7 | Bob | | Starbucks | May 16 10:10 | 18:10 | Cora | | |
| 8 | Lily | | Starbucks | May 20 14:00 |

Fig. 1. An example of potential COVID-19 transmission cluster discovery on spatial-temporal logs. The query consists of a patient user

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A Novel Graph Indexing Approach for Uncovering Potential COVID-19 Transmission Clusters | | | | | | | | | | 3 |
| **A query record** | | ID | Data | ID | | | Data | | |
| **(q, tq)** | |
| 1 | | | U1, L1, T1 | | |
| Add (q, tq) into Q | | 2 | | | U2, L2, T2 | | |
| 1 | U1, L1, T1 |
| 2 | | | U2, L2, T2 | .. | | | | ....... | |
| .. | | | ....... | Spatial-temporal Database D | | | | | |
| **Suspect records Q** | | Spatial-temporal Database D | | Construct | | | | | |
| Add Newq | New query (q', tq') Q | | Construct | **A query record (q, tq)** | Find BCG partitions | | | | BCG-Index |
| Return all records close contact to q' | | R-tree Index | Bi w.r.t. (q, tq) | B1 | | |
| B2 | | | |
| into Q |

B3

|  |  |  |
| --- | --- | --- |
| **New suspect** | R-tree Index | BCG-Index |
| **records Newq** |
| Scan BCG-Index  once |
| Newq = ø |

**Output: potential**   
**transmission cluster Cq**

(a) Classical online search framework

**Output: potential**   
**transmission cluster Cq**

(b) Our graph indexing framework

Fig. 2. The frameworks of potential transmission cluster discovery in spatial-temporal databases.

However, an efficient extraction of query-dependent potential transmission clusters is challenging. The reason has two-fold. First, given millions of people in a city and multiple spatial-temporal records visited by one person, it may incur combinatorial blow-ups for enumerating all possible transmission clusters. Second, the infectious diseases usually have a time window of incubation period, reflecting that the virus transmission already happen before the patient is identified. Thus, the discovery of direct close contact is not enough for potential transmission clusters, which needs the search of all users involved in the high-risk clusters by underlying transmission.

To tackle the problem efficiently, we consider two different approaches of classical online search and our graph indexing based search. First, we consider the classical online search framework as shown in Figure 2(a), which finds all potential infecting users for every suspected record in a spatial-temporal database. Specifically, it first adds a query record (



4 Xuliang Zhu, Xin Huang, Longxu Sun, and Jiming Liu

relationships between users and spatio-temporal records. Through compressing and partitioning a bipartite graph, we propose a compact BCG-Index to optimize the index construction and potential cluster queries in a fast way. Different with online search method, it only needs to search in the partitioned graph index once instead of scanning the suspected spatial-temporal logs multiple times (at the step 1○ and 2○ in Figure 2(b)). Overall, our proposed techniques have wide applications on disease expansion control and early outbreak prevention. More importantly, our potential cluster discovery algorithms not only work for COVID-19 but also benefit for other kinds of close-contact based infectious diseases.

In summary, this paper makes the following contributions:  
• We motivate and formulate the problem of potential cluster discovery to prevent virus transmis- sion chains. We formally define the close contact and contact reachability. Based on them, we propose a potential transmission cluster model (Section 3).

• We analyze the properties of our potential transmission cluster model, satisfying the good desiderata of close social distance, arbitrary structural shape, and incubation-aware transmis-sion in real applications. Moreover, we discuss the particular usability in details to show our model flexibility (Section 4).

• We first develop an online search approach to find potential transmission clusters in an on-the-fly manner. To achieve the efficient search, we propose an offline indexing approach to construct multigraph for keeping all the records of close contact in history (Section 5).

• We further optimize the multigraph index by constructing a compact BCG-Index in an efficient space cost. We develop the techniques of spatio-temporal entity compressions and graph partitions, which shrink the index into multiple small bipartite graphs, which can support the fast potential cluster discovery. Moreover, we theoretically compare and analyze the algorithm complexity of online search methods (based on binary search and R-tree) and graph index methods (MG-Indexing and BCG-Indexing) (Section 6).

• We conduct extensive experiments on four real-world datasets of check-in records in geo-social networks. We generate the ground-truth of virus transmissions and infected users via the propagation simulation using a classical independent cascade model. We also conduct a case study of COVID-19 transmission following real-world dataset. The results validate the effectiveness and efficiency of our potential transmission cluster model and proposed algorithms (Section 7).

We discuss related work in Section 2 and conclude the paper in Section 8.

**2**  **RELATED WORK**

Our work is related to COVID-19 transmission analytics and spatio-temporal mining.

COVID-19 transmission analytics. Under the COVID-19 pandemic, numerous studies have tried to discover suspected clusters in order to track the COVID-19 transmissions [1, 9, 11, 14, 16, 18, 26, 35–37, 41]. Recently, various COVID-19 relevant models are proposed for COVID-19 forecasting [1, 37] and spread prediction [16]. Luo et al. [26] developed a explorer system to monitor spatio-temporal data of COVID-19. A graph embedding approach is proposed to help identify COVID-19 cases [41]. A warning system is designed to predict the hazard area, by collecting data from websites and using machine learning approaches to analyze the relevant features [9]. Kim et al. proposed a deep learning approach and a hierarchical model for estimating the number of imported COVID-19 cases from abroad [18]. On other hand, some researchers aim to explore the impact of certain containment measures on the transmission of COVID-19. The system dynamic simulation model is used to discover how physical distance measures influence the infectious [36] . [14] compares the impact of different control policies on the spread of COVID-19 and investigates the influence of heterogeneity

ACM Trans. Knowl. Discov. Data.

A Novel Graph Indexing Approach for Uncovering Potential COVID-19 Transmission Clusters 5

of urban mobility during the propagation. [11] focus to lift mitigation measures using deep learning approaches. Previous tracking systems can be also found in [3, 46, 52]. This work of transmission cluster discovery is also related to clustering algorithms in graphs [20, 21]. Different from the above studies, our work leverages the spatio-temporal activities to efficiently find query-dependent suspected clusters using a graph indexing approach.

Spatio-temporal mining. There exist several studies on the spatio-temporal data mining [2, 4, 7, 8, 13, 23, 32, 33, 43]. A comprehensive survey of spatio-temporal data mining can be found in [4]. Wu et al. study the problem to mine the reachable area from a given location and temporal period and propose a data-driven method to tackle the problem based on historical trajectory dataset [43]. Francalanci et al. [8] propose to analyze the evolving information for spatio-temporal queries. Ahmed et al. [2] develop a tracking device to find the most frequent terms in spatio-temporal region for each query. The above spatio-temporal data mining studies work on the reachable region, evolving information analysis, and frequent terms discovery tasks. There exist also many works on developing the learning based models for spatio-temporal prediction [7, 23, 32]. In addition, there exists a line of work [45, 53] studying the reachability problem in temporal graphs. Wu et al. [45] proposed an indexing based technique to answer reachability and time-based path queries in a temporal graph. Zhang et al. [53] developed a labeling scheme of temporal vertex labeling over distributed temporal graphs. Both these two works use index based method to answer the reachability queries in the temporal graphs. R-tree [13, 28, 33] is a classical tree index to store spatial objects in database, which accelerates the nearest neighbor search or the objects within a given spatial interval. However, for detecting a transmission cluster in this paper, the R-tree-based online search approaches cannot avoid a large number of query times interacting with spatial-temporal database, leading to the inefficiency. A detailed comparison to R-tree-based approaches can be found in Section 1, Section 6.3, and Section 7, in terms of motivations, algorithm complexities, and also experimental evaluations, respectively. Different from these studies, we focus on the close contact modeling and develop fast querying algorithms for potential COVID-19 transmission cluster discovery, which uses spatio-temporal logs to build graph indexes.

**3**  **PRELIMINARIES**

We are given a database

6 Xuliang Zhu, Xin Huang, Longxu Sun, and Jiming Liu

contact, i.e.,

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| A Novel Graph Indexing Approach for Uncovering Potential COVID-19 Transmission Clusters | | | | | 7 |
| **q** |  | **q** | **q** |  |
| **(a)** Chain | | **(b)** Star | **(c)** Quasi-Clique | |

Fig. 3. Arbitrary structural shapes of transmission clusters

1. (Participation). The query user

8 Xuliang Zhu, Xin Huang, Longxu Sun, and Jiming Liu

• Second, the contagious effect w.r.t. the incubation period Δ

|  |  |
| --- | --- |
| A Novel Graph Indexing Approach for Uncovering Potential COVID-19 Transmission Clusters | 9 |

Algorithm 1 Online Potential Transmission Cluster Discovering

Input: Spatial-temporal database

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 10 | **Query:** (Amy, 'May 17 00:00') | | | | | | | Amy | Xuliang Zhu, Xin Huang, Longxu Sun, and Jiming Liu | | | | | | | | | | | |
| Bob | | | | | May 9 15:00 | | Ella | May 20 14:01 | | Lily | |
| Starbucks, May 5 18:00 | | | | | | | | | | | |
| ID | User | | | Location | Time | |
| Starbucks, May 5 18:05 | | | | | | | | | | | |
| 14 | 1 | Amy | | Starbucks | | May 5 18:00 | | Bob | Starbucks, May 5 18:10 | | | | | | | | | | | |
| 2 | Bob | | Starbucks | | May 5 18:05 | |
| Ikea, May 9 15:00 | Amy | | | May 5 18:10 | | Cora | Finn | | May 20 14:03 | Holly | |
| days | 3 | Cora | | | Starbucks | | May 5 18:10 | Cora |
| Ikea, May 9 15:00 | **(c)** Multigraph Index | | | | | | | | | | |
| 4 | Bob | | Ikea | | May 9 15:00 | |
| 5 | Ella | | Ikea | | May 9 15:00 | | Ella | Starbucks, May 16 10:00 | Index | **B1** | Amy | | | | Starbucks, May 5 18:05 **r1** | | | | 14 |
| 6 | Amy | | Starbucks | | May 16 10:00 | | Starbucks, May 16 10:10 | Construction | Bob | | | |
| Lily |
| Ikea, May 9 15:00 | | | **r2** |
| 7 | Bob | | Starbucks | | May 16 10:10 | | Starbucks, May 20 14:00 | Cora | | | | days |
| 8 | Lily | | Starbucks | | May 20 14:00 | | Finn | Starbucks, May 20 14:01 | Ella | | | | | | Starbucks, May 16 10:05 **r3** | | | | |
| 9 | Ella | | Starbucks | | May 20 14:01 | | Holly |
| Starbucks, May 20 14:02 | Ella | | | | | | | | | | |
| 10 | | Finn | | Starbucks | May 20 14:02 | |
| Starbucks, May 20 14:03 | **B2** | | Lily | | | | Starbucks, May 20 14:02 **r4** | | | | 14 |
| 11 | Holly Starbucks | | | | May 20 14:03 | | Finn | | | | days |

Holly

|  |  |  |
| --- | --- | --- |
| **(a)** Spatio-temporal Database D | **(b)** Bipartite Contact Relationship | **(d)** BCG-Index |

Fig. 4. An example of graph indexing framework

Algorithm 2 Multigraph Indexing Approach for Potential Cluster Identification

Input: A multigraph G(

|  |  |
| --- | --- |
| A Novel Graph Indexing Approach for Uncovering Potential COVID-19 Transmission Clusters | 11 |

Algorithm 3 BCG-Index Construction

Input: Spatio-temporal database

12 Xuliang Zhu, Xin Huang, Longxu Sun, and Jiming Liu

Bipartite Contact Graph. We present a new definition of bipartite contact graph. Let be the bipartite graph

|  |  |
| --- | --- |
| A Novel Graph Indexing Approach for Uncovering Potential COVID-19 Transmission Clusters | 13 |

Algorithm 4 BCG-Index based Potential Transmission Cluster Tracking

Input: BCG-Index B, a query user

14 Xuliang Zhu, Xin Huang, Longxu Sun, and Jiming Liu

For the online potential transmission cluster search in Algorithm 1, if we implement the close contact using the R-tree index instead of the binary search, the close contact of a new query takes

A Novel Graph Indexing Approach for Uncovering Potential COVID-19 Transmission Clusters 15

Table 2. The statistics of check-in datasets.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 104 | 1 | Name | | | | | | | | | | | | | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 16 | | 1 | Our Model | | | 30 | 100 | 1 | Our Model | | | 30 | Xuliang Zhu, Xin Huang, Longxu Sun, and Jiming Liu | | | | | | | | | | | | | |
| 300 | | 1000 | | | | | | | 1200 | | | | | | |
| Our Model | | | | | | | Our Model | | | | | | |
| 250 | | 1-Hop Close Contact | | | 80 | 1-Hop Close Contact | | | 800 | | | 1-Hop Close Contact | | | | 1000 | | 1-Hop Close Contact | | | | |
| 200 | | 60 | 600 | | | | | | | 800 | | | | | | |
| 150 | | 600 | | | | | | | | | | | | | |
| 40 |
| 100 | | 400 | | | | | | | | | | | | | |
| 400 | | | | | | | | | | | | | |
| 20 |
| # Suspected Users | 50 | # Suspected Users | 200 | | | | | | # Suspected Users | 200 | | | | | |
| 3 | 7 | 14 | # Suspected Users | 3 | 7 | 14 |
| 0 | | 0 | 0 | | | | | | | 0 | | | | | | |
| 1 | | 3 | | 7 | 14 | 30 | 1 | | | 3 | 7 | 14 | 30 |
| Incubation Period Δt (days) | | | Incubation Period Δt (days) | | | Incubation Period Δt (days) | | | | | | | Incubation Period Δt (days) | | | | | | |
| (a) Brightkite | | | (b) Gowalla | | | (c) Foursquare | | | | | | | (d) LBSN | | | | | | |

Fig. 6. The average number of potential infecting users in 1-hop close contacts and potential trans-

mission clusters on four real datasets varied by the incubation period Δ

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A Novel Graph Indexing Approach for Uncovering Potential COVID-19 Transmission Clusters | | | | | | | | | | | | | | | | | | | | | | | | | | | 30 | 17 |
| 8% | |  | | --- | | Our Model  1-Hop Close Contact | | | | | | | 8% | | |  | | --- | | Our Model  1-Hop Close Contact | | | | | | | 20% | | Our Model | | | 20% | | | Our Model | | | |
| 6% | 6% | | 15% | | 1-Hop Close Contact | | | 15% | | | 1-Hop Close Contact | | | |
| 4% | 4% | |
| 10% | | 10% | | |
| 2% | 2% | | 5% | | 5% | | |
| 0% | 0% | | 0% | | 0% | | |
| Rate of Missing Infected Users | 1 | 3 | 7 | 14 | 30 | Rate of Missing Infected Users | | 1 | | 3 | 7 | 14 | 30 | Rate of Missing Infected Users | | 1 | 3 | 7 | 14 | 30 | Rate of Missing Infected Users | 1 | | 3 | 7 | 14 |
| Incubation Period Δt (days) | | | | | | Incubation Period Δt (days) | | | | | | | | Incubation Period Δt (days) | | | | | | Incubation Period Δt (days) | | | | | | |
| (a) Brightkite | | | | | | (b) Gowalla | | | | | | | | (c) Foursquare | | | | | | (d) LBSN | | | | | | |

Fig. 7. The average rate of missing infected users in 1-hop close contacts and potential transmission

clusters on four real datasets varied by the incubation period Δ

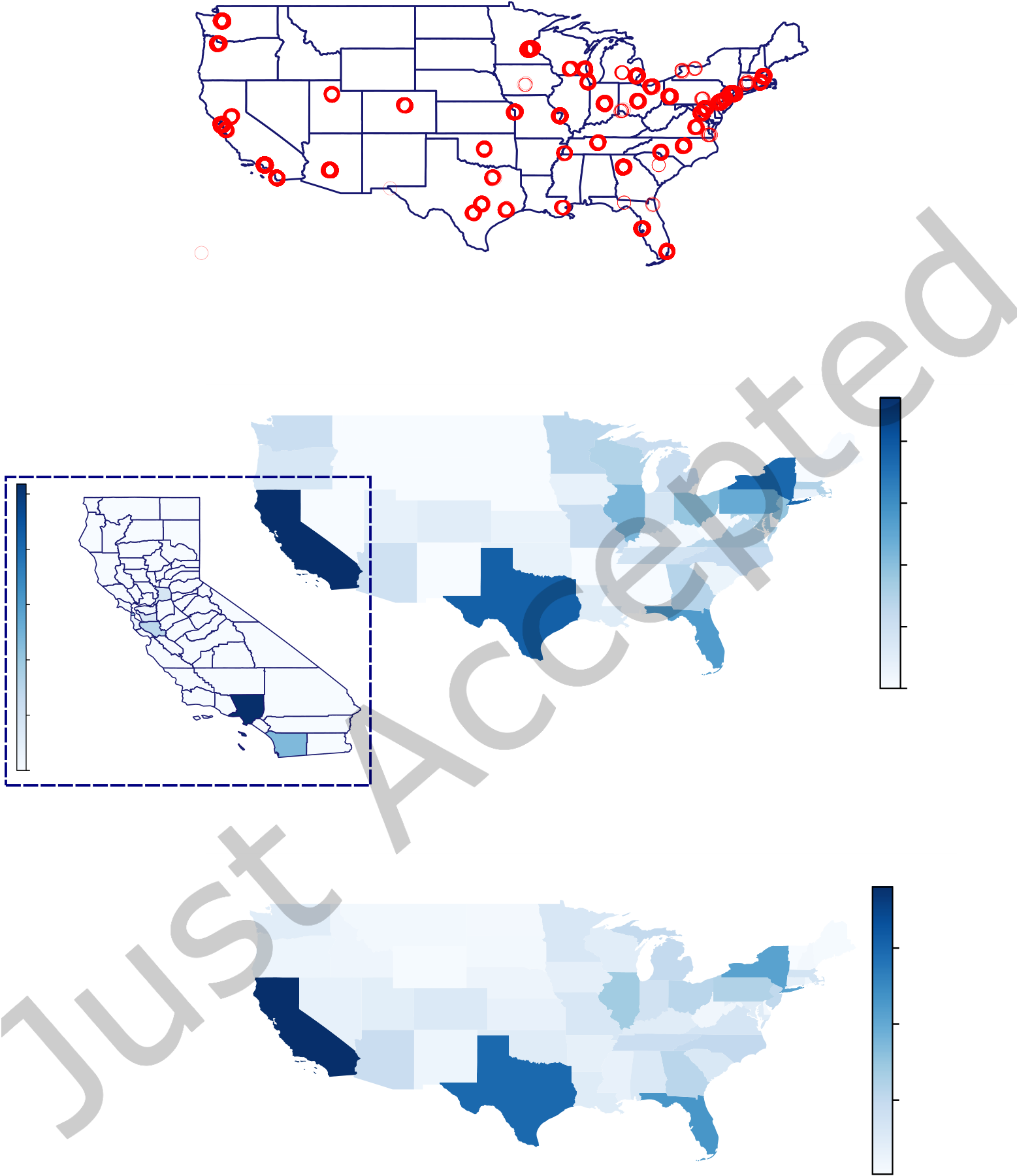
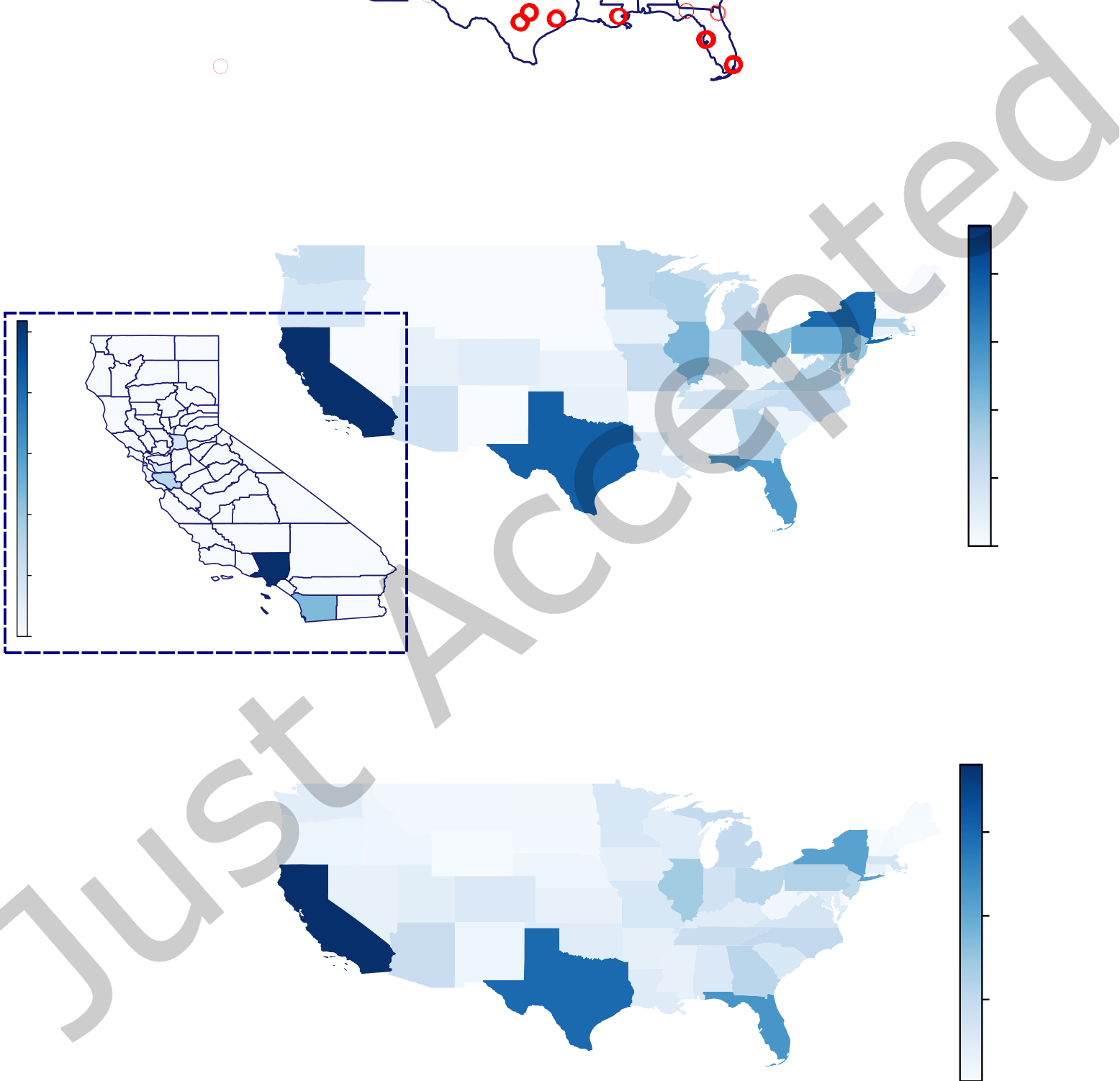
18 Xuliang Zhu, Xin Huang, Longxu Sun, and Jiming Liu

Table 3. The index size (in Megabytes) of different indexing methods and various parameters on all datasets. Here, **K**= 103and **M**=106.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A Novel Graph Indexing Approach for Uncovering Potential COVID-19 Transmission Clusters | | | | | | | | | | | | | | | | 19 |
| 6% | | BK | | | | | | 500 | | BK | | | | FQ | |
| 5% | | GW | | | | | |
| 4% | | GW | | | | LBSN | |
| FQ | | | | | |
| 3% | | LBSN | | | | | | 400 | | | | | | | |
| 2% | | | | | | | | 300 | | | | | | | |
| 1% | | | | | | | | 200 | | | | | | | |
| |  |  | | --- | --- | | 0% | 100 | | | | | | | | | | | | | | | | |
| Percent of Missing Suspected Cluster | 0% | | 2% | 4% | 6% | 8% | 10% | Size of Suspected Cluster | 0 | 0% | 2% | 4% | 6% | 8% | 10% |
| Missing Data | |
| Missing Data | |
| (a) The percentage of missing infecting users | | | | | | | | (b) The number of infecting users in potential | | | | | | | |
| in potential clusters | | | | | | | | clusters | | | | | | | |

Fig. 10. Quality evaluation of our potential transmission cluster model with missing records.

BCG-Indexing only takes nearly 1,000 seconds on Brightkite-Syn and Gowalla-Syn, and 5,000 seconds on Foursquare. In addition, Table 3 reports the index size on disk and various parameters in complexity analysis for R-tree, MG-Indexing, and BCG-Indexing. For the index size, BCG-Indexing takes about three times as the size of original dataset in the worst cases, which is much smaller than R-tree and MG-Indexing. Moreover, the parameters



20 Xuliang Zhu, Xin Huang, Longxu Sun, and Jiming Liu

A Transmission Source

(a) Initial transmission sources. Each transmission source is represented by a red circle.

The region located in a larger and darker red area contains more transmission sources.

|  |  |  |  |
| --- | --- | --- | --- |
| 25000 | Potential Infected Cases | 40000 | Potential Infected Cases |
| 30000 |
| 20000 | 20000 |
| 15000 | 10000 |
| 10000 | 0 |

5000

0

(b) Our potential infected cases

1e6

3

2

1

Infected Cases

(c) Real-world infected cases

Fig. 11. A case study of COVID-19 transmission using Foursquare dataset in United States.

of our model, we also report the real-world COVID-19 confirmed cases in Figure 11(c). Compare

the results in Figures 11(b) and 11(c), we observe that our model detect no potential transmission

ACM Trans. Knowl. Discov. Data.

A Novel Graph Indexing Approach for Uncovering Potential COVID-19 Transmission Clusters 21

clusters in some states, i.e. Idaho and Montana. Because the Foursquare dataset has no check-in records located in these states. Nevertheless, our distribution results in Figure 11(b) are very similar with the real US confirmed cases in Figure 11(c). The four states California, Texas, New York, and Florida have the maximum potential infected users, which are the same as the real-world confirmed cases for all time in US [39]. This confirms the effectiveness of our proposed potential transmission cluster model in the case of COVID-19 transmissions.

**8**  **CONCLUSION AND FUTURE WORK**

In this paper, we motivate and investigate the problem of discovering query-dependent COVID-19 transmission clusters on spatio-temporal logs, which finds all potential infected users to a given query of patient user and infection time. We propose three different methods including the online search approach and two indexing based solutions. Our novel BCG-indexing approach achieves a good balance of index construction and online query processing for fast suspected cluster discovery. Extensive experiments on real-world datasets validate the effectiveness of our suspected cluster model and query-dependent suspected cluster tracking algorithms.

Although the proposed BCG-indexing methods can efficiently uncover potential COVID-19 transmission clusters, the current algorithms leave three open issues for further improvements. First, the querying algorithms are designed only for one single query. In real applications, several patients may be confirmed simultaneously. Thus, it needs an efficient transmission cluster search for multiple queries. We could further explore the fast detection of potential transmission clusters for multiple queries in a batch. Second, the current algorithm of BCG-index construction is static, which lies on the predefined parameters



22 Xuliang Zhu, Xin Huang, Longxu Sun, and Jiming Liu

[7] Rui Dai, Shenkun Xu, Qian Gu, Chenguang Ji, and Kaikui Liu. 2020. Hybrid Spatio-Temporal Graph Convolutional Network: Improving Traffic Prediction with Navigation Data. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 3074–3082.

[8] Chiara Francalanci, Barbara Pernici, and Gabriele Scalia. 2017. Exploratory spatio-temporal queries in evolving information. In International Workshop on Mobility Analytics for Spatio-temporal and Social Data. 138–156.

[9] Zhenxin Fu, Yu Wu, Hailei Zhang, Yichuan Hu, Dongyan Zhao, and Rui Yan. 2020. Be Aware of the Hot Zone: A Warning System of Hazard Area Prediction to Intervene Novel Coronavirus COVID-19 Outbreak. In SIGIR. 2241–2250. [10] Fred Galvin. 1995. The list chromatic index of a bipartite multigraph. Journal of Combinatorial Theory, Series B 63, 1 (1995), 153–158.

[11] Salah Ghamizi, Renaud Rwemalika, Maxime Cordy, Lisa Veiber, Tegawendé F Bissyandé, Mike Papadakis, Jacques Klein, and Yves Le Traon. 2020. Data-driven Simulation and Optimization for Covid-19 Exit Strategies. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 3434–3442.

[12] Jonathan L Gross and Jay Yellen. 2005. Graph Theory and Its Applications. CRC press.

[13] Antonin Guttman. 1984. R-trees: A dynamic index structure for spatial searching. In ACM SIGMOD International Conference on Management of Data. 47–57.

[14] Qianyue Hao, Lin Chen, Fengli Xu, and Yong Li. 2020. Understanding the Urban Pandemic Spreading of COVID-19 with Real World Mobility Data. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 3485–3492.

[15] Frank Harary. 1991. Graph Theory. Addison-Wesley.

[16] Xiaoyong Jin, Yu-Xiang Wang, and Xifeng Yan. 2021. Inter-Series Attention Model for COVID-19 Forecasting. In Proceedings of the 2021 SIAM International Conference on Data Mining (SDM). 495–503.

[17] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 137–146.

[18] Minseok Kim, Junhyeok Kang, Doyoung Kim, Hwanjun Song, Hyangsuk Min, Youngeun Nam, Dongmin Park, and Jae-Gil Lee. 2020. Hi-COVIDNet: Deep Learning Approach to Predict Inbound COVID-19 Patients and Case Study in South Korea. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 3466–3473. [19] Jure Leskovec and Andrej Krevl. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. [http://snap.stanford.](http://snap.stanford.edu/data)  [edu/data](http://snap.stanford.edu/data).

[20] Hui-Jia Li, Zhan Bu, Zhen Wang, and Jie Cao. 2019. Dynamical clustering in electronic commerce systems via optimization and leadership expansion. IEEE Transactions on Industrial Informatics 16, 8 (2019), 5327–5334.

[21] Hui-Jia Li, Lin Wang, Yan Zhang, and Matjaž Perc. 2020. Optimization of identifiability for efficient community detection. New Journal of Physics 22, 6 (2020), 063035.

[22] Hui-Jia Li, Wenzhe Xu, Shenpeng Song, Wen-Xuan Wang, and Matjaž Perc. 2021. The dynamics of epidemic spreading on signed networks. Chaos, Solitons & Fractals 151 (2021), 111294.

[23] Ting Li, Junbo Zhang, Kainan Bao, Yuxuan Liang, Yexin Li, and Yu Zheng. 2020. Autost: Efficient neural architecture search for spatio-temporal prediction. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 794–802.

[24] Yang Liu, Zhonglei Gu, and Jiming Liu. 2021. Uncovering transmission patterns of COVID-19 outbreaks: A region-wide comprehensive retrospective study in Hong Kong. EClinicalMedicine (2021), 100929.

[25] Yang Liu, Zhonglei Gu, Shang Xia, Benyun Shi, X-N Zhou, Yong Shi, and Jiming Liu. 2020. What are the underlying transmission patterns of covid-19 outbreak?–an age-specific social contact characterization. EClinicalMedicine (2020), 100354.

[26] Yuyu Luo, Wenbo Li, Tianyu Zhao, Xiang Yu, Lixi Zhang, Guoliang Li, and Nan Tang. 2020. Deeptrack: Monitoring and exploring spatio-temporal data: a case of tracking COVID-19. Proceedings of the VLDB Endowment 13, 12 (2020), 2841–2844.

[27] Nikos Mamoulis, Huiping Cao, George Kollios, Marios Hadjieleftheriou, Yufei Tao, and David W Cheung. 2004.

Mining, indexing, and querying historical spatiotemporal data. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 236–245.

[28] Yannis Manolopoulos, Apostolos N Papadopoulos, Apostolos N Papadopoulos, and Yannis Theodoridis. 2006. R-Trees: Theory and Applications: Theory and Applications. Springer Science & Business Media.

[29] Lukas M Marti, Michael P Dal Santo, and Ronald Keryuan Huang. 2016. Modeling significant locations. US Patent 9,267,805.

[30] Mirco Nanni, Gennady Andrienko, Albert-László Barabási, Chiara Boldrini, Francesco Bonchi, Ciro Cattuto, Francesca Chiaromonte, Giovanni Comandé, Marco Conti, Mark Coté, et al. 2021. Give more data, awareness and control to individual citizens, and they will help COVID-19 containment. Ethics and Information Technology (2021), 1–6. [31] Bing Ni, Qiaomu Shen, Jiayi Xu, and Huamin Qu. 2017. Spatio-temporal flow maps for visualizing movement and contact patterns. Visual Informatics 1, 1 (2017), 57–64.

ACM Trans. Knowl. Discov. Data.



A Novel Graph Indexing Approach for Uncovering Potential COVID-19 Transmission Clusters 23

[32] Maya Okawa, Tomoharu Iwata, Takeshi Kurashima, Yusuke Tanaka, Hiroyuki Toda, and Naonori Ueda. 2019. Deep mix-ture point processes: Spatio-temporal event prediction with rich contextual information. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 373–383.

[33] Dimitris Papadias, Yufei Tao, P Kanis, and Jun Zhang. 2002. Indexing spatio-temporal data warehouses. In IEEE International Conference on Data Engineering. 166–175.

[34] Zhe Peng, Jinbin Huang, Haixin Wang, Shihao Wang, Xiaowen Chu, Xinzhi Zhang, Li Chen, Xin Huang, Xiaoyi Fu, Yike Guo, et al. 2021. BU-Trace: A Permissionless Mobile System for Privacy-Preserving Intelligent Contact Tracing. In DASFAA 2021 International Workshops: BDQM, GDMA, MLDLDSA, MobiSocial, and MUST. 381–397.

[35] Zhe Peng, Cheng Xu, Haixin Wang, Jinbin Huang, Jianliang Xu, and Xiaowen Chu. 2021. P2B-Trace: Privacy-Preserving Blockchain-based Contact Tracing to Combat Pandemics. In Proceedings of the 2021 International Conference on Management of Data. 2389–2393.

[36] Putsadee Pornphol and Suphamit Chittayasothorn. 2020. System Dynamics Model of COVID-19 Pandemic Situation: The Case of Phuket Thailand. In International Conference on Computer Modeling and Simulation. 77–81.

[37] Amray Schwabe, Joel Persson, and Stefan Feuerriegel. 2021. Predicting COVID-19 Spread from Large-Scale Mobility Data. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 3531–3539.

[38] Yufei Tao and Dimitris Papadias. 2001. The mv3r-tree: A spatio-temporal access method for timestamp and interval queries. In International Conference on Very Large Data Bases. 431–440.

[39] The New York Times. 2021. Coronavirus (Covid-19) Data in the United States. [https://github.com/nytimes/covid-19-](https://github.com/nytimes/covid-19-data) [data](https://github.com/nytimes/covid-19-data).

[40] Vincent S Tseng, Josh Jia-Ching Ying, Stephen TC Wong, Diane J Cook, and Jiming Liu. 2020. Computational intelligence techniques for combating COVID-19: a survey. IEEE Computational Intelligence Magazine 15, 4 (2020), 10–22.

[41] Bowen Wang, Yanjing Sun, Trung Q Duong, Long D Nguyen, and Lajos Hanzo. 2020. Risk-Aware Identification of Highly Suspected COVID-19 Cases in Social IoT: A Joint Graph Theory and Reinforcement Learning Approach. IEEE Access 8 (2020), 115655–115661.

[42] WHO. 2020. [https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200306-sitrep-46-covid-19.](https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200306-sitrep-46-covid-19.pdf?sfvrsn=96b04adf_4)  [pdf?sfvrsn=96b04adf\_4](https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200306-sitrep-46-covid-19.pdf?sfvrsn=96b04adf_4).

[43] Guojun Wu, Yichen Ding, Yanhua Li, Jie Bao, Yu Zheng, and Jun Luo. 2017. Mining spatio-temporal reachable regions over massive trajectory data. In IEEE International Conference on Data Engineering. 1283–1294.

[44] Huanhuan Wu, James Cheng, Yi Lu, Yiping Ke, Yuzhen Huang, Da Yan, and Hejun Wu. 2015. Core decomposition in large temporal graphs. In IEEE International Conference on Big Data. 649–658.

[45] Huanhuan Wu, Yuzhen Huang, James Cheng, Jinfeng Li, and Yiping Ke. 2016. Reachability and time-based path queries in temporal graphs. In IEEE International Conference on Data Engineering. 145–156.

[46] Marcin Wylot, Philippe Cudré-Mauroux, Manfred Hauswirth, and Paul Groth. 2017. Storing, tracking, and querying provenance in linked data. IEEE Transactions on Knowledge and Data Engineering 29, 8 (2017), 1751–1764.

[47] Xiaopeng Xiong, Mohamed F Mokbel, and Walid G Aref. 2005. Sea-cnn: Scalable processing of continuous k-nearest neighbor queries in spatio-temporal databases. In IEEE International Conference on Data Engineering. 643–654. [48] Zhe Xu, Lei Shi, Yijin Wang, Jiyuan Zhang, Lei Huang, Chao Zhang, Shuhong Liu, Peng Zhao, Hongxia Liu, Li Zhu, et al. 2020. Pathological findings of COVID-19 associated with acute respiratory distress syndrome. The Lancet Respiratory Medicine (2020).

[49] Dingqi Yang, Bingqing Qu, Jie Yang, and Philippe Cudre-Mauroux. 2019. Revisiting user mobility and social relation- ships in lbsns: a hypergraph embedding approach. In The world wide web conference. 2147–2157.

[50] Dingqi Yang, Daqing Zhang, Longbiao Chen, and Bingqing Qu. 2015. Nationtelescope: Monitoring and visualizing large-scale collective behavior in lbsns. Journal of Network and Computer Applications 55 (2015), 170–180.

[51] Dingqi Yang, Daqing Zhang, and Bingqing Qu. 2016. Participatory cultural mapping based on collective behavior data in location-based social networks. ACM Transactions on Intelligent Systems and Technology 7, 3 (2016), 1–23.

[52] Zhao Yang and Nathalie Japkowicz. 2017. Meta-morisita index: Anomaly behaviour detection for large scale tracking data with spatio-temporal marks. In IEEE International Conference on Data Mining Workshops. 675–682.

[53] Tianming Zhang, Yunjun Gao, Lu Chen, Wei Guo, Shiliang Pu, Baihua Zheng, and Christian S Jensen. 2019. Efficient distributed reachability querying of massive temporal graphs. The VLDB Journal 28, 6 (2019), 871–896.

ACM Trans. Knowl. Discov. Data.