Privacy Preservation of COVID-19 Contact Tracing Data

Anifat M. Olawoyin, Carson K. Leung(✉), and Qi Wen
Department of Computer Science
University of Manitoba
Winnipeg, MB, Canada
✉ kleung@cs.umanitoba.ca

***Abstract*—Big data are everywhere. Examples of big data include contact tracing data of patients who contracted coronavirus disease 2019 (COVID-19). On the one hand, mining these contact tracing data can be for social good. For instance, it helps slow down the spread of COVID-19. It also helps people diagnosed with COVID-19 get referrals for services and resources they may need to isolate safely. On the other hand, it is also important to protect the privacy of these COVID-19 patients. Hence, we present in this paper a solution for privacy preservation of COVID-19 contact tracing data. Specifically, our solution preserves the privacy of individuals by publishing only their spatio-temporal representative locations. Evaluation results on real-life COVID-19 contact tracing data from South Korea demonstrate the effectiveness and practicality of our solution in preserving the privacy of COVID-19 contact tracing data.**

***Index Terms*—computer science, information technology, database and data management, big data, data science and systems, data and informatics, information security, privacy, data mining, privacy preserving data mining, spatio-temporal data, spatial data, temporal data, spatio-temporal hierarchy, visualization**

I. Introduction

Modern technological advancements have contributed to the rapid generation and collection of very large volumes of data from a wide variety of rich data sources in numerous real-life applications. These include:

- networks (e.g., social networks [1–5], transportation networks [6–9]);
- financial time series [10];
- biomedical data (e.g. disease reports [11, 12], Genomic data [13–16], epidemiological data [17, 18]); and
- trajectory data [19, 20] (e.g., transportation trajectories, weather trajectories [21, 22]).

As a consequence, big data [23–25] are everywhere. These big data may be at different veracity levels (e.g., imprecise or uncertain data [26–30], precise data).

Coronavirus disease 2019 (COVID-19), which was caused by the acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has led to rapid generation and collection of valuable big data. Examples of big data related to COVID-19 include epidemiological data and statistics [31], disease reports and blood test results [32, 33], computed tomography (CT) scan images of lungs [34, 35], indicators for measuring impacts of COVID-19 on various socio-economic aspects of our daily life [36, 37], as well as contact tracing data and trajectories. Applying data science [38, 39]—which makes good use of a fusion of data mining [40–47], machine learning [48–51], mathematics and statistics [52, 53], informatics [54, 55], data analysis [56–61], and visualization [62–65]—to these big data (e.g., contact tracing data) can be for social good. For instance, it discovers implicit, previously unknown and potentially useful knowledge, which may help slow down the spread of COVID-19. It may also help people diagnosed with COVID-19 get referrals for services and resources they may need to isolate safely. In addition, advancement in communication and computing technologies [66–69] has made it easy for governments and research institutions to invent new solutions [70–72] to combat the spread of diseases. These solutions[1] include:

- alerting apps such as COVID Alert[2] (a COVID-19 exposure notification app used in Canada);
- contact tracing apps such as COVIDSafe[3] used in Australia;
- information apps such as Coronavírus - SUS[4] used in Brazil;
- medical reporting apps such as allertaLOM[5] used in the Italian region of Lombardia;
- quarantine enforcement apps such as Stay Home Safe[6] used in Hong Kong; and
- self-diagnostic apps such as CoronApp[7] used in Chile.

While applying data science to the big data (e.g., COVID-19 contact tracing data) can be for social good, it is also important to protect the privacy of these COVID-19 patients. A way to protect their privacy is to anonymize data. Generally, there are two broad categories of data anonymization techniques:

- syntactic models [73–75], which are characterized by generalization or suppression. However, this often result in loss of information.
- differential privacy [76–79], which is characterized by their application of a random noise so that any addition

[1] https://www.coe.int/en/web/data-protection/contact-tracing-apps
[2] https://www.canada.ca/en/public-health/services/diseases/coronavirus-disease-covid-19/covid-alert.html
[3] https://www.health.gov.au/resources/apps-and-tools/covidsafe-app
[4] https://www.gov.br/pt-br/apps/coronavirus-sus
[5] https://www.allertalom.regione.lombardia.it/homepage
[6] https://www.coronavirus.gov.hk/eng/stay-home-safe.html
[7] https://coronapp.gob.cl/

978-1-6654-6667-7/21/$31.00 ©2021 IEEE
DOI 10.1109/IUCC-CIT-DSCI-SmartCNS55181.2021.00055

of a data point to the dataset (or removal of a data point from the dataset) may not have significant effects on the outcome. However, addition of noise may distort the dataset beyond its usefulness.

Thus, utility-privacy trade off is a common concern in data anonymization.

To facilitate discovery of useful knowledge from COVID-19 contact tracing data via data science (or data mining) while preserving privacy of COVID-19 patients, we present in this paper a solution for privacy preservation of COVID-19 contact tracing data. Specifically, we present the concept of spatio-temporal representative point, which is a non-trivial extension of the representative point used in geographical information system (GIS) for the temporal correlated dataset. Our solution preserves the privacy of individuals by publishing only their spatio-temporal representative locations via differential privacy mechanism. It maintains a good balance of anonymization and data utility. The solution builds a spatio-temporal hierarchy and properly aggregates the counts of attributes at different levels of the hierarchy.

As a preview, when applying our solution to the dataset in Table I, we focus on anonymizing temporal and spatial attributes within the dataset. Other attributes (e.g., type, district, City, ID) may be quasi-identifiers or sensitive identifiers that can be aggregated. Generally, we select an identifier as a grouping reference for utility measurement, and aggregate all other attributes by a frequency count query like:

```
SELECT COUNT(*)
FROM dataset name
GROUP BY gr, date, temporal point
```

where *gr* is the grouping reference. We will apply our solution to a real-life South Korean COVID-19 patient route dataset.

We organize the rest of this paper as follows. The next section provides background and related works, and Section III presents our privacy-perserving solution . Implementation details are described in Section IV, and Section V describes the experimental evaluation of our proposed solution. Conclusions an future work are presented in Section VI.

II. Background and Related Works

Over the past few decades, data mining algorithms has focused on achieving space and time efficiency leading to research such as bitwise pattern mining [80] and scalable vertical mining [81]. However, recent development and more specifically, the most recent COVID-19 pandemic has made anonymization an important concept in data mining. Thus, in recent years researchers have designed models and techniques for anonymization in data mining including anonymization by surrogate vectors and LFP-tree [82], addition of noise to trajectory [83], privacy preservation in uncertain bid data mining [84] and keyword search on encrypted outsourced data [85]. Anonymization has been added as a step in data mining to privately preserve the mining outcomes.

Eom et al. [82] combined length-based frequent pattern tree (LFP-tree) and surrogate vectors models for publishing of anonymized trajectory databases. The study first transformed the location data into surrogate vectors using two-dimensional space grids and then employed the novel length-based frequent pattern tree (LFP-tree) to skip unnecessary task while searching for minimum violating sequences (MVS). The surrogate vectors approach can generally be used to identify the shape of trajectory, measure efficiency of trajectory, and grasp user tendency. At the same time, it ensures that no single point is released or disclosed based on trust. As a preview, our solution uses temporal and spatial correlation in grouping records to (a) eliminate individual association with a record and (b) keep the temporal and spatial utilities of the dataset.

Wang et al. [86] extend maximal frequent itemsets mining on sensitive data by utilizing sequence exponential mechanism (SEM). Huo et al. [87] generalized stay of points for preserving privacy of trajectories. Dwork [88] proposed differential privacy, to ensure any addition of a data point to the dataset (or removal of a data point from the dataset) does not have significant difference in the probability of outcomes of any aggregate function. Lee et al. [89] examined current issues on management and control of privacy level in big data de-identification, and suggested solutions to these issues.

III. Our Privacy Preserving Solution

In this section, we present the overview of our solution, as shown in Fig. 1. We also discuss the three components of the solution: temporal hierarchy, spatio-temporal representative point and Laplace mechanism of differential privacy.

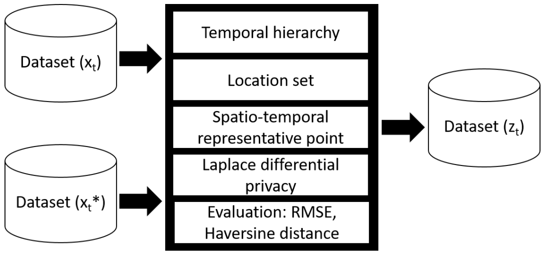

Dataset (x_t)
Dataset (x_t*)
Temporal hierarchy
Location set
Spatio-temporal representative point
Laplace differential privacy
Evaluation: RMSE, Haversine distance
Dataset (z_t)

Fig. 1: An overview of our solution

A. Temporal Hierarchy

We build temporal hierarchy, which aggregate time series data to a user-specified periodic level—like quarterly (Q), monthly (M), daily (D), and or hourly (H)—such that:

$$Y = \sum_{i=1}^{4} Q_i \tag{1a}$$

$$Q_i = \sum_{j=1}^{3} M_j \tag{1b}$$

$$M_j = \sum_{k=1}^{d} D_k \tag{1c}$$

$$D_n = \sum_{n=1}^{24} H_n \tag{1d}$$

TABLE I: South Korean patient route data

ID	Date	District ("gu")	City	Type	Latitude	Longitude
68	2020-Feb-25	Gangnam-gu	Seoul	public transportation	37.493601	127.079526
86	2020-Feb-25	Gangnam-gu	Seoul	public transportation	37.487468	127.101324
93	2020-Feb-25	Gangnam-gu	Seoul	public transportation	37.514236	127.031593
63	2020-Feb-26	Gangnam-gu	Seoul	restaurant	37.499907	127.037393
66	2020-Feb-26	Gangnam-gu	Seoul	restaurant	37.504689	127.043847
86	2020-Feb-26	Gangnam-gu	Seoul	restaurant	37.516567	127.021503
66	2020-Feb-26	Gangnam-gu	Seoul	cafe	37.505153	127.044054
86	2020-Feb-26	Gangnam-gu	Seoul	cafe	37.522466	127.037943
93	2020-Feb-26	Gangnam-gu	Seoul	pharmacy	37.509681	127.032382
86	2020-Feb-28	Gangnam-gu	Seoul	restaurant	37.516567	127.021503
93	2020-Feb-28	Gangnam-gu	Seoul	restaurant	37.514236	127.031593
105	2020-Feb-28	Gangnam-gu	Seoul	restaurant	37.504356	127.043652

TABLE II: Location set aggregated from the temporal hierarchy

Group#	#rec in group	Date	District	City	Type	Location set
1	3	2020-Feb-25	Gangnam-gu	Seoul	public transportation	{(37.493601, 27.079526), (37.487468, 127.101324), (37.514236, 127.031593)}
2	3	2020-Feb-26	Gangnam-gu	Seoul	restaurant	{(37.499907, 127.037393), (37.504689, 127.043847), (37.516567, 127.021503)}
3	2	2020-Feb-26	Gangnam-gu	Seoul	cafe	{(37.505153, 127.044054), (37.522466, 127.037943)}
4	1	2020-Feb-26	Gangnam-gu	Seoul	pharmacy	{(37.509681, 127.032382)}
5	3	2020-Feb-28	Gangnam-gu	Seoul	restaurant	{(37.516567, 127.021503), (37.514236, 127.031593), (37.504356, 127.043652)}

TABLE III: Spatio-temporal representative point

Group#	#rec in group	Date	District	City	Type	Spatio-temporal rep. pt.
1	3	2020-Feb-25	Gangnam-gu	Seoul	public transportation	(37.498435, 127.070814)
2	3	2020-Feb-26	Gangnam-gu	Seoul	restaurant	(37.507054, 127.034247)
3'	3	2020-Feb-26	Gangnam-gu	Seoul	**others**	(37.507054, 127.038126)
5	3	2020-Feb-28	Gangnam-gu	Seoul	restaurant	(37.511719, 127.032249)

B. *Spatio-Temporal Representative Points*

A way to attempt to preserve privacy of a collection spatial points appear within a temporal interval is to randomly select a point as a representative from the collection. However, a potentially problem is that such a selection may lead to bias (e.g., when a edge point is randomly selected).

A better way to preserve privacy of a collection n spatial points appear within a temporal interval is select their *centroid* as a *spatio-temporal representative point*. This centroid is a center of mass, a mean center of the coordinates, or an average of x- and y-coordinates over n observations:

$$R(x, y) = \left(\frac{\sum_i^n x_i}{n}, \frac{\sum_i^n y_i}{n} \right) \tag{2}$$

An alternative way to compute such a spatio-temporal representative point is to use the minimum square distance D_{min} between a set of coordinates (x, y) among all n points within the collection:

$$D_{min} = \min \sum \left[(x_i - x_j)^2 + (y_i - y_j)^2 \right] \tag{3}$$

We present the algorithm for finding spatio-temporal representative point in Algorithm 1. First, we generate the temporal hierarchy and create a location set (i.e., a matrix of coordinates at the same temporal level). Then, we compute the average of all coordinates as a spatio-temporal representative point using Eq. (2). To preserve privacy of individuals, if the computed centroid happens to match a real data point within the collection (i.e., within the location set), we add Laplace noise to preserve the privacy.

Algorithm 1 Spatio-temporal representative point with differential privacy

```
Input Dataset $D$, minimum support *minsup*
Output Differential-privacy aggregated dataset with temporal point $D'$
Compute Temporal Hierarchy $T$
Group by quasi-identifiers and create geometry list/matrix $M$ (i.e., location set)
Compute spatio-temporal representative point as an average of coordinates in $M$
```

Example Let us illustrate Algorithm 1 by considering route data between February 25-28, 2020, with 12 sample patient route information presented in Table I. All patient route records were from the South Korean capital city of Seoul, in particular, in the district of Gangnam-gu. The first three records occurred in February 25 when patients used public transportation. Among the next six records for February 26, three of them dined at restaurants, two visited a cafe, and the last record represents a visit to a pharmacy. The final three records occurred in February 28 when patients also dined at restaurants. Our algorithm builds a temporal hierarchy, groups these 12 records by *day*, district, city, type of establishment (e.g., public transportation, restaurant), and generates a coordinate list for the resulting group as shown in Table II. Afterwards, it computes the average of each of these four location sets as their spatio-temporal representative points for these four groups using Eq. (2). For any group not satisfying the *minsup* threshold (e.g., groups 3 and 4 do not satisfy *minsup* of 3 records), we change the type to "others", re-aggregate and test again. Specifically, for groups 3 and 4, the cafe and the pharmacy are grouped as "others" and an average of their coordinates becomes their spatio-temporal representative point for the resulting group. The final output of our sample dataset is presented in Table III where we have one point representing each group.

C. Differential Privacy via Laplace Mechanism

Our solution is guaranteed to preserve the privacy of both the actual timestamp and location by using:

- the temporal hierarchy to preserve the privacy of the actual timestamp, and
- the spatio-temporal representative point to preserve the privacy of the actual location.

If the computed centroid happens to be an actual location, it applies Laplace mechanism as shown in Algorithm 0 to add Laplace noise. The probability density function for this Laplace mechanism having mean μ and ϵ level of noise is:

$$f(x \mid \mu, \epsilon) = \frac{1}{2\epsilon} e^{-\left|\frac{x-\mu}{\epsilon}\right|} \tag{4}$$

Algorithm 2 Laplace Mechanism

```
Input Dataset $D$
Output Differential-privacy dataset $D'$
for each Temporal geometry formed by points in the location set do
    (x,y).add(Laplace noise)
```

By doing so, at any level t of the temporal hierarchy, a randomized mechanism A satisfies ϵ-differential privacy on spatio-temporal representative point provided that any location output z_t from a spatio-temporal dataset x_t and a neighbouring spatio-temporal dataset x_t* obtained by the addition or removal a record is not significantly different:

$$\frac{Pr\left(A(x_t = z_t)\right)}{Pr\left(A(x_t* = z_t)\right)} \leq e^{\epsilon} \tag{5}$$

See Fig. 1.

IV. Evaluation

To evaluate our solution, we compare it with existing baseline differential privacy by Laplace mechanism. For the Laplace mechanism, we generalize the timestamp to day level without aggregation and apply Laplace noise to each spatial record.

First, we evaluated the **impact of parameters**. We vary the value of ϵ from 0.01 to 0.1. The optimal difference is observed when ϵ = 0.09, which we consider as the optimal privacy suitable for protecting the published data with differential privacy mechanism. The spatio-temporal representative point differential privacy mechanism as shown in Fig. 2 provides protection for the patients such that their information is hidden in space and time. Individual patient could be traced in the original unprotected dataset using the spatial and temporal attributes. The identity of some patients could be traced by an adversary with background knowledge given the time and location visited. The patient visited different locations such as store and restaurants by public transportation before finally tested in the hospital. With our solution, the actual patient identity is unknown.

Moreover, we extend the notion of near object relationship to privacy by grouping all patients and their activities under temporal and spatial correlation. For instance, patients route trace in Fig. 3(a) shows all locations visited by 69 patients on February 17, 2020, with 128 points. Our version of the same patient route data after applying our solution is shown in Fig. 3(b), where no patient information is revealed. Privacy-preserved data are represented by 53 temporal representative points while the privacy is enhanced with Laplace differential privacy mechanism such that the knowledge of any adversary is limited to the trends within the aggregated time series data. The grouping of patients using temporal hierarchy still provides essential route information necessary for contact tracing without relating the information to any specific patient.

Next, we evaluated the **compression ratio**. We computed the spatial and file size compression ratio for our solution. Results presented in Table IV shows that our solution sub-

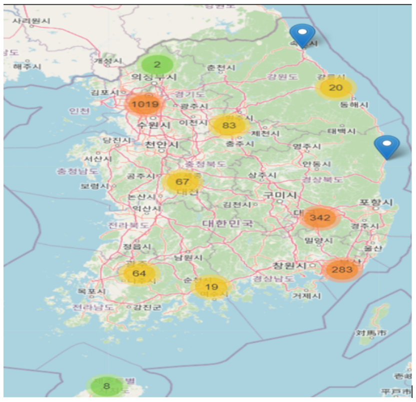
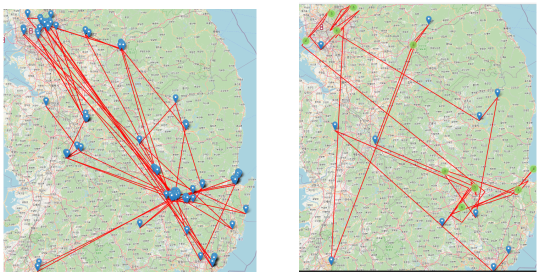


Fig. 2: SKCOVID19 points



(a) Patient trace Feb 17 (b) Protected trace Feb 17

Fig. 3: SKCOVID19 data before and after protection

TABLE IV: Spatial and File Compression Ratio

Model	#points	File size
Laplace	100%	100%
Our solution	34.7%	31.8%

stantially reduces both the number of points and file size when compared with the original number of points and original file size.

We also evaluated the **aggregate queries**. We considered aggregation without differential privacy by varying the temporal hierarchy level and aggregation parameter. When **aggregated by province, city and infection route**, we grouped the dataset by province, city and infection route while varying the temporal hierarchy by day, week and month. We then calculated the average for the spatial data to represent the group. The resulting count query result is presented in Table V. We observed that the aggregation by day after reclassifying all trajectories with less than *minsup*=3 occurrence as 'Others' has the lowest Pearson correlation and the highest root mean square error (RMSE) of 0.0483. This is expected due to the level of granularity of presenting daily data: Lot of the occurrence that are less than three are reclassified as 'others' thereby lowering the utility of the dataset. However, considering the disclosure risk, the daily aggregation is still very useful because of the time sensitive nature of the dataset. The data may become obsolete and less useful if such disclosure is delayed for instance by a month. *The weekly aggregation provided the balance of utilities and protection with RMSE of 0.0074*.

When **aggregated by province and infection route**, the resulting count query is presented in Table VI. Disclosure at provincial level presents a lesser risk in term of the RMSE and aggregation results. The resulting aggregation by week provided the balance of utilities and privacy with root mean square error of 0.0604 and good correlation with the original dataset (0.992). The details of the RMSE and Pearson correlation for all the different variation of our temporal hierarchy is presented in Tables VII and VIII.

Finally, we conducted a case study for the data in the capital city of Seoul. We investigated the impact of our solution using the City of Seoul having the highest route of infection.

TABLE V: Aggregation by province, city and infection route

Infection route	Dataset	Day	Week	Month
academy	11	4	5	5
airport	120	73	95	109
bakery	24	3	3	3
bank	28	3	12	14
beauty salon	14	0	5	7
cafe	85	13	48	49
church	120	68	86	101
gas station	12	0	0	3
gym	20	6	11	12
hospital	1496	614	1238	1392
lodging	29	3	14	17
pc cafe	46	0	24	28
pharmacy	200	21	74	132
post office	15	0	4	7
public transportation	382	95	255	307
restaurant	451	98	255	326
school	49	8	22	25
store	507	157	344	397
university	14	0	3	9
others	1698	4155	2823	2378
Total	5321	5321	5321	5321

TABLE VI: Aggregation by province and infection route

Infection route	Dataset	Day	Week	Month
academy	11	4	5	5
airport	120	87	108	116
bakery	24	3	7	16
bank	28	3	14	21
beauty salon	14	0	5	7
cafe	85	33	61	73
church	120	74	98	113
gas station	12	0	3	6
gym	20	6	11	12
hospital	1496	1258	1456	1493
lodging	29	3	17	22
pc cafe	46	10	34	43
pharmacy	200	82	166	188
post office	15	0	5	8
public transportation	382	261	358	372
restaurant	451	273	398	438
school	49	15	32	40
store	507	346	471	493
university	14	0	4	11
others	1698	2863	2068	1844
Total	5321	5321	5321	5321

TABLE VII: Pearson correlation

Aggregation level	Day	Week	Month
Type	0.994935	0.996386	0.999994
Province, type	0.935215	0.991849	0.998804
Province, city, type	0.796876	0.933743	0.972514

TABLE VIII: RMSE

Aggregation level	Day	Week	Month
Type	0.048308	0.007400	0.001209
Province, type	0.231026	0.060400	0.021108
Province, city, type	0.928283	0.231700	0.120472

Classifying infection route with less than 3 trajectories as 'others' route resulted in non-disclosure of some route such as bakery, bank, beauty cafe, gas station, gym, PC cafe (i.e., internet cafe), post office, salon, and school at the daily level of the temporal hierarchy. Beauty salon, gym and post office are reclassified as 'others' at both the weekly and monthly levels of the hierarchy. The resulting count query is presented in Table IX. However, non-disclosure does not necessarily translate to no contact tracing because the city may still trace the contact on those infection routes without disclosing the routes.

TABLE IX: Aggregation effect on Seoul data

Infection route	**Dataset**	**Day**	**Week**	**Month**
airport	18	13	18	18
bakery	6			4
bank	8		5	6
beauty salon	3			
cafe	30	10	24	30
church	37	17	31	37
gas station	3		3	3
gym	3			
hospital	603	582	601	603
lodging	8	3	6	8
PC cafe	19		14	19
pharmacy	53	30	51	53
post office	3			
public transportation	192	166	191	192
restaurant	150	120	146	150
school	5			3
store	146	117	146	146
others	424	653	475	439
Total	1711	1711	1711	1711

V. Conclusions

We presented in this paper a solution for privacy preservation of COVID-19 contact tracing data. Specifically, our solution preserves the privacy of individuals by (a) building a temporal hierarchy, (b) grouping similar data to preserve the actual timestamp and locations, (c) representing each group by their spatio-temporal representative points, and (d) adding Laplace noise the representative points happen to be actual locations. We evaluated our solution by measuring the spatial error with the Harvasine distance and root mean square error (RMSE). We also evaluated the utility of our solution using aggregate queries. The results on real-life COVID-19 contact tracing data collected from mobile devices in South Korea demonstrate the effectiveness and practicality of our solution in preserving the privacy of COVID-19 contact tracing data.

As *ongoing and future work*, we are exploring further enhancements of our solution using the generative adversarial network (GAN) [90]. GAN is a deep leaning model often for generating synthetic data, speech, image or text by training and learning patterns from the input dataset to preserve privacy of individuals. We are also interested in transfer knowledge learned to preserve privacy of other data.

Acknowledgment

This work is partially supported by (a) the Natural Sciences and Engineering Research Council of Canada (NSERC), and (b) the University of Manitoba.

References

[1] X. Chen, et al., "Degree of integration into social networks (DISN) evaluation model based on micro-blogging platforms and information dissemination prediction," IEEE CIT 2016, pp. 172-176.

[2] F. Jiang, et al., "Finding popular friends in social networks," CGC 2012, pp. 501-508.

[3] C.K. Leung, C.L. Carmichael, "Exploring social networks: a frequent pattern visualization approach," IEEE SocialCom 2010, pp. 419-424.

[4] C.K. Leung, et al., "Parallel social network mining for interesting 'following' patterns," CCPE 28(15), 2016, pp. 3994-4012.

[5] K. Zhang, et al., "A core theory based algorithm for influence maximization in social networks," IEEE CIT 2017, pp. 31-36.

[6] A.A. Audu, et al., "An intelligent predictive analytics system for transportation analytics on open data towards the development of a smart city," CISIS 2019, 224-236.

[7] P.P.F. Balbin, et al., "Predictive analytics on open big data for supporting smart transportation services," Procedia Computer Science 176, 2020, pp. 3009-3018.

[8] C.K. Leung, et al., "Urban analytics of big transportation data for supporting smart cities," DaWaK 2019, 24-33.

[9] M.A.M. Marinho, et al., "Antenna array based localization scheme for vehicular networks," IEEE CIT 2017, pp. 142-146.

[10] A.K. Chanda, et al., "A new framework for mining weighted periodic patterns in time series databases," ESWA 79, 2017, pp. 207-224.

[11] C.K. Leung, et al., "Explainable data analytics for disease and healthcare informatics," IDEAS 2021, 12:1-12:10.

[12] J. Souza, et al., "An innovative big data predictive analytics framework over hybrid big data sources with an application for disease analytics," AINA 2020, pp. 669-680.

[13] C.K. Leung, et al., "Predictive analytics on genomic data with high-performance computing," IEEE BIBM 2020, pp. 2187-2194.

[14] T. Pawliszak, et al., "Operon-based approach for the inference of rRNA and tRNA evolutionary histories in bacteria," BMC Genomics 21 (Supplement 2), 2020, pp. 252:1-252:14.

[15] O.A. Sarumi, et al., "Spark-based data analytics of sequence motifs in large omics data," Procedia Computer Science 126, 2018, pp. 596-605.

[16] O.A. Sarumi, C.K. Leung, "Adaptive machine learning algorithm and analytics of big genomic data for gene prediction," Tracking and Preventing Diseases with Artificial Intelligence, 2022, pp. 103-123.

[17] Y. Chen, et al., "Temporal data analytics on COVID-19 data with ubiquitous computing," in IEEE ISPA-BDCloud-SocialCom-SustainCom 2020, pp. 958-965.

[18] P. Gupta, et al., "Vertical data mining from relational data and its application to COVID-19 data," in Big Data Analyses, Services, and Smart Data, 2021, pp. 106-116.

[19] C.S. Eom, et al., "STDP: secure privacy-preserving trajectory data publishing," IEEE Cybermatics 2018, pp. 892-899.

[20] C.K. Leung, et al., "Privacy-preserving healthcare analytics of trajectory data," APWeb-WAIM 2021, Part II, pp. 414-420.

[21] T.S. Cox, et al., "An accurate model for hurricane trajectory prediction," IEEE COMPSAC 2018, vol. 2, 534-539.

[22] X. Wang, et al., "Micro-scale severe weather prediction based on region trajectories extracted from meteorological radar data," IEEE CIT 2016, pp. 335-338.

[23] A. Kobusinska, et al., "Emerging trends, issues and challenges in Internet of Things, big data and cloud computing," FGCS 87, 2018, pp. 416-419.

[24] C.K. Leung, "Big data analysis and mining," Encyclopedia of Information Science and Technology, 4e, 2018, pp. 338-348.

[25] T.A. Shaikh, R. Ali, "Quantum computing in big data analytics: a survey," IEEE CIT 1016, pp. 112-115.

[26] F. Jiang, C.K. Leung, "A data analytic algorithm for managing, querying, and processing uncertain big data in cloud environments," Algorithms 8(4), 2015, 1175-1194.

[27] C.K. Leung, "Uncertain frequent pattern mining," Frequent Pattern Mining, 2014, pp. 417-453.

[28] C.K. Leung, et al., "Fast algorithms for frequent itemset mining from uncertain data," IEEE ICDM 2014, 893-898.

[29] C.K. Leung, F. Jiang, "Frequent pattern mining from time-fading streams of uncertain data," DaWaK 2011, pp. 252-264.

[30] N.A. Othman, et al., "Enhancing aggregation over uncertain databases," IEEE CIT-IUCC-DASC-PICom 2015, pp. 132-139.

[31] C. Zhao, et al., "Analyzing COVID-19 epidemiological data," IEEE DASC-PICom-CBDCom-CyberSciTech 2021, pp. 985-990.

[32] C.K. Leung, et al., "A digital health system for disease analytics," IEEE ICDH 2021, pp. 70-79.

[33] C.K. Leung, et al., "Health analytics on COVID-19 data with few-shot learning," DaWaK 2021, pp. 67-80.

[34] D.L.X. Fung, et al., "Self-supervised deep learning model for COVID-19 lung CT image segmentation highlighting putative causal relationship among age, underlying disease and COVID-19," BMC Journal of Translational Medicine 19, 2021, pp. 318:1-318:18.

[35] Q. Liu, et al., "A two-dimensional sparse matrix profile DenseNet for COVID-19 diagnosis using chest CT images," IEEE Access 8, 2020, pp. 213718-213728.

[36] Y. Chen, et al., "A data science solution for supporting social and economic analysis," IEEE COMPSAC 2021, pp. 1689-1694.

[37] C.K. Leung, et al., "Data mining on open public transit data for transportation analytics during pre-COVID-19 era and COVID-19 era," INCoS 2020, pp. 133-144.

[38] K.E. Dierckens, et al., "A data science and engineering solution for fast k-means clustering of big data," IEEE TrustCom-BigDataSE-ICESS 2017, pp. 925-932.

[39] C.K. Leung, F. Jiang, "A data science solution for mining interesting patterns from uncertain big data," IEEE BDCloud 2014, pp. 235-242.

[40] P. Braun, et al., "Game data mining: clustering and visualization of online game data in cyber-physical worlds," Procedia Computer Science 112, 2017, pp. 2259-2268.

[41] P. Braun, et al., "Pattern mining from big IoT data with fog computing: models, issues, and research perspectives," IEEE/ACM CCGrid 2019, pp. 854-891.

[42] A. Fariha, et al., "Mining frequent patterns from human interactions in meetings using directed acyclic graphs," PAKDD 2013, Part I, pp. 38-49.

[43] L.V.S. Lakshmanan, et al., "The segment support map: scalable mining of frequent itemsets," ACM SIGKDD Explorations 2(2), 2000, pp. 21-27.

[44] C.K. Leung, "Frequent itemset mining with constraints," Encyclopedia of Database Systems, 2e, 2018, 1531-1536.

[45] C.K. Leung, C.L. Carmichael, "FpViz: a visualizer for frequent pattern mining," ACM KDD-VAKD 2009, pp. 30-39.

[46] J. Liu, et al., "Efficient mining of extraordinary patterns by pruning and predicting," ESWA 125, 2019, pp. 55-68.

[47] J. Ruohonen, V. Leppänen, "Investigating the agility bias in DNS graph mining," IEEE CIT 2017, pp. 253-260.

[48] S. Ahn, et al., "A fuzzy logic based machine learning tool for supporting big data business analytics in complex artificial intelligence environments," FUZZ-IEEE 2019, pp. 1259-1264.

[49] J. de Guia, et al., "DeepGx: deep learning using gene expression for cancer classification," IEEE/ACM ASONAM 2019, pp. 913-920.

[50] C.K. Leung, et al., "A machine learning approach for stock price prediction," in IDEAS 2014, pp. 274-277.

[51] K.J. Morris, et al., "Token-based adaptive time-series prediction by ensembling linear and non-linear estimators: a machine learning approach for predictive analytics on big stock data," IEEE ICMLA 2018, pp. 1486-1491.

[52] C.K. Leung, Mathematical model for propagation of influence in a social network, in Encyclopedia of Social Network Analysis and Mining, 2e, 2018, pp. 1261-1269.

[53] S. Smrithy, R. Balakrishnan, "A statistical technique for online anomaly detection for big data streams in cloud collaborative environment," IEEE CIT 2016, 108-111.

[54] W. Lee, et al., "Reducing noises for recall-oriented patent retrieval," IEEE BDCloud 2014, pp. 579-586.

[55] C.K. Leung, et al., "Information technology-based patent retrieval model," Springer Handbook of Science and Technology Indicators, 2019, pp. 859-874.

[56] P. Braun, et al., "MapReduce-based complex big data analytics over uncertain and imprecise social networks,"

DaWaK 2017, pp. 130-145.

[57] R.C. Camara, et al., “Fuzzy logic-based data analytics on predicting the effect of hurricanes on the stock market,” FUZZ-IEEE 2018, pp. 576-583.

[58] D. Deng, et al., “An innovative framework for supporting cognitive-based big data analytics for frequent pattern mining,” IEEE ICCC 2018, pp. 49-56.

[59] K. Hoang, et al., “Cognitive and predictive analytics on big open data,” ICCC 2020, pp. 88-104.

[60] C.K. Leung, F. Jiang, “Big data analytics of social networks for the discovery of ”following” patterns,” DaWaK 2015, pp. 123-135.

[61] M. Mai, et al., “Big data analytics of Twitter data and its application for physician assistants: who is talking about your profession in Twitter?” Data Management and Analysis, 2020, pp. 17-32.

[62] K.E. Barkwell, et al., “Big data visualisation and visual analytics for music data mining,” IV 2018, pp. 235-240.

[63] C.K. Leung, C.L. Carmichael, “FpVAT: a visual analytic tool for supporting frequent pattern pattern mining,” ACM SIGKDD Explorations 11(2), 2009, pp. 39-48.

[64] C.K. Leung, et al., “Visual analytics of social networks: mining and visualizing co-authorship networks,” HCII-FAC 2011, pp. 335-345.

[65] C.K. Leung, F. Jiang, “RadialViz: an orientation-free frequent pattern visualizer,” PAKDD 2012, Part II, pp. 322-334.

[66] K. Huang, et al., “EBD-MLE: enabling block dynamics under BL-MLE for ubiquitous data,” IEEE ISPA-IUCC 2017, pp. 1281-1288.

[67] E. Serrano, et al., “A cloud environment for ubiquitous medical image reconstruction,” IEEE ISPA-IUCC-BDCloud-SocialCom-SustainCom 2018, pp. 1048-1055.

[68] M. Badawy, et al., “Verification in mobile communication during the change of IP address,” IEEE IUCC-DSCI-SmartCNS 2019, pp. 101-105.

[69] X. Cheng, et al., “Post-evaluation model of telecommunication network construction based on AHP,” IEEE IUCC-DSCI-SmartCNS 2019, pp. 616-620.

[70] N. Ahmed, et al., “Survey of COVID-19 contact tracing apps,” IEEE Access 8, 2020, pp. 134577-134601.

[71] R. Raskar, et al., “Contact tracing: holistic solution beyond bluetooth,” IEEE Data Eng. Bull. 43(2), 2020, pp. 67-70.

[72] N. Trieu, et al., “Epione: lightweight contact tracing with strong privacy. IEEE Data Eng. Bull. 43(2), 2020, pp. 95-107.

[73] K. LeFevre, et al., “Incognito: efficient full-domain k-anonymity,” ACM SIGMOD 2005, pp. 49-60.

[74] N. Li, et al., “t-closeness: privacy beyond k-anonymity and l-diversity,” IEEE ICDE 2007, pp. 106-115.

[75] A. Machanavajjhala, et al., “l-diversity: privacy beyond k-anonymity,” ACM TKDD 1(1), 2007, pp. 3:1-3:52.

[76] G. Acs, C. Castelluccia, “A case study: privacy preserving release of spatio-temporal density in Paris,” in ACM KDD 2014, pp. 1679-1688.

[77] M.E. Andrés, et al., “Geoindistinguishability: differential privacy for location-based systems,” in ACM CCS 2013, pp. 901-914.

[78] Y. Cao, et al., “Quantifying differential privacy under temporal correlations,” IEEE ICDE 2017, pp. 821-832.

[79] Y. Xiao, L. Xiong, “Protecting locations with differential privacy under temporal correlations,” ACM CCS 2015, pp. 1298-1309.

[80] F. Jiang, et al., “Web page recommendation based on bitwise frequent pattern mining,” IEEE/WIC/ACM WI 2016, pp. 632-635.

[81] C.K Leung, et al., “Scalable vertical mining for big data analytics of frequent itemsets,” DEXA 2018, pp. 3-17.

[82] C.S. Eom, et al., “Effective privacy preserving data publishing by vectorization,” Information Sciences 527, 2020, pp. 311-328.

[83] R. Tojiboev, et al., “Adding noise trajectory for providing privacy in data publishing by vectorization,” IEEE BigComp 2020, pp. 432-434.

[84] C.K. Leung, et al., “Privacy-preserving frequent pattern mining from big uncertain data,” IEEE BigData 2018, pp. 5101-5110.

[85] B.H. Wodi, et al., “Fast privacy-preserving keyword search on encrypted outsourced data,” IEEE BigData 2019, pp. 6266-6275.

[86] N. Wang, et al., “Privsuper: a superset-first approach to frequent itemset mining under differential privacy,” IEEE ICDE 2017, pp. 809-820.

[87] Z. Huo, et al., “You can walk alone: trajectory privacy-preserving through significant stays protection,” DASFAA 2012, pp. 351-366.

[88] C. Dwork, “Differential privacy,” ICALP 2006, pp. 1-12.

[89] H. Lee, et al., “De-identification and privacy issues on bigdata transformation,” IEEE BigComp 2020, 514-519.

[90] A. Creswell, et al., “Generative adversarial networks: an overview,” IEEE Signal Processing Magazine, 35(1), 2018, pp. 53-65.