Supporting Privacy-Preserving Big Data Analytics on Temporal Open Big Data

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 198 (2022) 112–121

Procedia Computer Science

www.elsevier.com/locate/procedia

The 12th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2021)
November 1-4, 2021, Leuven, Belgium

Supporting Privacy-Preserving Big Data Analytics on Temporal Open Big Data

Alfredo Cuzzocrea[a,*], Carson K. Leung[b], Anifat M. Olawoyin[b], Edoardo Fadda[c]

[a] *iDEA Lab, University of Calabria, Rende, Italy & LORIA, Nancy, France*
[b] *Department of Computer Science, University of Manitoba, Winnipeg, MB, Canada*
[c] *DAUIN, Politecnico di Torino & ISIRES, Torino, Italy*

Abstract

Nowadays, valuable big data are generated and collected rapidly from numerous rich data sources. Following the initiatives of open data, many organizations including municipal governments are willing to share their data such as open big data regarding parking violations. While there have been models to preserve privacy of sensitive personal data like patient data for health informatics, privacy of individuals who violated parking regulations should also be protected. Hence, in this article, we present a model for supporting privacy-preserving big data analytics on temporal open big data. This temporally hierarchical privacy-preserving model (THPPM) adapts and extends the traditional temporal hierarchy to generalize spatial data generated within a time interval with an aim to preserve privacy of individuals who violated parking regulations during some time intervals at certain geographic locations. Evaluation on open big data from two North American cities demonstrates the usefulness of our model in supporting privacy-preserving big data analytics on temporal open big data.

Keywords: Big Data; Privacy; Privacy-Preserving Data Mining; Spatio-Temporal Data; Temporal Hierarchy

1. Introduction

In the current technological era, big data are everywhere. From a wide variety of rich data sources huge amounts of data have been produced. Different levels of veracity (e.g., precise data, imprecise and uncertain data [1, 25, 36]) characterized these big data. Examples of big data include financial time series [8, 33, 45], social network data [23, 26, 32, 37, 41], transportation data [6, 2, 34, 35, 40], omic (e.g., genomic) data [3, 44], as well as disease reports

* Corresponding author. Tel.: +39-0984-492501
E-mail address: alfredo.cuzzocrea@unical.it

1877-0509
10.1016/j.procs.2021.12.217

[21, 46, 48], epidemiological data and statistics. Big data management and mining provide the methods to deal with these data [10, 22, 11, 12, 17].

Open data are becoming more and more popular in the recent years. Consequently, more big data are openly available on open data platforms or portals. Transparency and enabling a better understanding of the services provided by government to its citizens and/or by organizations to their service recipients are two of the advantages provided by the accessibility of open data. In open data, sensitive attributes that may uniquely identify individuals within the published records (e.g., such as names and identification numbers, social insurance number and social security number) are usually de-identified. This leads to the demand for privacy-preserving data publishing.

The privacy-utility trade off is a major setback in privacy-preserving model. In fact, techniques used utilized by most syntactic models such as suppression of attributes and/or generalization may not prevent re-identification of individual records [30, 42, 43].

In particular, temporal and spatial attributes may be joined with other data from heterogeneous sources to re-identify the individual records. For instance, it is possible to re-identification of individual owners of the parking tickets by considering a combination of the plotted spatial data of parking ticket and data from other different sources over the internet. Nevertheless, in some application the volumes of records may big enough to make individual records too specific for providing meaningful information related to the big data. Aggregated data may provide generalization of specific data for more meaningful knowledge for big data analytics.

Some noise is added to preserve data privacy in techniques such as differential privacy models [19, 5]. This may alter the utility of the data and in extreme case lead them to be useless. To keep a good balance of data privacy and utility, we present *temporal hierarchy privacy-preserving model* (*THPPM*) in this paper. Specifically, the privacy of temporal big data is preserved by the THPPM while maintains utility of the data. It is worth noting that despite the presented use case on parking ticket data, the same model can be used in all the real-life applications and services needing to preserve privacy of temporal big data. As an example, consider activities of an individual for a particular day, which starts with a visit to a pharmaceutical store, a retail store, and a health insurance agency; then parking; and ends the day with airline booking. By using temporal and/or spatial attributes from each activity analysts can link together these disjointed activities. Nevertheless, in several cases the individual records does not provide useful information. A pharmacist and a store manager, for example, are more interested in the shopper purchase patterns (e.g., for inventory purposes). Similarly, booking patterns are the main interest of travel agencies (for promotional purposes of some tourist attractions and/or tourist activities). Thus, privacy preserving of temporal big data is the main topic of the present work. In particular, we focus on preserving the association of individual to a record by using temporal hierarchy. Enabling aggregated ad-hoc queries over anonymized dataset without revealing the underlying associations to any individual is the mian characteristic of the proposed THPPM. The contributions of this paper are several. First, we describe the THPPM model. Second, we also extend the concept of temporal hierarchy to other areas of data publishing. Third, we show that by using the proposed methodology, it is possible to maintain data utility while preserving the privacy of individual entity within the dataset. Finally, the THPPM is extendable to other forms of generalization with and without the spatial attribute. For instance, we combine the temporal hierarchy model with location generalization for the city of Toronto dataset. The temporal aggregation of activities by using the frequency count query is the basic idea of the framework:

SELECT COUNT(∗)
FROM *dataset name*
GROUP BY *temporal aggregation level*

The paper is organised as follows: in Section 2 we provide background and related works. We presents the THPPM model in Section 3. Finally, Section 4 describes the experimental evaluation and Section 5 draws the conclusions.

2. Background and Related Work

Over the past decades, space and time efficiency are the motivation for many data mining algorithms such as the efficient frequent itemset mining [31], bitwise pattern mining [27], and scalable vertical mining [39, 28]. Nowadays, anonymization is included in data mining as a step within the data mining tasks to ensure the outcome is privately preserved. For instance, in [20] the authors propose a surrogate vectors and length-based frequent pattern tree (LFP-

tree) for anonymizing data; instead in [47] authors add noise to trajectory preserving privacy dealing with uncertain data mining while [38] and [50] study the keyword search on encrypted outsourced data. They combined LFP-tree and surrogate vectors models to publish anonymized trajectory database. In contrast, our approach is not using surrogate vectors or LFP-tree but it make good usage of temporal hierarchy to preserve privacy of temporal big data.

As another instance, Wang et al. [49] utilized sequence exponential mechanism (SEM) to extend maximal frequent itemsets mining on sensitive data. In [24] a generalization of stay of points to preserve privacy of trajectories is presented. The *differential privacy* is proposed in [19]; its main idea is ensuring that the probability outcome of any aggregate function by adding or removing individual record from the original database must not have significant differences.

Lee et al. [29] discussed current issues and solutions to manage and control the privacy level (e.g., k-anonymity [18], l-diversity [43], t-closeness [42], differential privacy) in big data de-identification, with focus on preserving privacy of health related applications. Instead, parking applications are considered in this paper.

3. An Innovative Temporal Hierarchy Privacy-Preserving Model

In this section, we present our *temporal hierarchy privacy-preserving model* (*THPPM*). In particular, the key concept behind the THPPM are presented; they are: temporal hierarchy (with extension to time generalization), temporal representative points, and location generalization. By using temporal hierarchy, THPPM is able to preserve privacy of temporal data of fine granularity while maintaining the information of the data by aggregating temporal data to an appropriate temporal hierarchical level. Moreover, temporal representative points enable the THPPM to represent a collection of points within a specific level of temporal hierarchy. Moreover, temporal representative points are adapted to spatial data and representative points over spatial data to generalize spatial locations are computed.

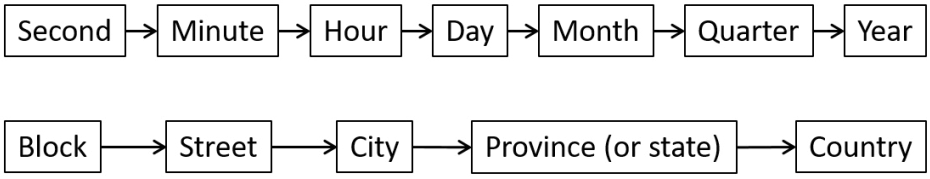

Fig. 1: Architecture for our THPPM, with extended temporal hierarchy for time generalization (up), and extended hierarchy for spatial data for location generalization (down)

3.1. Temporal Hierarchy and Time Generalization

Temporal hierarchy is the aggregation of time series data to a predefined periodic level. For example, we can aggregate daily information d_k to obtain monthly information m_j which can be further be aggregated to get quarterly information q_i. Finally, the quarterly information q_i can be aggregated to obtain yearly information y, and so on with the various level of time hierarchy (see Fig. 1). Mathematically, temporal hierarchy can be represented as:

$$y = \sum_{i=1}^{4} d_i \tag{1}$$

$$q_i = \sum_{j=1}^{3} m_j \tag{2}$$

$$m_j = \sum_{k=1}^{d} d_k \tag{3}$$

where d is the number of days within a month (e.g., $d \in \{28, 29, 30, 31\}$). In order to preserve privacy of health related real-life applications, temporal hierarchy from day up to year (i.e., *date generalization*) are effective. Nevertheless,

details of data are need to a much finer granularity (e.g., with timestamp) for non-health related applications such as parking. Thus, we extend our temporal hierarchy to *time generalization* for each day. The extended temporal hierarchy for time generalization is shown in Fig. 1. There, both daytime *dt* and nighttime *nt* are aggregated to obtain daily information d_k. In particular aggregating 12 hourly information hh_i from daytime (say, 06:00-18:00) we compute *dt* while aggregating the remaining 12 hourly information hh_i from nighttime (say, 18:00-06:00) we compute *nt*. By following this direction, we compute hourly information h_i by aggregating information from its 60 minutes mm_j, each of which is obtained by aggregating information from its 60 seconds ss_k. Mathematically, this extended temporal hierarchy can be represented as:

$$d_k = dt + nt \tag{4}$$

$$dt \text{ (and } nt) = \sum_{i=1}^{12} hh_i \tag{5}$$

$$hh_i = \sum_{j=1}^{60} mm_j \tag{6}$$

$$mm_j = \sum_{k=1}^{60} ss_k. \tag{7}$$

In order to preserve the privacy of individual record within the dataset, the detailed time of the events is generalized to the next level in the hierarchy.

3.2. Temporal Representative Points

A *temporal representative point* is a point representing all points (a line, multipoints or polygon) within a specific level of temporal hierarchy. We define weighted and unweighted temporal representative point. In particular, a *weighted representative point* $R(x, y)$ is computed by using weighted mean centre of the n coordinates x- and y-coordinates; when all the weights are equal we obtain the sample means of the set of points as in Eq. (8).

$$R(x, y) = \left(\frac{\sum_{i=1}^{n} x_i}{n}, \frac{\sum_{i=1}^{n} y_i}{n} \right) \tag{8}$$

A weighted representative can also be computed using the minimum square distance within existing points. Eq. (9) recall the minimum square distance $Dist_{min}$.

$$Dist_{min} = \min_{j=1,\ldots,n} \sum_{i=1}^{n} \left[\left(x_j - x_i \right)^2 + \left(y_j - y_i \right)^2 \right] \tag{9}$$

From a geometric point of view, a weighted representative point is a centroid if it is the centre of gravity of the geographical region considered regardless of whether the representative point is inside the region or not.

An *unweighted representative point* is a loosely computed point that optimally represents geographical points, polygon or multipoints. In contrast to weighted representative point, it guarantee to be inside the considered region. Usually, It is chosen as a randomly selected point among the one belonging to the region.

3.3. Location Generalization

Temporal representative point (weighted or unweighted) can be computed over spatial data. However, using the proposed THPPM, other location attributes such as street, blocks or city, can be generalized and aggregate. Fig. 1

depict a typical location generalization. A country is depicted as an hierarchy of locations: blocks, street, city, and state/province. For instance, the attributes attributes census block (which is is the smallest geographic unit used by the United States Census Bureau) and street information are the ones of the parking contravention dataset of the city of Buffalo. The raw dataset has several other location information. By using the data, the reader can observe that for a parking violation occurred on Washington Street located at (42.891484°N, 78.871482°W) within the US census block 1006 the summons number A5508230 was issued at 01:04 on January 01, 2007.

3.4. Temporal Hierarchy Model

After describing key components of our THPPM, let us non-trivially integrating them to form our model. Given a dataset D (e.g., parking dataset), first a temporal hierarchy for date generalization is created, then it is extended temporal hierarchy for time generalization and a extended hierarchy for spatial data for location generalization is created. A general representation of the architecture econmpassing these three key components in the temporal hierarchy is provided in Fig. 1. For each item (e.g., each parking violation), our THPPM generalizes date, time and location for preserving privacy of individuals (e.g., parking violators). Then, the frequency of the generalized data, time and location to report the aggregated information for big data analytics is counted by the THPPM. By doing so, the proposed THPPM preserves privacy of temporal big data while maintaining the sufficient information for the analytics. The pseudocode of the model model can be found in Algorithm 1.

Algorithm 1: Temporal Hierarchy and Location Generalization

```
Input: Dataset B
Result: Privacy-preserving dataset B'
Date = create temporal hierarchy (for date generalization);
Time = create time generalization ;
Location = create location generalization ;
foreach temporal hierarchy do
    foreach item in temporal hierarchy in B do
        select location generalization;
        group by Time, Location;
        compute frequency count;
    end
end
```

4. Experimental Assessment and Analysis

4.1. Setup: Development Environment and Tools

For evaluation, we implemented our THPPM in Python on Spyder scientific development environment. A laptop computer with 2.6 GHz Intel(R) core^{TM} i7 64-bit operating system and 8 GB installed memory has been used for the experimental evaluations.

4.2. Datasets

We conducted our experimental evaluation by using two real-world datasets for our evaluations. The first dataset considers non-identifiable information relating to parking tickets for the city of Toronto in the Canadian province of Ontario [1]. It consists of 1.47 GB dataset containing 11,096,751 violation records registered in the period from January 2014 to December 2018 by the Toronto Police Services as well as by people certified and authorized to issue tickets. Table 1 provides some detailed descriptions of the dataset, enabling the readers to interpret data about infractions.

[1] https://open.toronto.ca/dataset/parking-tickets/

The second dataset records parking tickets for the city of Buffalo in the New York State, USA [2]. This 413 MB dataset contains 2,283,838 parking summonses for a 13-year period from January 2007 to December 2019, issued by the Division of Parking Enforcement, the Buffalo Police Department, and other agencies permitted to write summonses on behalf of the City of Buffalo.

Fig. 1 represent the temporal hierarchy, together with its time generalization and location generalization. The two datasets are preprocessed as follows. Each data is aggregated with respect to daytime and nighttime levels capturing tickets issued from the 12-hour periods of 06:00-18:00 and 18:00-06:00, respectively. Since the census block is a geographic unit used only in the US, while census blocks are used for the Buffalo parking tickets, street are used for the Toronto parking. An aggregation of address in group of hundred addresses is create in order to mimic the census blocks.

Table 1: Record in dataset

Hierarchy Level	**Experiment**		
	1	***2***	***3***
Year	5735185	5364835	4232982
Quarter	2094890	1748291	1214532
Month	1364165	1111379	760382
Day	794882	632295	426096

4.3. Evaluation Results

For evaluation of our THPPM, we conducted three sets of experiments. In the first experiment, the location details are kept constant, generalized the time component, and varied the temporal hierarchy. The corresponding SQL query is

```
SELECT province, street, block, time, (S1|S2|S3|S4), COUNT(*)
FROM ...
WHERE ...
GROUP BY province, street, block, time (S1|S2|S3|S4)
```

where S1, S2, S3 and S4 are the temporal hierarchy level representing day, month, quarter and year, respectively. In the second experiment, the location generalization is combined with a temporal hierarchy *without* the time generalization:

```
SELECT province, street, block, (S1|S2|S3|S4), COUNT(*)
FROM ...
WHERE ...
GROUP BY province, street, block, (S1|S2|S3|S4)
```

Finally, in the third experiment, we used only the street for location generalization and varied the temporal hierarchy. In the three sets of aforementioned experiments, we measure the impact of variation in hierarchy level. For the various hierarchical level, the reduction in number of record in the dataset is measured at each level and the uniqueness percentage for each level is computed.

For the Toronto parking dataset, the number of records at each temporal hierarchy for Experiment 1 is presented in Fig. 2. By moving up in the hierarchy level, the number of records in the aggregated dataset decreases. In particular, there are 5,735,185 aggregated daily observation, that becomes 2,094,890 if aggregated monthly, 1,364,165 if aggregated quarterly and finally, 794,882 if yearly aggregated.

As shown in Fig. 2, the number further reduces in Experiment 2 because of the suppression of the time generalization. The number of aggregated data is further reduced by suppressing the time generalization (i.e., not distinguishing

[2] https://data.buffalony.gov/Transportation/Parking-Summonses/yvvn-sykd

the daytime and nighttime data). In particular, we have 5,364,835 aggregated daily counts, that can be reduced to 1,748,291 aggregated monthly counts, to 1,111,379 aggregated quarterly counts, and finally to 632,295 aggregated yearly counts. This number of aggregated yearly counts is less than 80% of that with time generalization. In the third experiment, the number of aggregated is further reduced by using the location generalization to street level; the results are shown in Fig. 2. Specifically, it drops to 4,232,982 aggregated daily counts, then to 1,214,532 aggregated monthly counts, further 760,382 aggregated quarterly counts, and finally to 426,096 aggregated yearly counts of infractions in Toronto. Thus, a reduction in the number of aggregated of more than the 32% and of more than the 46% with respect to the first and second experiments can be observed. This is a major outcomes of the benefits coming from the location generalization. These aggregated counts for the three experiments are summarized in Fig. 2.

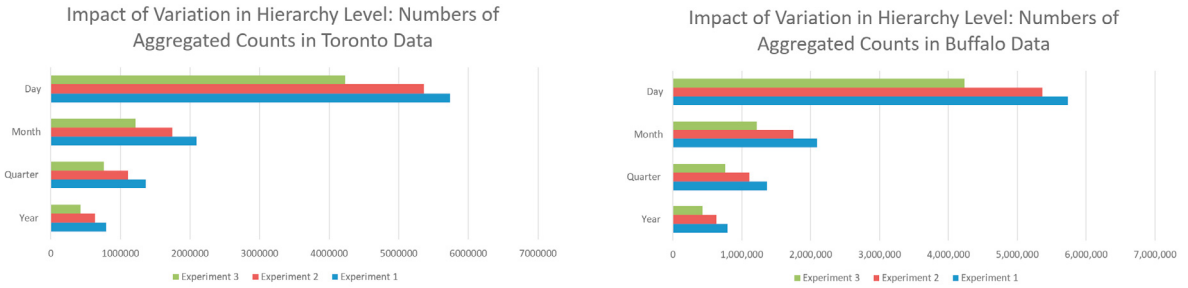

Fig. 2: Impact of Variation in Hierarchy Level: Numbers of Aggregated Counts in Toronto Data (left) and Buffalo Data (right)

For Buffalo parking data, Table 1 shows the reduction in the number of record for each experiment. As aforementioned, the number of records decreases as we consider higher hierarchical levels. Specifically, in the first experiment with time generalization, the number of aggregated data drops from 1,519,895 aggregated daily counts to 899,988 aggregated monthly counts, then to 646,959 aggregated quarterly counts, and finally to 411,032 aggregated yearly counts when we move up the temporal hierarchy level for day to month to quarter and finally to year. Similarly, in the second experiment is shown that, without time generalization, the number of aggregated data decreases further to 1,501,133 aggregated daily counts, to 839,468 aggregated monthly counts, to 579,298 aggregated quarterly counts, and to 347,334 aggregated yearly counts. Furthermore, the third experiment reports that applying the location generalization, the number of aggregated data further decreases to 1,117,809 aggregated daily counts, to 442,392 aggregated monthly counts, to 279,799 aggregated quarterly counts, and to 152,305 aggregated yearly counts. This number of aggregated yearly counts is less than 44% and less than 37% of those in Experiment 2 and Experiment 1, respectively, which both without location generalization. Experiment 3 again shows the benefits of location generalization. Fig. 2 summarizes these quantities.

The unique record percentage is also measured. Note that the unique record is the frequency count of record where the count of ticket remains one despite aggregation, which can be obtained by running a SQL query:

COUNT(*) ... HAVING COUNT(ticket) < 2

The unique record percentage drops decrease fom 30% to 1.8% for the Toronto dataset when varying the hierarchy level, as shown in Fig. 3. Instead for the Buffalo dataset, Fig. 3 reports that the unique percentage reduced from 48% in the lowest level of temporal hierarchy to 3% for the year level.

As our third evaluation metrics, we measured data compression ratio. The data compression ratio *DCR* is computed as the ratio of the anonymized dataset B' to the original dataset B at a specific temporal hierarchy:

$$DCR(B, B') = \frac{B'}{B} \tag{10}$$

Fig. 4 shows that by aggregating to the finer temporal hierarchy resulted in preserving privacy of individual record and at least half of data compression.

Fig. 3: Unique record percentage in Toronto Data (left) and Buffalo Data (right)

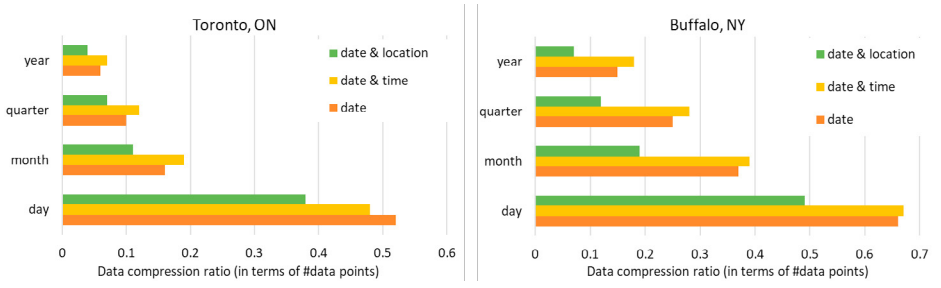

Fig. 4: Data compression ratio in Toronto Data (left) and Buffalo Data (right)

Moreover, for each temporal hierarchy level to measure the utility of the anonymized published dataset count queries is measured by performing series of count queries (histogram) over the aggregated dataset. It is worth noting that the results of the aggregation is the same over the protected and over the original dataset since temporal hierarchy is a bottom-up aggregation of record. The query execution time for each experiment at various levels of the hierarchy is presented in Table 2 and Table 3. The computed query result for the lowest level of the temporal hierarchy is used such that the original dataset is read only once. The remaining queries are computed from the lower level of the temporal hierarchy, this reduces the query execution time.

Table 2: Query execution time: Toronto data

Hierarchy Level	Time[s]		
	Date	**Date and Time Hierarchy**	**Location**
Year	2.72%	1.53	0.83
Quarter	4.25	4.33	1.41
Month	4.9	4.4	1.5
Day	8.57	5.91	4.2

Table 3: Query execution time: Buffalo data

Hierarchy Level	Time[s]		
	Date	**Date and Time Hierarchy**	**Location**
Year	2.13	1.62	1.41
Quarter	2.11	1.58	1.31
Month	2.14	2.03	1.31
Day	3.41	3.43	2.73

5. Conclusions and Future Work

Nowadays, valuable big data are generated and collected rapidly from numerous rich data sources. Following the initiatives of open data, many organizations including municipal governments are willing to share their data such as open big data regarding parking violations. While there have been models to preserve privacy of sensitive personal data like patient data for health informatics, privacy of individuals who violated parking regulations should also be protected. Hence, in this article, we presented a model for supporting privacy-preserving big data analytics on temporal open big data. To preserve privacy of individuals who violated parking regulations during some time intervals at certain geographic locations, our temporally hierarchical privacy-preserving model (THPPM) adapts and extends the traditional hierarchy (which generalizes the day dimension) to generalize spatial data generated within a time interval (i.e., to generalize both the time dimension and location dimension) into temporal representative points. Evaluation on open big data from (a) City of Toronto (in province of Ontario, Canada) and (b) City of Buffalo (in New York state, USA) demonstrated the usefulness of our model in supporting privacy-preserving big data analytics on temporal open big data.

As *ongoing and future work*, we exploit and transfer the learned knowledge from the current work to support privacy-preserving big data analytics on other open big data, for instance by exploiting OLAP methodologies (e.g, [16, 15, 9, 13, 14, 7]). Moreover, we incorporate relevant techniques into the model to further enhance results of our privacy-preserving big data analytics, for instance following recent and correlated approaches (e.g., [4]).

6. Acknowledgments

This research has been made in the context of the Excellence Chair in Computer Engineering — Big Data Management and Analytics.

This research is partially supported by (1) the French PIA project “Lorraine Université d’Excellence” (reference ANR-15-IDEX-04-LUE), (2) the Natural Sciences and Engineering Research Council of Canada (NSERC), (3) the University of Manitoba.

References

[1] A. Alim, et al., “Uncertainty-aware opinion inference under adversarial attacks,” in IEEE BigData 2019, pp. 6-15.
[2] 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.
[3] D. Barh, et al., “Multi-omics-based identification of SARS-CoV-2 infection biology and candidate drugs against COVID-19,” Comput. Biol. Medicine 126, 2020, pp. 104051:1-104051:13.
[4] A. Campan, A. Cuzzocrea, T.M. Truta, “Fighting fake news spread in online social networks: Actual trends and future research directions,” in 2017 IEEE Big Data, pp. 4453-4457.
[5] Y. Cao, et al., “Quantifying differential privacy under temporal correlations,” in IEEE ICDE 2017, pp. 821-832.
[6] P. Castrogiovanni, E. Fadda, G. Perboli, A. Rizzo, “Smartphone Data Classification Technique for Detecting the Usage of Public or Private Transportation Modes,” IEEE Access 8, 2020, pp. 58377–58391.
[7] M. Ceci, A. Cuzzocrea, D. Malerba, “Effectively and efficiently supporting roll-up and drill-down OLAP operations over continuous dimensions via hierarchical clustering,” Journal of Intelligent Information Systems 44(3), 2015, pp. 309-333.
[8] A.K. Chanda, et al., “A new framework for mining weighted periodic patterns in time series databases,” ESWA 79, 2017, pp. 207-224.
[9] A. Cuzzocrea, “Improving range-sum query evaluation on data cubes via polynomial approximation,” Data & Knowledge Engineering 56(2), 2006, pp. 85-121.
[10] A. Cuzzocrea, “Big data provenance: state-of-the-art analysis and emerging research challenges,” in EDBT/ICDT Workshops 2016, pp. 37:1-37:4.
[11] A. Cuzzocrea, E. Bertino, “A secure multiparty computation privacy preserving OLAP framework over distributed XML data,” in ACM SAC 2010, pp. 1666-1673.
[12] A. Cuzzocrea, S.L. Francis, M.M. Gaber, “An information-theoretic approach for setting the optimal number of decision trees in random forests,” in IEEE SMC 2013, pp. 1013-1019.
[13] A. Cuzzocrea, U. Matrangolo, “Analytical synopses for approximate query answering in OLAP environments,” in DEXA 2004, pp. 359-370.
[14] A. Cuzzocrea, R. Moussa, G. Xu, “OLAP*: effectively and efficiently supporting parallel OLAP over big data,” in MEDI 2013, pp. 38-49.
[15] A. Cuzzocrea, D. Saccà, P. Serafino, “A hierarchy-driven compression technique for advanced OLAP visualization of multidimensional data cubes,” in DaWaK 2006, pp. 106-119.
[16] A. Cuzzocrea, P. Serafino, “LCS-Hist: taming massive high-dimensional data cube compression,” in EDBT 2009, pp. 768-779.

[17] A. Cuzzocrea, I.-Y. Song, "Big graph analytics: The state of the art and future research agenda," in DOLAP 2014, pp. 99-101.
[18] J. Domingo-Ferrer, "*k*-anonymity," in Encyclopedia of Database Systems, p. 1585, 2009.
[19] C. Dwork, "Differential privacy, in automata, languages and programming," in ICALP 2006, pp. 1-12.
[20] C.S. Eom, et al., "Effective privacy preserving data publishing by vectorization," Information Sciences 527, 2020, pp. 311-328.
[21] P. Gupta, et al., "Vertical data mining from relational data and its application to COVID-19 data," Big Data Analyses, Services, and Smart Data, 2021, pp. 106-116.
[22] M. Hassani, P. Spaus, A. Cuzzocrea, T. Seidl, "Adaptive stream clustering using incremental graph maintenance," in BigMine 2015, pp. 49-64.
[23] C. He, et al., "Finding mutual X at WeChat-scale social network in ten minitues," in IEEE BigData 2019, pp. 288-297.
[24] Z. Huo, et al., "You can walk alone: trajectory privacy-preserving through significant stays protection," in DASFAA 2012, Part I, pp. 351-366.
[25] F. Jiang, C.K. Leung, "A data analytic algorithm for managing, querying, and processing uncertain big data in cloud environments," Algorithms 8(4), 2015, pp. 1175-1194.
[26] F. Jiang, et al., "Finding popular friends in social networks," in CGC 2012, pp. 501-508.
[27] F. Jiang, et al., "Web page recommendation based on bitwise frequent pattern mining," in IEEE/WIC/ACM WI 2016, pp. 632-635.
[28] L.V.S. Lakshmanan, et al., "The segment support map: Scalable mining of frequent itemsets," ACM SIGKDD Explorations 2(2), 2000, pp. 21-27.
[29] H. Lee, et al., "De-identification and privacy issues on bigdata transformation," in IEEE BigComp 2020, pp. 514-519.
[30] K. LeFevre, et al., "Incognito: efficient full-domain *k*-anonymity," in ACM SIGMOD 2005, pp. 49-60.
[31] C.K. Leung, "Frequent itemset mining with constraints," Encyclopedia of Database Systems, 2e, 2018, pp. 1531-1536.
[32] C.K. Leung, C.L. Carmichael, "Exploring social networks: a frequent pattern visualization approach," in IEEE SocialCom 2010, pp. 419-424.
[33] C.K. Leung, et al., "A machine learning approach for stock price prediction," in IDEAS 2014, pp. 274-277.
[34] C.K. Leung, J.D. Elias, S.M. Minuk, A. Roy R. de Jesus, A. Cuzzocrea, "An innovative fuzzy logic-based machine learning algorithm for supporting predictive analytics on big transportation data," in FUZZ-IEEE 2020, pp. 1-8.
[35] C.K. Leung, et al., "Data mining on open public transit data for transportation analytics during pre-COVID-19 era and COVID-19 era," in INCoS 2020, pp. 133-144.
[36] C.K. Leung, et al., "Fast algorithms for frequent itemset mining from uncertain data," in IEEE ICDM 2014, pp. 893-898.
[37] C.K. Leung, et al., "Personalized DeepInf: enhanced social influence prediction with deep learning and transfer learning," in IEEE BigData 2019, pp. 2871-2880.
[38] C.K. Leung, et al., "Privacy-preserving frequent pattern mining from big uncertain data," in IEEE BigData 2018, pp. 5101-5110.
[39] C.K. Leung, et al., "Scalable vertical mining for big data analytics of frequent itemsets," in DEXA 2018, Part I, pp. 3-17.
[40] C.K. Leung, et al., "Urban analytics of big transportation data for supporting smart cities," in DaWaK 2019, pp. 24-33.
[41] C.K. Leung, F. Jiang, "Big data analytics of social networks for the discovery of "following" patterns," in DaWaK 2015, pp. 123-135.
[42] N. Li, et al., "*t*-closeness: privacy beyond *k*-anonymity and *l*-diversity," in IEEE ICDE 2007, pp. 106-115.
[43] A. Machanavajjhala, et al., "*l*-diversity: privacy beyond *k*-anonymity," ACM TKDD 1(1), 2007, pp. 3:1-3:52.
[44] O.A. Sarumi, C.K. Leung, "Exploiting anti-monotonic constraints for mining palindromic motifs from big genomic data," in IEEE BigData 2019, pp. 4864-4873.
[45] R. Sharma, et al., "Tale of three states: analysis of large person-to-person online financial transactions in three Baltic countries," in IEEE BigData 2019, pp. 1497-1505.
[46] J. Souza, et al., "An innovative big data predictive analytics framework over hybrid big data sources with an application for disease analytics," in AINA 2020, pp. 669-680.
[47] R. Tojiboev, et al., "Adding noise trajectory for providing privacy in data publishing by vectorization," in IEEE BigComp 2020, pp. 432-434.
[48] S. Tsumoto, et al., "Estimation of disease code from electronic patient records, in IEEE BigData 2019, pp. 2698-2707.
[49] N. Wang, et al., "PrivSuper: a superset-first approach to frequent itemset mining under differential privacy," in IEEE ICDE 2017, pp. 809-820.
[50] B.H. Wodi, C.K. Leung, A. Cuzzocrea, S. Sourav, "Fast privacy-preserving keyword search on encrypted outsourced data," in IEEE BigData 2019, pp. 1-10.