A General Overview of Privacy-Preserving Big Data Management and Analytics Models, Methods and Techniques in Specific Domains: Static and Dynamic Distributed Environments

Alfredo Cuzzocrea
University of Trieste and ICAR-CNR
Italy
alfredo.cuzzocrea@dia.units.it

Carlo Mastroianni
ICAR-CNR
Italy
carlo.mastroianni@icar.cnr.it

***Abstract*—Privacy-preserving big data management and analytics is gaining the momentum within the research community, and several current research efforts aim to provide solutions to the challenges that emerge when models, techniques and algorithms must be delivered on top of massive, distributed big data repositories, especially with regards to emerging distributed settings such as Clouds and social networks. In this paper, at the convergence of the contexts of static and dynamic distributed environments, we provide a general overview of models, issues and approaches, along with some reference frameworks. Indeed, both static and dynamic distributed environments are relevant cases of settings where the privacy of big data turns to be critical. Finally, we discuss emerging research directions.**

I. Introduction

While several research efforts have been developed in the context of *privacy-preserving big data management and analytics* recently, relevant challenges arise when such models, techniques and algorithms must be delivered on top of massive, distributed big data repositories. This problem opens the door to the design of innovative models, techniques and algorithms that, contrary to actual proposals, are able to inject the *scalability* feature during the privacy-preserving big data management and analytics phase. On the basis of these considerations, this paper provides an overview on actual problems and limitations of state-of-the-art techniques (e.g., [44]).

Privacy-preserving big data management and analytics (e.g., [15], [20], [18]) is gaining the momentum within the research community. Indeed, a lot of research efforts have been invested with the goal of providing solutions to this emerging challenge recently, especially with regards to emerging distributed settings such as *Clouds* and *social networks*. Nevertheless, these approaches expose several problems when they are applied to massive, distributed big data repositories, hence innovative models, techniques and algorithms are necessary to deal with the *scalability issues* deriving from ensuring privacy-preserving big data management and analytics.

As a matter of fact, scalable solutions are not only a question of high-performance architectures (e.g., [71]), which may be still delivered on powerful Cloud Computing platforms, but also a question of models, techniques and algorithms devoted to the desired goal. From this main motivation, several proposals appeared in literature recently. For instance, [75] presents a *proximity privacy model* which allows us to support semantic proximity of sensitive values and multiple sensitive attributes, and models the problem of local recoding as a *proximity-aware clustering problem*. Therefore, a scalable two-phase clustering approach is proposed. This approach predicates a *t-ancestors clustering* algorithm, and a proximity-aware agglomerative clustering algorithm is proposed to address the derived problem. Experiments demonstrate an effective scalability of the algorithm. Similarly, [51] focuses on privacy concerns when sharing large-scale transactional databases and moves to the big data context. With this goal in mind, the authors present an *optimal procedure leveraging intuition from linear programming based column generation*, and identify a common structure that exists in these problems. Based on this main intuition, the authors show how an approach based on *sorting and column generation* can make the process more efficient. Finally, they illustrate how this structure can be incorporated into the column generation based procedure to develop an effective, scalable heuristic. Experiments show a critical gain over classical methods.

On the other hand, another emerging data management context for Big Data research is represented by the issue of effectively and efficiently supporting *Data Warehousing and OLAP over Big Data* (e.g., [68], [24], [19]), as multidimensional data analysis paradigms are likely to become an "enabling technology" for *analytics over Big Data* (e.g., [14], [29]), a collection of models, algorithms and techniques oriented to extract useful knowledge from Cloud-based Big Data repositories for decision making and analysis purposes.

At the convergence of the three axioms introduced above (i.e., security and privacy of Big Data, Data Warehousing and OLAP over Big Data, analytics over Big Data), a critical research challenge is represented by the issue of *effectively* and *efficiently computing privacy-preserving OLAP data cubes*

978-1-5386-5035-6/18/$31.00 ©2018 IEEE

over Big Data (e.g., [25], [27]). It is easy to foresee that this problem will become more and more important in future years, as it not only involves into relevant theoretical and methodological aspects, not all explored by actual literature, but also it regards significant modern scientific applications, such as *biomedical tools over Big Data* (e.g., [10], [53]), *e-science and e-life Big Data applications* (e.g., [9], [34]), *intelligent tools for exploring Big Data repositories* (e.g., [11], [35]), and so forth.

By analyzing the state-of-the-art, a relevant number of critical challenges emerge. The following are among the most relevant ones:

- *devising scalable models, techniques and algorithms for supporting scalable privacy-preserving big data management and analytics*: starting from classical proposals, innovative solutions need to be devised, in order to deal with scalability issues deriving from generating, accessing and processing privacy-preserving big data;
- *formal models for measuring the privacy of big data over massive repositories*: in order to assess the quality and the accuracy of proposed solutions, a strident need for measuring the (achieved) privacy of such solutions arises – currently, no relevant proposals that focus on this critical problem are available in the literature;
- *accuracy support*: scalable privacy-preserving big data management and analytics techniques must also provide enough guarantees on the accuracy of the upper-lying big data analytics processes – indeed, very often privacy and accuracy of big data are conflicting properties (e.g., [27]), hence supporting the accuracy of such processes, especially on massive, distributed big data repositories becomes problematic;
- *integration with big data processing platforms*: at a pure practical level, integrating innovative scalable models, techniques and algorithms for supporting scalable privacy-preserving big data management and analytics with latest big data processing platforms, such as *Apache Spark* [73], is a critical issue – this because proposed approaches must be made *fully-parallelizable* in order to be successfully delivered on top of these platforms.

Another interesting context that is related to this research is the *privacy of big data streams*. Indeed, with the relevant growth of big data (e.g., [49], [33], [28]) observed recently, the problem of *mining and extracting knowledge from such kind of data is gaining momentum* (e.g., [30], [14], [19]). Among the various characteristics of big data [42], *velocity* is, without doubts, one among the most relevant ones, hence conferring to *big data streams* (e.g., [57], [56], [67]) the first-class role for such data. Therefore, *mining big data streams* is relevant and necessary, as confirmed by recent initiatives in this research context (e.g., [59], [37], [63], [3], [76], [7]). With the mining problem, another relevant problem arises: the issue of *preserving the privacy of big data stream sources while mining such data* (e.g., [1], [41], [54], [46], [6], [52], [74]). It is easy to understand how the so-depicted problem has a relevant number of real-life application scenarios, ranging from *trajectory data stream management* to *electronic health data stream processing*, from *fraud detection and analysis of business data streams* to *surveillance and emergency management*, and so forth.

Among several privacy-preservation strategies (e.g., [69]), three relevant approaches are:

- *data anonymization*;
- *privacy-preserving data publishing*;
- *differential privacy*.

Data anonymization consists in meaningfully erasing/masking critical stream attributes that may breach the privacy of the target streams. Well-known approaches in the static case are: *k-anonymization* [65], *l-diversity* [48], *t-closeness* [45]. Relevant extensions to the streaming (i.e., dynamic) case are: [8], [39]. Privacy-preserving data publishing (e.g., [36]) is a *generalization method* that consists in making published data as much as possible *hostile* so that knowledge expressed by such data is still useful while individual privacy is preserved. Significant approaches that appear in the streaming context are presented in [50] and [1]. Finally, differential privacy is a modern technique allowing us to obtain privacy preservation of data mining algorithms. Basically, the differential privacy axioms [32] argue that a *differentially-private algorithm* does not change behavior if information related to a single individual in the target dataset is modified or deleted. Differential privacy has been applied to the privacy-preserving data stream mining as well, being [6], [52], [12] some noticeable initiatives.

In this paper, at the convergence of the two main topics, we focus the attention on privacy-preserving research in two distinct specific domains: *static and dynamic distributed environments*. These are relevant cases of settings where studying the privacy of big data turns to be critical. In particular, the contribution of this paper is two-fold. The first one, related to the issue of supporting privacy-preserving big data management and analytics in distributed static environments, is the description of `DRIPROM`, *an innovative framework for supporting privacy-preserving big data via aggregate-provenance big data analysis* [21], in Section II. The second one, related to the issue of supporting privacy-preserving big data management and analytics in distributed dynamic environments, is an overview, enriched by discussion, of *state-of-the-art privacy-preserving big data stream management and mining techniques* [17], in Section III. Finally, we discuss relevant research directions in Section IV.

II. DRIPROM: A Framework for Supporting Privacy-Preserving Big Data via Aggregate-Provenance Big Data Analysis in Distributed Settings

`DRIPROM` [22] is a relevant implementation of the general guidelines provided in Section I where the scalable privacy-preserving big data management and analytics goal is achieved via the so-called *aggregate-provenance big data analysis*.

In particular, in DRIPROM, suitable *aggregation-based summaries* are computed on top of the input big data sources. Then, during the big data management and analytics phase, at each step, the framework checks, by means of a data-driven method, whether these summaries have the correct "*pedigree*" for supporting a safe privacy-preserving analytical computation. This allows us to avoid resource-intensive computational overheads due to classical *protocol-based privacy-preserving big data mechanisms* (e.g., [62], [72], [47]).

One of the most relevant research challenges arising in the depicted reference application scenario is represented by the method used to check the "pedigree" of summary data representatives. In our research, we propose to apply well-known provenance recognition methods (e.g., [5], [40], [13], [58], [64], [61]). Given two data sets D_i and D_j, the provenance recognition problem consists in detecting if D_j has been "produced" from D_i via some arbitrary processing procedures. Formally, we denote this property as follows:

$$D_j = \mathcal{P}(D_i) \tag{1}$$

such that $\mathcal{P}$ models the procedure that computes D_j from D_i. The final goal is to "synthesize" $\mathcal{P}$ from the analysis of D_j and D_i. When analyzing the literature, it emerges that provenance is a relevant problem in the context of security and privacy of databases, traditionally, and, more recently, in the context of security and privacy of big data (e.g., [16], [66], [60], [55], [2], [70]). In our proposal, we make use of these results as baseline tools of our proposed framework.

Indeed, DRIPROM is a relevant realization for supporting data-driven privacy-preserving big data management in distributed environments. DRIPROM works on *big multidimensional data* [29] and supports the two following fundamental procedures:

- given a big multidimensional data source D_{B_i}, the summary representative of D_{B_i}, $D^P_{B_i}$, is obtained by computing a *privacy-preserving sample* of D_{B_i}, for instance by applying the approach proposed in [26];
- given a summary representative $D^P_{B_i}$, the problem of recognizing if $D^P_{B_i}$ is a reliable summary of D_{B_i} without accessing the entire big multidimensional data source D_{B_i}, is addressed and solved by means of a provenance recognition method, for instance by applying the approach proposed in [4].

Our methodology is orthogonal to the specific algorithms used to obtain the big multidimensional data representative and to check the provenance relation. This means that any algorithm available in the literature can be exploited to this end. This gives a clear *openess* nature to our proposed framework.

Figure 1 reports the conceptual scheme that is at the basis of the privacy-preserving phase of DRIPROM. We propose using sampling-based techniques as several studies have already demonstrated the nice *flexibility* ensured to the privacy-preserving goal by this class of techniques (e.g., [26]). As an interesting extension, considering this issue in the context of *uncertain multidimensional data* (e.g., [23]) represents a relevant research challenge.

Figure 2 reports the conceptual scheme that is at the basis of the provenance-checking phase of DRIPROM. It is worth noticing that, contrary to the privacy-preserving phase, here multiple proposals can be used (e.g., [4], [31], [43], [38]); this further confirms the effectiveness of our proposed framework when combined and integrated with several big data processing algorithms and techniques.

The logical architecture of DRIPROM is represented in Figure 3: it combines component-oriented and scalable organizations of modern Cloud-based applications and systems. Every node that contributes to implement the data-driven privacy-preserving big data management framework must adhere to this logical architecture. As shown in Figure 3, the architecture introduces the following layers/modules:

- *Big Multidimensional Data Layer*: it is the layer where the big multidimensional data sources are located;
- *Big Multidimensional Data Access Module*: this module is responsible for providing the necessary access routines and procedures over the target big multidimensional data sources;
- *Privacy-Preserving Big Multidimensional Data Module*: this module provides the algorithms and techniques for supporting the privacy-preserving phase of DRIPROM;
- *Big Multidimensional Data Representative Layer*: it is the layer where the big multidimensional data representatives are located;
- *Provenance-Checking Module*: this module is in charge of providing algorithms and techniques for supporting the provenance-checking phase of DRIPROM;
- *Cloud-Based Service-Oriented Interface*: it is the component by which the target big multidimensional data sources are interconnected with the overlying Cloud-aware big analytics functions.

III. Privacy-Preserving Big Data Stream Management and Mining: State-Of-The-Art

A plethora of proposals appear in the research literature, which focus on the privacy-preserving big data stream mining problem. In the following, we report on some of the recent most-noticeable ones.

[17] proposes a novel algorithm for *anonymizing trajectory data streams*. In particular, the proposed approach, called *Incremental Trajectory Stream Anonymizer* (ITSA), is incremental in nature, and it makes use of suitable sliding windows in order to process target streams. Such windows are updated as soon as stream individuals join and leave, dynamically. To

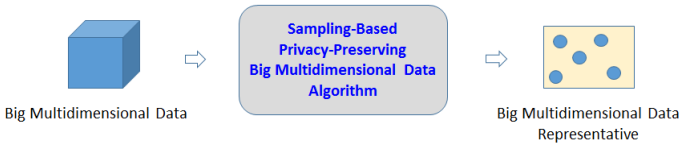

Fig. 1: The Privacy-Preserving Phase in DRIPROM

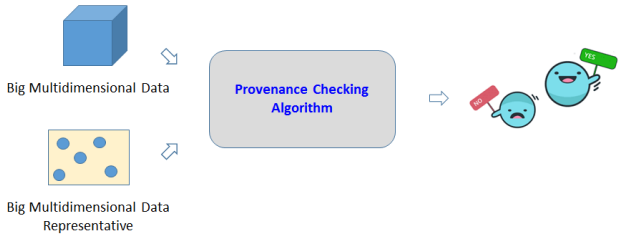

Fig. 2: The Provenance-Checking Phase in DRIPROM

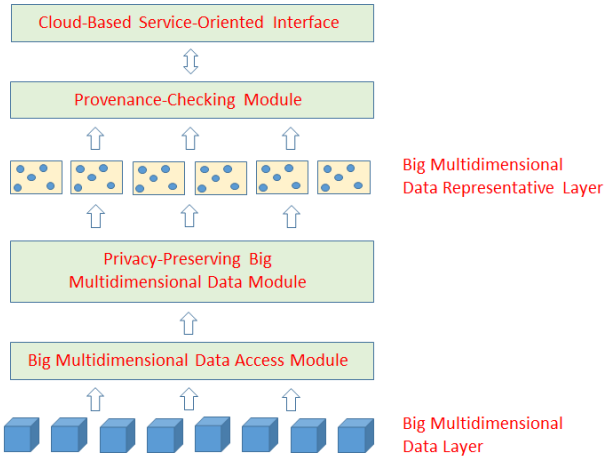

Fig. 3: DRIPROM Logical Architecture

this end, efficient data structures are exploited, so as to comply with massive sizes of big data streams. Extensive experiments confirm the benefits coming from the proposed algorithm.

[18] moves the attention to the context of *electronic health data streams*, whose privacy is deeply studied. The final proposal is called *delay-free anonymization*. The main property of this proposal consists in the amenity that input streams are anonymized *immediately* with counterfeit values. In order to further improve the data utility of the anonymization phase, the authors propose an innovative *late validation* methodology that magnifies the overall privacy-preservation effect.

[19] focuses on the privacy of trajectory streams. In particular, the solution is conducted via proper *access control mechanisms over data streams*, by devising an innovative *Privacy Protection Mechanism* (PPM). The authors prove that PPM meets solid privacy requirements as those dictated by the *k*-anonymity and *l*-diversity schemes over data streams, respectively, via *generalization*. However, since the PPM methodology is based on delaying the publishing of data streams, this may introduce inaccuracies due to false-negatives that affect query processing. Therefore, the authors recognize a new problem, called *precision-bounded access control for privacy-preservation stream mining*, and provide hardness results, enriched by comprehensive experimental evaluation.

[20] proposes *Shadow Coding*, a method that preserves the privacy in *data transmission* and ensures the recovery in *data collection* in distributed data stream settings. The authors prove that the proposed method achieves privacy-preserving computation in a data-recoverable, efficient, and scalable way, being scalability a first-class requirement for big data processing (e.g., [35,36]). Practical techniques that make *Shadow Coding* efficient and safe over data streams are also provided. The authors complete their analytical contributions by means of an extensive experimental study on a large-scale real-life dataset.

[21] focuses on the general problem of *estimating the sortedness of data streams in a privacy-preserving way* by computing the length of the *Longest Increasing Subsequence* (LIS) of target data streams. The authors show that this relevant property of data streams can turn to be extremely useful in a plethora of modern applications, such as finance data stream monitoring, surveillance data stream processing, and so forth. The basic idea consists in exploiting *block decomposition* and *local approximation* techniques.

[22] addresses the problem of supporting *privacy-preserving geo-spatial data stream publishing* via differential privacy techniques. The method consists in supporting accurate query processing over geo-spatial data streams with well-defined *dynamic scopes* (hence the privacy requirement derives), by computing *suitable synopsis with high utility*. The proposed technique, called *Realtime Geospatial Publish* (RGP), is theoretically-sound and experimentally-solid.

[23] considers *categorical data streams*. In this respect, the authors propose a novel anonymization technique for providing a strong privacy protection to safeguard against *privacy disclosure* and *information tampering*. The proposed technique introduces an innovative *two-phase anonymization approach*. The authors prove that such an approach is highly efficient in terms of speed and communication, and robust against possible tampering from adversaries. Extensive experimental evaluation confirms the goodness of the proposed technique.

Finally, [40] investigates the interesting problem of protecting the *output-privacy* of classification algorithms (e.g., the privacy of the classifiers results) over data streams. The authors propose a systematic method implemented by the so-called *Diverse and k-Anonymized HOeffding Tree* (DAHOT) algorithm, which is a meaningful combination of the popular data stream classification algorithm *Hoeffding tree* and suitable variants of k-anonymity and l-diversity principles.

IV. Future Emerging Research Directions

In this Section, we focus the attention on future research directions in the context of scalable privacy-preserving big data management and analytics in static and dynamic distributed environments.

Privacy-Preserving Indexing Data Structures. While effective and efficient algorithms and techniques can be devised, a relevant problem is still represented by the issue of *indexing big data*, for both management and analytics purposes, *while preserving the privacy of big data themselves*. There is an emerging call for supporting privacy-preserving massive data management operations, via extending traditional database-oriented indexing data structures (e.g., *B*-trees, *R*-trees, and so

forth) and enriching them with privacy-preservation features. A possible solution is represented by the so-called *cluster indexes*, which can well-adapt to the enormous size of data.

Privacy-Preserving Partioned-Based Big Data Management. In order to achieve parallel solutions for supporting privacy-preserving big data management and analytics, it is mandatory to enable models and techniques allowing us to manage big data repositories based on *intelligent partitioning paradigms* that are also able to protect the privacy of target big data. Partitions are derived from the target big data set on the basis of a certain criterion (e.g., minimizing a given approximation error, or maximizing a given privacy degree), and then processed on top of the reference distributed environment via suitable parallel algorithms and techniques. Models and techniques for supporting algorithm-oriented and task-oriented big data partitioning that preserves the privacy of big data, will play a relevant role in future years.

Scalable Privacy-Preserving Big Data Analytics. Applying privacy-preserving knowledge discovery and prediction over big data repositories can represent a critical issue, due to specific nature of such data. Therefore, classical KDD (*Knowledge Discovery in Databases*) solutions, which have later been extended by means of privacy-preserving procedures, cannot be directly applied to this end, and supporting scalable big data analytics represents a first-class challenge for next years. A possible research direction could consist in devising *incremental analytical procedures* that run over big data while considering both the need for big data privacy preservation, and the need for big data management optimization.

Scalable Provenance Methods over Big Data. Provenance was discussed in Section II as one possible solution to obtain privacy-preserving big data management and analytics. Nevertheless, obtaining *scalable provenance methods* over big data is still an open research issue, due to the fact that applying provenance algorithms is really resource-consuming. In this context, a lot of work is still to be done, also by adopting advanced tools such as *probabilistic models* and *heuristic algorithms*.

Secure, Scalable Privacy-Preserving Methods over Big Data. Scalability is not the sole challenge to be faced-off when dealing with big data privacy-preservation. Another critical requirement is represent by the need for obtaining privacy-preserving big data management and analytics methods while *ensuring the security of target big data sources*, still with an eye over scalability. Combining privacy with security and scalability represents a critical challenge for future research efforts in the field.

Integration with NoSQL Architectures. The *NoSQL movement* is gaining the scene, actually. Indeed, NoSQL databases expose several features, among which: high-scalability, schema-lessness, distributed architectures, and so forth, which well-marry with the applicative requirements of supporting scalable privacy-preserving big data management and analytics. Therefore, it is easy to foresee the integration of both paradigms, with relevant innovations.

Application Scenarios. Last but not least, studying, devising, and prototyping *interesting application scenarios* where the scalable privacy-preserving big data management and analytics requirements can be fixed and formally defined, represents another milestone to be considered in future years.

Emerging Domains. It is clear enough that, due to the specific research focus, i.e. privacy-preserving big data stream mining, practical applications and systems *drive* and *determine* the effective requirements for corresponding privacy-preserving big data stream mining algorithms. Hence, emerging domains, such as social networks, intelligent TV provisioning and intelligent transportation systems, will play a first-class role in the future.

Accuracy vs Privacy. Accuracy and privacy are *conflict properties* for big data stream mining algorithms. Indeed, determining the correct trade-off between these two properties is a fundamental research issue. How to increase privacy while preserving accuracy? This is a relevant question for future research activities.

Concept-Drift Issues. Big data streams are affected by *concept-drift problems*. Matching the privacy-preserving requirement can be hard because preserving the privacy of data is performed in dependence on a *pre-determined* set of attributes/concepts of the target data model.

Security Issues. Preserving privacy when accessing big data streams involves a problematic side-effect: how to ensure the *security* of big data streams while accessing them? Combining security and privacy is an annoying problem for the big data stream mining research.

Cryptography. Traditionally, the privacy-preserving big data stream mining problem is addressed by means of model-based or algorithmic-based approaches. Along with these initiatives, the usage of *cryptographic methodologies* is emerging as a promising approach to be explored by future research efforts.

Quality and Utility of Data. Ensuring the privacy of data streams while mining big data streams can deteriorate the *quality* and *utility* of such data. This critical issue needs to be tackled in the future.

Stream Analytics. Models, techniques and algorithms proposed in the literature must converge in suitable *unifying frameworks* for finally supporting *privacy-preserving big data stream analytics*, a critical research challenge at present. Several issues arise, ranging from architectural requirements to parameter tuning, framework trade-offs, performance, and so forth.

Performance. Last but not least, *performance issues* always

arise when processing big data streams (to preserve their privacy, in this case). As a consequence, devising models and optimizations that allow us to ensure performance of privacy-preservation mining methods over big data streams is a relevance challenge for the future.

V. Conclusions

This paper has provided an overview on the issues and limitations of state-of-the-art scalable privacy-preserving big data management and analytics techniques, in both static and dynamic distributed environments. Possible research directions have been discussed. We hope this contribution can become a useful reference for future research efforts.

References

[1] K. Al-Hussaeni, B. C. M. Fung, and W. K. Cheung. Privacy-preserving trajectory stream publishing. *Data Knowl. Eng.*, 94:89–109, 2014.

[2] A. Albatli, D. McKee, P. Townend, L. Lau, and J. Xu. PROV-TE: A provenance-driven diagnostic framework for task eviction in data centers. In *Third IEEE International Conference on Big Data Computing Service and Applications, BigDataService 2017, Redwood City, CA, USA, April 6-9, 2017*, pages 233–242, 2017.

[3] D. Amagata and T. Hara. Mining top-k co-occurrence patterns across multiple streams. *IEEE Trans. Knowl. Data Eng.*, 29(10):2249–2262, 2017.

[4] Y. Amsterdamer, D. Deutch, and V. Tannen. Provenance for aggregate queries. In *Proceedings of the 30th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2011, June 12-16, 2011, Athens, Greece*, pages 153–164, 2011.

[5] D. W. Archer, L. M. L. Delcambre, and D. Maier. User trust and judgments in a curated database with explicit provenance. In *In Search of Elegance in the Theory and Practice of Computation - Essays Dedicated to Peter Buneman*, pages 89–111, 2013.

[6] L. Bonomi and L. Xiong. On differentially private longest increasing subsequence computation in data stream. *Transactions on Data Privacy*, 9(1):73–100, 2016.

[7] P. Braun, J. J. Cameron, A. Cuzzocrea, F. Jiang, and C. K. Leung. Effectively and efficiently mining frequent patterns from dense graph streams on disk. In *18th International Conference in Knowledge Based and Intelligent Information and Engineering Systems, KES 2014, Gdynia, Poland, 15-17 September 2014*, pages 338–347, 2014.

[8] J. Cao, B. Carminati, E. Ferrari, and K. Tan. CASTLE: continuously anonymizing data streams. *IEEE Trans. Dependable Sec. Comput.*, 8(3):337–352, 2011.

[9] Y. Cheah, S. R. Canon, B. Plale, and L. Ramakrishnan. Milieu: Lightweight and configurable big data provenance for science. In *IEEE International Congress on Big Data, BigData Congress 2013, Santa Clara, CA, USA, June 27 2013-July 2, 2013*, pages 46–53, 2013.

[10] X. Chen, H. Chen, N. Zhang, J. Chen, and Z. Wu. OWL reasoning over big biomedical data. In *Proceedings of the 2013 IEEE International Conference on Big Data, 6-9 October 2013, Santa Clara, CA, USA*, pages 29–36, 2013.

[11] D. Cheng, P. Schretlen, N. Kronenfeld, N. Bozowsky, and W. Wright. Tile based visual analytics for twitter big data exploratory analysis. In *Proceedings of the 2013 IEEE International Conference on Big Data, 6-9 October 2013, Santa Clara, CA, USA*, pages 2–4, 2013.

[12] M. Cheng, Y. Sun, B. Zhao, and J. Su. An event grouping approach for infinite stream with differential privacy. In *Advances in Services Computing - 10th Asia-Pacific Services Computing Conference, APSCC 2016, Zhangjiajie, China, November 16-18, 2016, Proceedings*, pages 106–116, 2016.

[13] F. Costa, V. S. Sousa, D. de Oliveira, K. A. C. S. Ocaña, and M. Mattoso. Towards supporting provenance gathering and querying in different database approaches. In *Provenance and Annotation of Data and Processes - 5th International Provenance and Annotation Workshop, IPAW 2014, Cologne, Germany, June 9-13, 2014. Revised Selected Papers*, pages 254–257, 2014.

[14] A. Cuzzocrea. Analytics over big data: Exploring the convergence of datawarehousing, OLAP and data-intensive cloud infrastructures. In *37th Annual IEEE Computer Software and Applications Conference, COMPSAC 2013, Kyoto, Japan, July 22-26, 2013*, pages 481–483, 2013.

[15] A. Cuzzocrea. Privacy and security of big data: Current challenges and future research perspectives. In *Proceedings of the First International Workshop on Privacy and Secuirty of Big Data, PSBD@CIKM 2014, Shanghai, China, November 7, 2014*, pages 45–47, 2014.

[16] A. Cuzzocrea. Big data provenance: State-of-the-art analysis and emerging research challenges. In *Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference, EDBT/ICDT Workshops 2016, Bordeaux, France, March 15, 2016.*, 2016.

[17] A. Cuzzocrea. Privacy-preserving big data stream mining: Opportunities, challenges, directions. In *2017 IEEE International Conference on Data Mining Workshops, ICDM Workshops 2017, New Orleans, LA, USA, November 18-21, 2017*, pages 992–994, 2017.

[18] A. Cuzzocrea. Scalable olap-based big data analytics over cloud infrastructures: Models, issues, algorithms. In *Proceedings of the 2017 International Conference on Cloud and Big Data Computing, ICCBDC 2017, London, United Kingdom, September 17 - 19, 2017*, pages 17–21, 2017.

[19] A. Cuzzocrea, L. Bellatreche, and I. Song. Data warehousing and OLAP over big data: current challenges and future research directions. In *Proceedings of the sixteenth international workshop on Data warehousing and OLAP, DOLAP 2013, San Francisco, CA, USA, October 28, 2013*, pages 67–70, 2013.

[20] A. Cuzzocrea and E. Bertino. Privacy preserving OLAP over distributed XML data: A theoretically-sound secure-multiparty-computation approach. *J. Comput. Syst. Sci.*, 77(6):965–987, 2011.

[21] A. Cuzzocrea and E. Damiani. Pedigree-ing your big data: Data-driven big data privacy in distributed environments. In *18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, CCGRID 2018, Washington, DC, USA, May 1-4, 2018*, pages 675–681, 2018.

[22] A. Cuzzocrea and E. Damiani. Pedigree-ing your big data: Data-driven big data privacy in distributed environments. In *18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, CCGRID 2018, Washington, DC, USA, May 1-4, 2018*, pages 675–681, 2018.

[23] A. Cuzzocrea and D. Gunopulos. A decomposition framework for computing and querying multidimensional OLAP data cubes over probabilistic relational data. *Fundam. Inform.*, 132(2):239–266, 2014.

[24] A. Cuzzocrea, R. Moussa, and G. Xu. Olap*: Effectively and efficiently supporting parallel OLAP over big data. In *Model and Data Engineering - Third International Conference, MEDI 2013, Amantea, Italy, September 25-27, 2013. Proceedings*, pages 38–49, 2013.

[25] A. Cuzzocrea and V. Russo. Privacy preserving OLAP and OLAP security. In *Encyclopedia of Data Warehousing and Mining, Second Edition (4 Volumes)*, pages 1575–1581. 2009.

[26] A. Cuzzocrea, V. Russo, and D. Saccà. A robust sampling-based framework for privacy preserving OLAP. In *Data Warehousing and Knowledge Discovery, 10th International Conference, DaWaK 2008, Turin, Italy, September 2-5, 2008, Proceedings*, pages 97–114, 2008.

[27] A. Cuzzocrea and D. Saccà. Balancing accuracy and privacy of OLAP aggregations on data cubes. In *DOLAP 2010, ACM 13th International Workshop on Data Warehousing and OLAP, Toronto, Ontario, Canada, October 30, 2010, Proceedings*, pages 93–98, 2010.

[28] A. Cuzzocrea, D. Saccà, and J. D. Ullman. Big data: a research agenda. In *17th International Database Engineering & Applications Symposium, IDEAS '13, Barcelona, Spain - October 09 - 11, 2013*, pages 198–203, 2013.

[29] A. Cuzzocrea, I. Song, and K. C. Davis. Analytics over large-scale multidimensional data: the big data revolution! In *DOLAP 2011, ACM 14th International Workshop on Data Warehousing and OLAP, Glasgow, United Kingdom, October 28, 2011, Proceedings*, pages 101–104, 2011.

[30] A. Cuzzocrea, I. Song, and K. C. Davis. Analytics over large-scale multidimensional data: the big data revolution! In *DOLAP 2011, ACM 14th International Workshop on Data Warehousing and OLAP, Glasgow, United Kingdom, October 28, 2011, Proceedings*, pages 101–104, 2011.

[31] R. Q. Dividino, G. Gröner, S. Scheglmann, and M. Thimm. Ranking RDF with provenance via preference aggregation. In *Knowledge Engineering and Knowledge Management - 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012. Proceedings*, pages 154–163, 2012.

[32] C. Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation, 5th International Conference,*

TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings, pages 1–19, 2008.

[33] C. Eaton, D. DeRoos, T. Deutsch, G. Lapis, and P. Zikopoulos. *Understanding big data : analytics for enterprise class Hadoop and streaming data*. McGraw-Hill, New York, NY, USA, 2012.

[34] A. G. Erdman, D. F. Keefe, and R. Schiestl. Grand challenge: Applying regulatory science and big data to improve medical device innovation. *IEEE Trans. Biomed. Engineering*, 60(3):700–706, 2013.

[35] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2149–2158, 2013.

[36] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42(4):14:1–14:53, 2010.

[37] M. M. Gaber, J. Gama, S. Krishnaswamy, J. B. Gomes, and F. T. Stahl. Data stream mining in ubiquitous environments: state-of-the-art and current directions. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 4(2):116–138, 2014.

[38] F. Giunghiglia and M. Reyad. Provenance in open data entity-centric aggregation. In *Provenance and Annotation of Data and Processes - 5th International Provenance and Annotation Workshop, IPAW 2014, Cologne, Germany, June 9-13, 2014. Revised Selected Papers*, pages 232–234, 2014.

[39] K. Guo and Q. Zhang. Fast clustering-based anonymization approaches with time constraints for data streams. *Knowl.-Based Syst.*, 46:95–108, 2013.

[40] G. Karvounarakis, T. J. Green, Z. G. Ives, and V. Tannen. Collaborative data sharing via update exchange and provenance. *ACM Trans. Database Syst.*, 38(3):19:1–19:42, 2013.

[41] S. Kim, M. K. Sung, and Y. D. Chung. A framework to preserve the privacy of electronic health data streams. *Journal of Biomedical Informatics*, 50:95–106, 2014.

[42] D. Laney. 3D Data Management: Controlling Data Volume, Velocity, and Variety, Feb. 2001.

[43] C. Lettner, M. Pichler, W. Kirchmayr, F. Kokert, and M. Habringer. Rdfreduce: Customized aggregations with provenance for RDF data based on an industrial use case. In *The 15th International Conference on Information Integration and Web-based Applications & Services, IIWAS '13, Vienna, Austria, December 2-4, 2013*, page 336, 2013.

[44] K. Li, H. Jiang, L. T. Yang, and A. Cuzzocrea, editors. *Big Data - Algorithms, Analytics, and Applications*. Chapman and Hall/CRC, 2015.

[45] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007*, pages 106–115, 2007.

[46] S. Liu, Q. Qu, L. Chen, and L. M. Ni. SMC: A practical schema for privacy-preserved data sharing over distributed data streams. *IEEE Trans. Big Data*, 1(2):68–81, 2015.

[47] Z. Liu, K. R. Choo, and M. Zhao. Practical-oriented protocols for privacy-preserving outsourced big data analysis: Challenges and future research directions. *Computers & Security*, 69:97–113, 2017.

[48] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. *L*-diversity: Privacy beyond *k*-anonymity. *TKDD*, 1(1):3, 2007.

[49] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity, May 2011.

[50] N. Medforth and K. Wang. Privacy risk in graph stream publishing for social network data. In *11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14, 2011*, pages 437–446, 2011.

[51] S. Menon and S. Sarkar. Privacy and big data: Scalable approaches to sanitize large transactional databases for sharing. *MIS Quarterly*, 40(4):963–981, 2016.

[52] Y. Nie, L. Huang, Z. Li, S. Wang, Z. Zhao, W. Yang, and X. Lu. Geospatial streams publish with differential privacy. In *Collaborate Computing: Networking, Applications and Worksharing - 12th International Conference, CollaborateCom 2016, Beijing, China, November 10-11, 2016, Proceedings*, pages 152–164, 2016.

[53] M. Paoletti, G. Camiciottoli, E. Meoni, F. Bigazzi, L. Cestelli, M. Pistolesi, and C. Marchesi. Explorative data analysis techniques and unsupervised clustering methods to support clinical assessment of chronic obstructive pulmonary disease (COPD) phenotypes. *Journal of Biomedical Informatics*, 42(6):1013–1021, 2009.

[54] Z. Pervaiz, A. Ghafoor, and W. G. Aref. Precision-bounded access control using sliding-window query views for privacy-preserving data streams. *IEEE Trans. Knowl. Data Eng.*, 27(7):1992–2004, 2015.

[55] I. Portugal, P. S. C. Alencar, and D. D. Cowan. Towards a provenance-aware spatial-temporal architectural framework for massive data integration and analysis. In *2016 IEEE International Conference on Big Data, BigData 2016, Washington DC, USA, December 5-8, 2016*, pages 2686–2691, 2016.

[56] D. Puthal, S. Nepal, R. Ranjan, and J. Chen. Dlsef: A dynamic key-length-based efficient real-time security verification model for big data stream. *ACM Trans. Embedded Comput. Syst.*, 16(2):51:1–51:24, 2017.

[57] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Wozniak, J. M. Benítez, and F. Herrera. Nearest neighbor classification for high-speed big data streams using spark. *IEEE Trans. Systems, Man, and Cybernetics: Systems*, 47(10):2727–2739, 2017.

[58] A. Rani, N. Goyal, and S. K. Gadia. Data provenance for historical queries in relational database. In *Proceedings of the 8th Annual ACM India Conference, Ghaziabad, India, October 29-31, 2015*, pages 117–122, 2015.

[59] P. P. Rodrigues and J. Gama. Distributed clustering of ubiquitous data streams. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 4(1):38–54, 2014.

[60] R. J. Sandusky. Computational provenance: Dataone and implications for cultural heritage institutions. In *2016 IEEE International Conference on Big Data, BigData 2016, Washington DC, USA, December 5-8, 2016*, pages 3266–3271, 2016.

[61] P. Senellart. Provenance and probabilities in relational databases. *SIGMOD Record*, 46(4):5–15, 2017.

[62] M. Sepehri, S. Cimato, E. Damiani, and C. Y. Yeun. Data sharing on the cloud: A scalable proxy-based protocol for privacy-preserving queries. In *2015 IEEE TrustCom/BigDataSE/ISPA, Helsinki, Finland, August 20-22, 2015, Volume 1*, pages 1357–1362, 2015.

[63] A. Silva and C. Antunes. Multi-relational pattern mining over data streams. *Data Min. Knowl. Discov.*, 29(6):1783–1814, 2015.

[64] S. Sultana and E. Bertino. A distributed system for the management of fine-grained provenance. *J. Database Manag.*, 26(2):32–47, 2015.

[65] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[66] Y. Tas, M. J. Baeth, and M. S. Aktas. An approach to standalone provenance systems for big social provenance data. In *12th International Conference on Semantics, Knowledge and Grids, SKG 2016, Beijing, China, August 15-17, 2016*, pages 9–16, 2016.

[67] A. Valsamis, K. Tserpes, D. Zissis, D. Anagnostopoulos, and T. A. Varvarigou. Employing traditional machine learning algorithms for big data streams analysis: The case of object trajectory prediction. *Journal of Systems and Software*, 127:249–257, 2017.

[68] M. Weidner, J. Dees, and P. Sanders. Fast OLAP query execution in main memory on large data in a cluster. In *Proceedings of the 2013 IEEE International Conference on Big Data, 6-9 October 2013, Santa Clara, CA, USA*, pages 518–524, 2013.

[69] M. Weidner, J. Dees, and P. Sanders. Fast OLAP query execution in main memory on large data in a cluster. In *Proceedings of the 2013 IEEE International Conference on Big Data, 6-9 October 2013, Santa Clara, CA, USA*, pages 518–524, 2013.

[70] D. Wu, S. Sakr, and L. Zhu. HDM: optimized big data processing with data provenance. In *Proceedings of the 20th International Conference on Extending Database Technology, EDBT 2017, Venice, Italy, March 21-24, 2017.*, pages 530–533, 2017.

[71] C. Yang, J. Liu, C. Hsu, and W. Chou. On improvement of cloud virtual machine availability with virtualization fault tolerance mechanism. *The Journal of Supercomputing*, 69(3):1103–1122, 2014.

[72] X. Yang, R. Lu, H. Liang, and X. Tang. SFPM: A secure and fine-grained privacy-preserving matching protocol for mobile social networking. *Big Data Research*, 3:2–9, 2016.

[73] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin, A. Ghodsi, J. Gonzalez, S. Shenker, and I. Stoica. Apache spark: a unified engine for big data processing. *Commun. ACM*, 59(11):56–65, 2016.

[74] J. Zhang, H. Li, X. Liu, Y. Luo, F. Chen, H. Wang, and L. Chang. On efficient and robust anonymization for privacy protection on massive streaming categorical information. *IEEE Trans. Dependable Sec. Comput.*, 14(5):507–520, 2017.

[75] X. Zhang, W. Dou, J. Pei, S. Nepal, C. Yang, C. Liu, and J. Chen. Proximity-aware local-recoding anonymization with mapreduce for scalable big data privacy preservation in cloud. *IEEE Trans. Computers*, 64(8):2293–2307, 2015.

[76] M. Zihayat, Y. Chen, and A. An. Memory-adaptive high utility sequential pattern mining over data streams. *Machine Learning*, 106(6):799–836, 2017.