Protecting Personally Identifiable Information (PII) in Critical Infrastructure Data Using Differential Privacy

Asma Alnemari, Rajendra K. Raj, Carol J. Romanowski and Sumita Mishra
Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, NY 14623, USA
[ama8457, rajendra.k.raj, cjrcms, sumita.mishra]@rit.edu

***Abstract*—In critical infrastructure (CI) sectors such as emergency management or healthcare, researchers can analyze and detect useful patterns in data and help emergency management personnel efficaciously allocate limited resources or detect epidemiology spread patterns. However, all of this data contains personally identifiable information (PII) that needs to be safeguarded for legal and ethical reasons. Traditional techniques for safeguarding, such as anonymization, have shown to be ineffective. Differential privacy is a technique that supports individual privacy while allowing the analysis of datasets for societal benefit. This paper motivates the use of differential privacy to answer a wide range of queries about CI data containing PII with better privacy guarantees than is possible with traditional techniques. Moreover, it introduces a new technique based on Multiple-attribute Workload Partitioning, which does not depend on the nature of the underlying dataset and provides better protection for privacy than current differential privacy approaches.**

***Index Terms*—Data privacy, protecting critical infrastructure data, data analytics.**

I. Introduction

Data analytics applied to critical infrastructure data can potentially lead to a variety of societal benefits but also comes with legal and ethical problems because such data typically contains personally identifiable information (PII). In the US healthcare sector, the Health Insurance Portability and Accountability Act (HIPAA) places strict requirements for protecting such information [1]. In the case of 911 calls, such data can include PII involving healthcare data too (for example, "a caller may have reported a cardiac emergency") or report a sensitive event such as a domestic violence situation.

Obscuring methods are applied to such datasets typically during data cleaning and preparation to mask key identifiers, quasi-identifiers, and other types of sensitive data and generate a separate, cleaned dataset for mining [2]. However, this dataset may still, when analyzed using different feature sets, expose sensitive PII [3].

Maintaining individual privacy while making data available for statistical purposes continues to be a major focus for data scientists. When some individuals have sensitive information about particular datasets, many ways exist to compromise individual privacy. For example, Sweeney showed how such datasets could threaten individuals' privacy when some datasets are joined to each other, and proposed the k-anonymity model to prevent re-identification of any individuals data contained in a dataset [4]. Under k-anonymity, a version of a particular dataset can be released with privacy guarantees that any individual whose data is a part of the original version cannot be re-identified while preserving the usefulness of the released data. To satisfy the k-anonymity property, any individual cannot be re-identified between at least $k-1$ individuals whose data within the released dataset [4]. On the other hand, Machanavajjhala et al. [5] demonstrated that k-anonymity is not sufficient to preserve individuals' privacy in cases where adversaries have additional background knowledge about their targets. Even when the access to a particular dataset is restricted, adversaries may be able to reconstruct some parts of the sensitive data using a sequence of counting queries. Multiple attribute counting queries could be more dangerous because they can be used to target small fractions of individuals' data using unique attribute values that have been known or obtained from public resources. For example, shingles disease is unlikely to be found in adults under 60 years old. Suppose an adversarial actor accesses a particular dataset containing health data of a 31 year-old woman that is known to him, and he wants to know if this woman is diagnosed with shingles. Using the answers of a set of correlated counting queries can easily answer his question. For instance, assuming the hidden dataset is the one that shown in table I, the adversary may ask firstly about the number of females who are between 30 and 40 years. Based on this dataset, the answer is 1–which indicates that the targeted woman is the only one who satisfies this condition. So, based on this response, his second query will be 'How many females are in the dataset in the age between 30 and 40, and are diagnosed with 'Shingles'. The response of 1 would verify that the targeted woman has been diagnosed with this disease. Thus, the sequential queries, coupled with some already known data, expose this patient's private information.

Differential privacy models have received considerable attention, as they seem to be a more effective models at preventing PII exposure while maintaining dataset utility [6], including in the healthcare sector [7]. These models preserve

978-1-7281-5092-5/19/$31.00 ©2019 IEEE

TABLE I
AN ANONYMIZED PATIENT DATASET

ID	Gender	Age	Disease
1	M	[25,35]	Flu
2	F	[55,100]	Flu
3	F	[25,35]	Shingles
4	M	[45,55]	Bronchitis
5	M	[55,100]	Shingles

individual privacy while considering background knowledge that adversaries may possess; for instance, anyone who can query the dataset may be able to learn sensitive information about individuals.

Another difference stems from the kind of interaction researchers have with datasets. Traditional techniques for privacy preservation, such as anonymizing datasets and releasing them for analysis, are fundamentally *non-interactive*. Differential privacy, however, can be applied to be an *interactive* model, which is a more realistic scenario that restricts data usage and provides more privacy than a *non-interactive* setting while providing more accurate analysis of the datasets.

This paper discusses how differential privacy can be used in CI sectors and introduces an *interactive* mechanism that can be incorporated with infrastructure systems to obscure PII while making their data available for the use of statistical analysis.

II. MOTIVATION

Although PII is present in data used or generated in every CI sector, this section explores PII in two sectors–emergency management and healthcare–to show that current anonymization does not protect individual privacy, thus motivating the use of differential privacy for protecting PII in data.

PII in 911 Data

Emergency calls contain PII in several different forms. While we automatically assume PII is present in 911 calls that result in ambulance dispatch, other types of calls also expose personal information. For example, a traffic accident, even without injuries, captures a person's name, address, driver's license number, phone number, and insurance carrier. Fire emergencies expose the person's name, address, and medical information about injuries sustained by the residents. Even security checks on industrial buildings contain names and phone numbers of the building's owner or caretaker. PII about first responders are also included in these records: for example, beats, badge numbers, and workplaces of the police officers, EMTs, and firemen are often included in the full transcript of an event. Eliminating or obscuring PII in these records is difficult to automate due to the variety of ways used to record this information. Figure 1 shows a part of the data that have been collected from 911 calls. Observe that the obvious PII such as names, phone numbers, and addresses were removed from this dataset snippet. However, releasing the other information may still expose individual PII. For instance, the combination of code (which is the category of emergency) and lat/long can expose specific neighborhoods or even homes, as in the domestic violence call in the third record .

PII in Healthcare Data

In addition to the pieces of PII mentioned above (e.g., name, address, phone number, social security number, or insurance carrier), healthcare data may include other identifiable information such as biometrics (fingerprints, retina scan, voice signature, facial geometry) or x-rays. It may also include the date of birth, weight, activity levels, information about specific health conditions, medical history, and prescribed medications. Sometimes, individual pieces of information might seem harmless on their own, but in combination with other types of data, they can compromise the privacy of an individual.

Alnemari et al. [8] provide a motivating example that shows patient privacy is easily lost based on queries that can uniquely identify a patient with a certain disease, even if their names are anonymized. With access to other databases or attacker knowledge, the individual will be totally identified. The paper also shows that range searches on anonymized data can reduce privacy loss but not remove it altogether. The paper uses a dataset of patients who were diagnosed with 'cervical cancer' [9] as an example for these privacy issues. Even though this dataset has more than 800 records, an adversary can easily find unique records by querying the 'age', and the 'number of pregnancies' attributes. For example, only one record each matched the 'age' attribute for values of 13, 79, and 84. So, with some additional background about the dataset, such as the name of the hospital that provided this data, an adversary could identify the individuals satisfying these values even though age is not an obviously sensitive attribute.

III. USING DIFFERENTIAL PRIVACY FOR PROTECTING PII IN CRITICAL INFRASTRUCTURE DATA

Differential privacy was originally proposed by Dwork [6] as a property of a randomized algorithm such that no individual whose data is included in a particular data set has a noticeable effect in the distributional outcome produced by this algorithm. As a result, an adversarial analyst who monitors the differentially private release should not be able to learn any information that is too specific to any individual in the dataset. Thus, the outcome produced by a differentially private algorithm should remain almost the same while adding or removing a record to the underlying dataset. Given an arbitrary query q and an ϵ-differently private algorithm M that is designed to answer q, then for any two neighboring datasets D and D' that differ in one element, and any output $S \in Range(M)$ the algorithm M should satisfy the following property:

$$Pr[M_q(D) = S] \leq e^{\varepsilon}.Pr[M_q(D') = S]$$

Differential privacy can be achieved while answering queries by adding amounts of noise to their true responses. So, for a given query q and a dataset D, if y is the true answer of q, then the private answer of q is $y + \alpha$, where α is a random noise. To satisfy differential privacy, the random noise can be drawn from Laplace distribution with mean 0

978-1-7281-5092-5/19/$31.00 ©2019 IEEE

EventID	Date	Time	Lat	Long	Name	Phone	Address	Dispatched	Details	CrossRef
1	4/3	12:30 a.m.	43.21508	-77.4224	*****	***-****	*****	Unit 4	Vehicle on fire	E1
2	4/3	12:40 a.m.	43.21508	-77.4224	*****	***-****	*****		Vehicle on fire	E1
3	4/4	10:30 p.m	43.1686	-77.6215	*****	***-****	*****	Unit 2	Says husband *** ** hit her with fist. Officers *** (Badge **) and Schell (Badge **) responding. Complainant has black eye, does not want to press charges	E3
4	4/5	1:10 a.m.	43.1686	-77.4313	*****	NA	NA	Pump 1	****. Called manager Wm. Price ***-****	E4
5	4/5	8:00 a.m.	43.1686	-77.6218	*****	***-****	*****	Unit 3	Collision with auto driven by ****. NYS drivers license********; Progressive insurance. No injuries.	E5
6	4/5	8:00 a.m.	43.1686	-77.6218	*****	***-****	*****	Unit 3	Rear-ended by vehicle driven by*****. NYS drivers license*******; Allstate Insurance. Has pre-existing back condition. EMS called by Officer (Badge**)	E5

Fig. 1. Part of a dataset collected from 911 calls. PII is represented by the symbol *

and scale Δf, where Δf is the sensitivity of q which is the maximum difference between $M_q(D)$ and $M_q(D')$, and ϵ is the privacy parameter (to achieve more privacy a smaller value should be chosen for ϵ but the accuracy would be decreased as ϵ became smaller) [10].

Differential privacy models assume a trusted curator holds a particular dataset and releases statistical information about the data while maintaining individuals' privacy. For the *non-interactive* models, the curator may prepare some sort of summary about the underlying dataset and release it for analytical use. The curator may receive analytical requests *interactively* as queries and answer them privately. Although the sensitive datasets remain under the control of the curator, ensuring that the released data does not leak sensitive information about the underlying dataset is not a trivial task. According to Vadhan, releasing private responses for an unlimited number of sequential queries may allow reconstructing the sensitive dataset from anonymized responses [11]. Therefore, limiting the number of the given queries or, giving all of them as a workload ensures better privacy while generating the private responses.

Counting queries allow statistical analysis to be performed on a dataset. The expected response of a counting query is the fraction of the entries in a dataset that meet certain conditions [12]. Counting queries may lead to the development of statistical learning models [13]. Therefore, designing models that answer this kind of queries accurately under differential privacy is a critical issue that has attracted many researchers recently. Since most of the infrastructure databases are highly dynamic, using an *interactive* deferentially private model would be more practical than the *non-interactive* models which need to update their releases periodically.

Differential Privacy in CI Systems

CI systems are built over databases that contain different kinds of data. Although most of these systems restrict PII access, some parts of the data are considered to be insensitive and may be released with no access constraints. At the same time, with auxiliary information known to an adversary about the underlying dataset, the released parts of the dataset may lead to individual data loss. Differential privacy over data with no obvious PII may contribute to the preservation of individual privacy. Unlike traditional anonymization techniques, differential privacy does not perform any change over the original data to hide individual data. Instead, differential privacy models generate or return statistical summaries about the underlying dataset such as, for example, anonymized histograms that represent the frequencies of each attribute. Moreover, this approach can be implemented to work interactively while preserving needed privacy for individual data by allowing data analysts to provide their statistical queries and returning answers after ensuring that no sensitive information is revealed.

Our proposed mechanism can be used over CI databases to enable data analysts to interact with these databases and obtain the information needed to conduct their studies. Data analysts must prepare their requests as a workload of counting queries and give it to our mechanism. Based on the sensitivity of the given queries, the mechanism performs the needed anonymization over the actual response and returns it back to the analyst. Our proposed mechanism is effective for CI databases, as it does not require data storage for private information that gets updated often. The responses do not require performing a complex set of operations as do other interactive mechanisms.

IV. The Workload Partitioning Mechanism for Multi-dimensional Datasets

In this section, we discuss an extension to the *Workload Partitioning Mechanism* [8] that takes a workload of counting queries and anonymizes their responses based on their sensitivity. The mechanism identifies the sensitive areas of the data distribution from the ranges of the given queries. The frequencies within these areas are injected with sufficient noise to ensure individual privacy. This strategy contributes to enhanced utility of the produced results [8].

Multiple attribute range queries are more useful for analysts since they capture the relationships between attributes values and those relationships help to discover significant

978-1-7281-5092-5/19/$31.00 ©2019 IEEE

TABLE II
TINY PATIENT DATASET

Age	Number of Pregnancies
18	1
15	1
52	4
26	3
59	0
79	5
84	11

TABLE III
THE PARTITIONED HISTOGRAM OF THE PATIENT DATASET

	Age	Number of Pregnancies	Count
1	[15,40]	[0,4]	3
2	[15,40]	[4,6]	0
3	[15,40]	[6,11]	0
4	[40,50]	[0,4]	0
5	[40,50]	[4,6]	0
6	[40,50]	[6,11]	0
7	[50,84]	[0,4]	1
8	[50,84]	[4,6]	2
9	[50,84]	[6,11]	1

contributing factors. However, adopting the workload partitioning mechanism to handle multiple attributes counting queries is likely to be prohibitive since the mechanism works over a histogram representing the underlying dataset, and the size of the required histogram grows exponentially by the dataset dimensions. A multiple-attribute histogram would have $\prod_{i=1}^{n} m_i$ records where n is the number of attributes and m_i is the number of values that attribute i has. Many researchers have developed models to handle multiple attribute counting range queries [14]–[16]. However, in addition to being data dependent, these models also suffer from the curse of dimensionality, where the histogram size increases exponentially as the number of attributes increases. To tackle this issue, we propose a mechanism that handles multiple attribute counting queries without the need to build a full histogram to represent the underlying dataset.

Our proposed mechanism takes the workload first and then builds a partial histogram based on the ranges of the given queries. This histogram only includes the involved areas in the given workload. That approach shrinks the size of the produced histogram because instead of considering all the attributes' values we only need to involve the given ranges which are much fewer than the attributes' values. Since the proposed mechanism adapts the noise to the given workload's ranges, we can build and partition the histogram at the same time using the given queries' ranges instead of building the histogram and then partitioning it. The ranges of the given queries are extracted firstly, and then the ranges of each attribute are intersected to generate a set of disjoint ranges. This process ensures that there are no overlapping areas of the data distribution involved in the given workload, preventing using the anonymized responses to infer private information about any specific area. Thus, the disjoint ranges of each attribute are the actual partition of the dimension representing that attribute and each range represents an entry of the private histogram. An example of generating the private responses to a given workload W of two queries q_1, and q_2 over the tiny dataset shown in table II. The first query asks about the count of the patients whose age is between 15 and 50, with fewer than 6 pregnancies. The second query asks about the counts of the patients aged 40 years or more with more than 3 pregnancies. We would represent q_1 and q_2 as follows: $q_1 = \{[15, 50]; [0, 6]\}$, and $q_2 = \{[40, 84]; [4, 11]\}$. Based on these two queries, we have two ranges over each attribute. Over the age attribute we have the ranges $[15, 50]$ and $[40, 84]$. Therefore, after intersecting these ranges we now have three ranges $[15, 40]$, $[40, 50]$, and $[50, 84]$. Repeating the same process over the extracted ranges for the 'Number of Pregnancies' attribute would produce these three range $[0, 4]$, $[4, 6]$, and $[6, 11]$. The next step is to generate the partitioned histogram using these ranges. The histogram should consider all possible combinations of the attributes' values that fall within the generated ranges. Table III illustrates the partitioned histogram that has been built based on the intersected range of each attribute. To answer the given queries under differential privacy, the counts of the partitioned histogram need to be anonymized by injecting small amounts of noise to each count. Our mechanism generates the needed amounts of noise from the Laplace distribution to ensure differential privacy. After anonymizing the histogram frequencies, we can answer the first query by summing up the counts of record number 1, 2, 4, and 5. The second query also can be answered using the counts of the record number 5, 6, 8, and 9.

Complexity, Privacy, and Utility

Based on the previous example, our proposed model generated a histogram of 9 records to answer the given workload of two queries, since each attributes was partitioned into three buckets. On the other hand, the naive histogram [10] would have $\prod_{i=1}^{2} m_i = 70 * 11$ records in order to answer the same given workload over the same tiny dataset. Even if a data independent mechanism was used, the complexity would still be worse than our proposed mechanism and the partitioned histogram would have more than $\prod_{i=1}^{2} m_i = 7 * 6$ records, since it needs to be built based on the distribution of the underlying dataset.

Unlike data dependent partitioning mechanisms which consume the privacy budget in order to build the required data structure, our partitioning strategy does not require any information about the underlying dataset. Therefore, we only need to use the privacy budget while building the private histogram when retrieving the counts that represent its frequencies. That is, as long as this histogram is built, the private response of each given query can be retrieved from this histogram.

Our partitioning strategy provides more accurate estimated statistics than the data dependent partitioning strategies especially for the multiple attribute range queries because in the case of the data dependent partitioning, each dimension is partitioned according to the uniformity of its frequencies. Therefore, the estimated count of the involved dimensions

978-1-7281-5092-5/19/$31.00 ©2019 IEEE

would include the total of the estimated counts of the involved parts (buckets) of each dimension. Observe that the beginning or the end points of each given range may fall inside one of the partition buckets; thus the count corresponding to this range would involve extra noise corresponding to the unwanted values that have been included in that bucket. On the other hand, since our strategy partitions the data dimensions based on the given ranges, we are sure that only the exact frequencies will be involved in the generated response, which provides better utility.

V. Related Work

Since differential privacy was proposed, many differentially private algorithms have been developed to preserve output accuracy [10], [13]–[26]. However, these algorithms were proposed for the non-interactive approach. Moreover, the released data should be suitable for random numbers of queries and that requires considering the very sensitive situations when a query focuses on small areas of the sensitive dataset. Therefore, these algorithms guarantee the privacy of the released data by injecting amounts of noise to each part of the data histogram. As a result, the utility of the released data may be lost for some aggregated queries that involve large areas of the data histogram.

Many researchers consider the *interactive* setting to be more protective for individuals' privacy [24], [27]–[29], [29], [30]. Even though these proposed mechanisms generate their private releases based on the sensitivity of the given queries, they require expensive operations, and the usefulness of the provided answers can not be guaranteed for some kind of queries. Alnemari et al. [8] proposed an interactive mechanism that is able to identify the sensitive areas of the underlying dataset through the ranges of the given queries to insure individuals' privacy within these areas. Their results shows that this mechanism could overcome the accuracy issue of the non-interactive models while maintaining the privacy of the provided responses. However, this mechanism works over a histogram that represents the underlying dataset which makes adopting this mechanism for multi-dimensional datasets very costly. In this paper, we present our interactive mechanism that lowers the cost of handling multiple attribute counting queries over infrastructure data.

VI. Concluding Remarks

This paper presented and analyzed a differential privacy based scheme for protecting personally identifiable information in critical infrastructure data. As the described method does not depend on the underlying dataset, it provides better privacy guarantees than those provided by several existing data dependent partitioning schemes, especially for multiple attribute range queries.

Future work includes adapting the proposed scheme to handle other kinds of queries such as counting queries over categorical attributes. We also intend to investigate the effect of increased workload size on overall performance and range size while ensuring lower error rates. All of these improvements can make differential privacy viable in the protection of PII within critical infrastructure data.

Acknowledgment

Asma Alnemari acknowledges the support of Taif University, and the Ministry of Higher Education, Kingdom of Saudi Arabia. This paper is also based upon work partly supported by the National Science Foundation under Awards DGE-1433736 and DGE-1922169.

References

[1] US Department of Health & Human Services – Privacy Rule, "Standards for Privacy of Individually Identifiable Health Information; Final Rule," 2002. http://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/administrative/privacyrule/privrulepd.pdf, Accessed May 2, 2017.

[2] X. Jiang, S. Cheng, and L. Ohno-Machado, "Quantifying fine-grained privacy risk and representativeness in medical data," in *Proceedings of the 2011 Workshop on Data Mining for Medicine and Healthcare*, DMMH '11, (New York, NY, USA), pp. 64–67, ACM, 2011.

[3] P. Shi, L. Xiong, and B. C. Fung, "Anonymizing data with quasi-sensitive attribute values," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, (New York, NY, USA), pp. 1389–1392, ACM, 2010.

[4] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[5] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Discov. Data*, vol. 1, Mar. 2007.

[6] C. Dwork, "Differential privacy," in *Automata, Languages and Programming* (M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, eds.), Springer Verlag, 2006.

[7] F. K. Dankar and K. El Emam, "The application of differential privacy to health data," in *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, EDBT-ICDT '12, (New York, NY, USA), pp. 158–166, ACM, 2012.

[8] A. Alnemari, C. J. Romanowski, and R. K. Raj, "An adaptive differential privacy algorithm for range queries over healthcare data," in *Healthcare Informatics (ICHI), 2017 IEEE International Conference on*, (Park City, UTAH, USA), pp. 397–402, IEEE, Aug. 2017.

[9] K. Fernandes, J. S. Cardoso, and J. Fernandes, "Transfer learning with partial observability applied to cervical cancer screening," in *Iberian conference on pattern recognition and image analysis*, pp. 243–250, Springer, 2017.

[10] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference*, pp. 265–284, Springer, 2006.

[11] S. Vadhan, "The complexity of differential privacy," in *Tutorials on the Foundations of Cryptography*, pp. 347–450, Springer, 2017.

[12] M. Bun, J. Ullman, and S. Vadhan, "Fingerprinting codes and the price of approximate differential privacy," in *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pp. 1–10, ACM, 2014.

[13] C. Dwork, A. Roth, *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2014.

[14] Y. Xiao, L. Xiong, and C. Yuan, "Differentially private data release through multidimensional partitioning," in *Workshop on Secure Data Management*, pp. 150–168, Springer, 2010.

[15] T.-H. H. Chan, E. Shi, and D. Song, "Private and continual release of statistics," *ACM Transactions on Information and System Security (TISSEC)*, vol. 14, no. 3, p. 26, 2011.

[16] C. Dwork, M. Naor, O. Reingold, and G. N. Rothblum, "Pure differential privacy for rectangle queries via private partitions," in *Proceedings, Part II, of the 21st International Conference on Advances in Cryptology — ASIACRYPT 2015 - Volume 9453*, (Berlin, Heidelberg), pp. 735–751, Springer-Verlag, 2015.

[17] G. Acs, C. Castelluccia, and R. Chen, "Differentially private histogram publishing through lossy compression," in *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pp. 1–10, IEEE, 2012.

[18] X. Xiao, G. Wang, and J. Gehrke, "Differential privacy via wavelet transforms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, pp. 1200–1214, Aug 2011.

978-1-7281-5092-5/19/$31.00 ©2019 IEEE

[19] J. Xu, Z. Zhang, X. Xiao, Y. Yang, G. Yu, and M. Winslett, "Differentially private histogram publication," *The VLDB Journal*, vol. 22, no. 6, pp. 797–822, 2013.

[20] G. Cormode, C. Procopiuc, D. Srivastava, E. Shen, and T. Yu, "Differentially private spatial decompositions," in *Data engineering (ICDE), 2012 IEEE 28th international conference on*, pp. 20–31, IEEE, 2012.

[21] Y. Xiao, J. Gardner, and L. Xiong, "Dpcube: Releasing differentially private data cubes for health information," in *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pp. 1305–1308, IEEE, 2012.

[22] A. Inan, M. Kantarcioglu, G. Ghinita, and E. Bertino, "Private record matching using differential privacy," in *Proceedings of the 13th International Conference on Extending Database Technology*, pp. 123–134, ACM, 2010.

[23] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar, "Privacy, accuracy, and consistency too: a holistic solution to contingency table release," in *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 273–282, ACM, 2007.

[24] M. Hay, V. Rastogi, G. Miklau, and D. Suciu, "Boosting the accuracy of differentially private histograms through consistency," *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 1021–1032, 2010.

[25] A. Blum, K. Ligett, and A. Roth, "A learning theory approach to noninteractive database privacy," *Journal of the ACM (JACM)*, vol. 60, no. 2, p. 12, 2013.

[26] D. Feldman, A. Fiat, H. Kaplan, and K. Nissim, "Private coresets," in *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 361–370, ACM, 2009.

[27] M. Hardt and K. Talwar, "On the geometry of differential privacy," in *Proceedings of the forty-second ACM symposium on Theory of computing*, pp. 705–714, ACM, 2010.

[28] C. Li, M. Hay, V. Rastogi, G. Miklau, and A. McGregor, "Optimizing linear counting queries under differential privacy," in *Proceedings of the Twenty-ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '10, (New York, NY, USA), pp. 123–134, ACM, 2010.

[29] M. Hardt, K. Ligett, and F. McSherry, "A simple and practical algorithm for differentially private data release," in *Advances in Neural Information Processing Systems*, pp. 2339–2347, 2012.

[30] C. Li, M. Hay, G. Miklau, and Y. Wang, "A data- and workload-aware algorithm for range queries under differential privacy," *Proc. VLDB Endow.*, vol. 7, pp. 341–352, Jan. 2014.

978-1-7281-5092-5/19/$31.00 ©2019 IEEE