Shifting Dataset To Preserve Data Privacy

Muhammad Syafiq Mohd Pozi[b,c], Asmidar Abu Bakar[a,b], Roslan Ismail[a,b], Salman Yussof[a,b], Fiza Abdul Rahim[a,b], and Ramona Ramli[a,b]

[a]College of Computer Science and Information Technology, Universiti Tenaga Nasional, Kajang, Malaysia
[b]Institute of Informatics and Computing in Energy (IICE), Universiti Tenaga Nasional, Kajang, Malaysia
[c]School of Computing, Universiti Utara Malaysia, 06010, UUM Sintok, Kedah, Malaysia
[a,b]{*syafiq.pozi, asmidar, roslan, salman, fiza, ramona*}*@uniten.edu.my*
[c]*syafiq.pozi@uum.edu.my*

***Abstract*—Data analytic is very valuable in any domain that produces large amount of data making demands on full datasets to be revealed for analytic purposes are rising. Regardless, the privacy of the released dataset should be preserved. New techniques using synthetic data as a mean to preserve the privacy has been identified as appropriate approach to fulfill the demand. In this paper, a privacy-preserving data synthetic framework for data analytic is proposed. Using a generative model that captures the density function of data attributes, the privacy-preserving synthetic data is produced. We performed classification task through various machine learning classifiers in measuring the data utility of the new privacy-preserving synthesized data.**

***Index Terms*—Privacy Preservation, Data Mining, Kernel Density Estimation, Dataset Shift**

I. Introduction

THE privacy law for individual customer, such as Malaysia Personal Data Protection Act 2010 [1], as well as in other nations [2], [3], has enforced organizations to act accordingly in order to ensure any kind of data processing should not be used to directly or indirectly invading user privacy. With rapid advancement in statistic and machine learning, as well as their large number of optimized software that are readily available, privacy preservation task such as de-identification can be easily broken. Variations of inference attacks, such as on Netflix dataset [4], location data [5] and social networks data [6], shows that simple modification of sensitive data by removing identifiers or by generalizing or suppressing data features can result in major information leakage and cannot guarantee meaningful privacy for data owners.

Proposing a suitable privacy preservation model is not a trivial task. One has to consider many aspects such as privacy specification [7], i.e. the kind of data need to be masked, computing performance [8] especially on large continuous data, latency [9] in which the data transaction occurs between two parties e.g. customer and utility provider and of course, security [10] such as the risk of data forging or denial of service attacks. Currently, all techniques that have been proposed to preserve the data privacy can be grouped into three categories, such as follows:

1) **Cryptography**. Public key infrastructure [11] has been used extensively to provide simple encrypted communication, such as data authentication between utility provider and utility customer. For example, the utility provider can formalize a clearance level model according to the customer status. However, cryptography requires significant computing time. As a result, the data transmission time to the endpoint will also increase.
2) **Statistical interference**, which includes data anonymization [12], data randomization [13], and/or data perturbation [14], in which these processes are used to improve the generalization property, thus, increasing the difficulty of performing microanalytic of particular data. However, statistical interference must be modeled properly. Too much generalization on dataset will hinder the process of obtaining the useful analytic model. Hence, it is important to ensure the process present very minimal impact to the whole series of data in term of probabilistic perspective.
3) **User intervention**. In this case, end user directly interferes with the data collection process [15], [16]. For example, in electric reading, user illegally tampered the electric meter for malicious purpose, such as for reducing billing cost.

Data owner can protect the data privacy through rigorous privacy definitions, such as differential privacy [17]. However, this could reduce the data utility in extracting the useful information. For example, differential privacy is limited to interactive count queries on statistical databases [18], and not the full records. In non-interactive setting, where there is a middleman that facilitate the data exchange between data owner and data user for releasing generic data, these mechanisms is simply not feasible, especially on large amount of high dimensional dataset.

Demands on full datasets are increasing, yet the privacy of the released dataset should be preserved. Hence, many researchers proposed to use synthetic data as a mean to preserve the data privacy [19]. Privacy-preserving synthetic data not only can be used by targeted data user, but can also be used in educational purpose in various sectors, and for data application development such as developing advanced machine learning and pattern recognition model.

Thus, we are proposing a privacy-preserving data synthetic generator framework for data analytic, without downgrading the utility of the synthesized data. Our framework consists of three main tasks: estimating the density function for a given dataset, generating new synthetic dataset and finally utilizing the new synthetic dataset. We measured the data utility of the

978-1-5386-7263-1/18/$31.00 ©2018 IEEE

synthetic data based on performing classification task through various machine learning classifiers.

However, our work is unique, such that, instead of directly sampling new records from existing distribution function [20], we project the original data into other distribution function, and perform the sampling on that new distribution function. Unlike only generating partial synthetic data such as in [21], we managed to generate new synthetic data with similar data utility.

This paper is organized as follows: Section II discusses recently proposed privacy preservation techniques in literature. Section III describes our privacy preservation techniques, that mainly derived from statistical model. Section IV outlines our experimental results. Then, Section VI discusses the limitation and how to address the limitation of the proposed privacy-preservation model. Finally, Section VII concludes this paper, together several future works that can be done to extend this privacy-preservation model.

II. Privacy Preservation Techniques

Data analytic is very useful in any domain that produces large amount of data, for example in power distribution, in which can be used to solve many problems, such as in fraud detection system [22], system balancing and transmission on power flow network [23], pricing and billing [24], measuring voltage and power quality [25], as well as outage and fault detection [26]. Data analytic also can be used for illegitimate purpose, if, the data have been exposed or leaked to the irresponsible parties. Therefore, preserving the data privacy is a critical task especially in preventing known or unknown malicious activities in the future.

Hence, besides cryptography, which is computationally expensive, and has almost zero data utility value, one can derived a statistical model to improve the privacy preservation on the data. In [27], a low sampling rate method has been proposed to preserve the data privacy without reducing anomaly detection accuracy. However, even though, anomaly detection is a very important task, it is actually a rare event. Thus, low sampling rate could provide very imbalanced data, which might not suit properly with most of the analytic tasks.

Improving the data generalization based on k-anonymity [28], such as for visualization task [29] or monitoring task [30] has an additional complexity, where prior knowledge of what need to be hidden must be carefully predefined. Again, improper data generalization through k-anonymity metric alone could result introduce bias to the dataset, which makes pattern recognition task increasingly hard.

Another method is data perturbation, where the data is distorted in order to mask the original distribution, such as in [31]. However, data perturbation only make sense if the distortion can be properly defined, because carelessly performing data distortion will make any relation exists in the data will be disturbed, hence resulting nonsensical pattern being learned by data modeler. Regardless, there are many well-studied and sophisticated machine learning algorithms that can identify any distorted data, in term of anomalies, novelties or outliers detection tasks [32].

Finally, data randomization involve data swapping between records or fields or both [33]. Again, this broke the structure of data, and also computationally expensive, depending on the choice of data randomization algorithms. Misguided data will just increase the training time in order to come with sensible model on that data. Furthermore, when doing data randomization, one need to track the randomization process, in order to recover the data back to the original form.

Regardless, these techniques, which can be grouped as *syntactic privacy protection*; their success rate depends on the knowledge of the adversaries. Again, these process can significantly reduce the data utility if these techniques are combined together to achieve the privacy metric, as some of the data attributes are sometime suppressed to achieve that.

Another group of approaches that has gain some popularity in literature for preserving the data privacy is through *synthetic generator mechanism* [34]. This approach is used with respect to its input data, in order to produce privacy-preserving synthetic data. Unlike syntactic privacy protection, synthetic generator mechanism keeps the data in original format. This allow the data to be used by several applications or programming code, especially anything that require processing on raw data, which obviously can't be achieved by excessive data sanitation.

Nevertheless, even though synthetic generator mechanism provide almost similar level of data utility with original dataset, the mechanism of producing those data is less understood. Furthermore, syntactic privacy protection provides clear and tangible privacy preservation effect compared to synthetic generator mechanism, at the cost of data utility, in significant way. This is is the reason why syntactic privacy protection is widely used in smart meter application and various other domains that require privacy-preserving task.

III. Generative Model

In this section, we describe our generative model in synthesizing new privacy preserving synthetic data. Our synthesizer is a probabilistic model that capture the density function of data attributes with privacy-preservation in mind. The synthesizer must learn from the original data in order to produce new synthetic data.

Let dataset $X = \{x_1^d, x_2^d, \ldots, x_n^d\}$ consists of n records with d random variables associated with the records' attributes, be the input vectors in input space. Our generative model, should map X into new privacy-preserving dataset X' with respect to X. Then, X' will be the one to be utilized by the data user.

If the new record $\boldsymbol{x}'$ is produced based on generative model, and achieve the desired privacy properties in this case unlinkability hence, it proves that $\boldsymbol{x}'$ meet the privacy requirements in the required process. If $\boldsymbol{x} \neq \boldsymbol{x}'$, then unlinkability is achieved. In this case, unlinkability means lack of ability to determine the source of $\boldsymbol{x}'$ is from $\boldsymbol{x}$.

A. Distribution Shifting

Most of the time, dataset X comes with a target variables Y, usually in form of a column matrix, $\{y_1, y_2, \ldots, y_n\}$. Each

value in Y correspond to attributes in X, in respected order. The joint distribution between $\boldsymbol{x}$ and y is defined as follows:

$$P(y, \boldsymbol{x}) = P(y|\boldsymbol{x})P(\boldsymbol{x}) \tag{1}$$

The goal is to estimate $P(\boldsymbol{x})$, then shift it into $P_{new}(\boldsymbol{x})$, before generating new synthetic dataset from $P_{new}(\boldsymbol{x})$.

In this work, we assume every $x \in \boldsymbol{x}$ is independent to each other. Let $x_1, x_2, \ldots, x_n$ be several records drawn from a common distribution described by the density function f. The kernel density estimate is defined such as follows:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big) \tag{2}$$

where

- x is a single data point;
- $\boldsymbol{h}$ is the bandwidth, also known as smoothing parameter;
- K is the kernel function;

The choice of K is not crucial. In this paper, we opt for Gaussian kernel, defined as follows:

$$K(x; h) \propto \exp\Big(\frac{x^2}{2h^2}\Big) \tag{3}$$

The inferred distribution then is shifted to another distribution such that:

$$f_{new}(\boldsymbol{x}) = z\hat{f}_h(\boldsymbol{x}) \tag{4}$$

where z is the user-defined weights. As data usually comes with several unique target variables, $\{y_1, y_2, \ldots, y_n\}$, hence, there will be n kernel density estimate models.

In this paper, we implement z as a predefined weight of each sample, directly to each feature of that sample, according to n different kernel density estimators, which is the number of classes of that dataset. Algorithm 1 describes how the dataset is being prepared before being estimated by kernel density estimator in Algorithm 2.

Algorithm 1 Dataset Preparation

procedure SHIFT($\boldsymbol{X}$)
Require:
Dataset $\boldsymbol{X}$ with label $\boldsymbol{Y}$.
  $classes \leftarrow \{\boldsymbol{Y}\}$
  $\boldsymbol{X'} \leftarrow \emptyset$
  for $y \in classes$ **do**
    $\boldsymbol{x'} \leftarrow \emptyset$
    for $\boldsymbol{x} \in \boldsymbol{X}$ **do**
      if $y \in (\boldsymbol{x}, y)$ **then**
        $\boldsymbol{x'} \leftarrow z.\boldsymbol{x}$
        $\boldsymbol{X'} \leftarrow \boldsymbol{x'}$
      end if
    end for
  end for
end procedure

Algorithm 2 Density Estimation

procedure KDE($\boldsymbol{X'}$)
Require:
Dataset $\boldsymbol{X'}$ with label $\boldsymbol{Y}$.
  $classes \leftarrow \{\boldsymbol{Y}\}$
  for $y \in classes$ **do**
    $\boldsymbol{X'_y} \leftarrow \boldsymbol{X'} \cap \boldsymbol{X'}_{\neg y}$
    $F_y \leftarrow model(\boldsymbol{X'_y})$ ▷ Equation 2
  end for
end procedure

B. Data Synthesization

Once the probability distribution has been approximately inferred, new data $\boldsymbol{x}$ will be generated through resampling over the weighted samples by the new distribution model, such in Equation 4. Hence, the new synthetic dataset will be different compared to the original samples.

1) Given X, multiply each $x \in \boldsymbol{x}$, for every $\boldsymbol{x} \in X$ with z, as described in Algorithm 1.
2) Estimate the new probability density function with kernel density estimator, as described in Algorithm 2.
3) Generate new dataset X' using the new probability density function. This process is simply to get another data point that lies within the density function. Total of generated data will be similar to the original data. The data however is not similar to the original data.

Finally, the new X' will be used by data user for various data analytic applications, such as clustering, regression or classification tasks. In this paper, we limit our scope only to classification task.

C. Classification Model

In this paper, we will be analyzing the utility of our synthetic data based on classification task. The classification model, which combined together with kernel density estimator and data synthetic are illustrated in Figure 1.

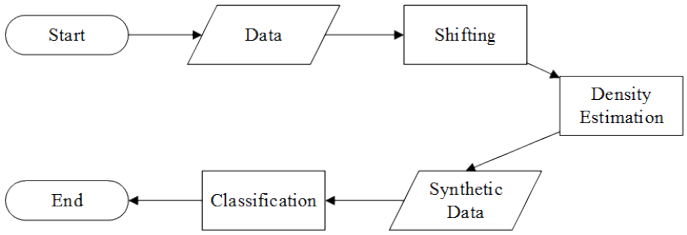

Fig. 1. The classification model, starting from shifting the dataset until the classification process.

Figure 1 shows the sequence of classification task using synthetic data. The classification task can be divided into two stage. The first stage is training task, in which the classification model is trained using existing data. The second stage is the prediction task. For any unseen data, the task is to classify which class the unseen data belong to. Both stages will undergo the same process illustrated in Figure 1. However, for prediction task, process is a little bit different compared to training task. The prediction task is illustrated in Algorithm 3.

Algorithm 3 Prediction Task

1: **procedure** PREDICT(x)
Require:
Record x.
2: $x' \leftarrow z.x$
3: $F \leftarrow estimate(x')$
4: $\{x_1, x_2, x_3, \ldots, x_n\} \leftarrow generatePoint(F) \setminus x'$
5: $x'' \leftarrow pickOne(\{x_1, x_2, x_3, \ldots, x_n\})$ ▷ Randomly
6: $\hat{y} \leftarrow f(x'')$ ▷ Prediction
7: **end procedure**

In Algorithm 3, the prediction will be performed on synthetic data, estimated by the predefined KDE models which is obtained from the training task in the first stage. Hence, any prediction task will always be accompanied by the KDE models such as defined in Algorithm 2, including the z value defined in Algorithm 1.

IV. Experimentation and Analysis

The original and synthetic data is evaluated through classification task. Table I describes the data used in this experiment. The datasets were retrieved from UCI machine learning repository [35]. The dataset are **Fertility** Diagnosis [36], **Iris**, **Haberman**, **Liver Disorder**, **Breast Cancer**, **Pima** Indians Diabetes, **Thyroid** and **Bank Marketing** [37]. Prior to classification task, each dataset is being preprocessed such that any non-numeric attributes are transformed into numeric attributes, through unique mapping, starting from 0 to n unique non-numeric values.

TABLE I
DATASET USED IN THIS EXPERIMENT

Dataset	# Records	# Attributes	# Classes
Fertility	100	10	2
Iris	150	4	3
Haberman	306	3	2
Liver Disorder	345	7	2
Breast Cancer	699	10	2
Pima	768	8	2
Thyroid	3,772	21	2
Bank Marketing	45,211	17	2

The classification experimentation is evaluated through 10-fold cross-validation method. We used common classifiers such as follows: Naive Bayes [38], Support Vector Machine [39], and Decision Tree [40]. The parameters of the experiment are defined as follows:

1) z values for each dataset is defined such that the deviation between classes is not obvious, but in the same time, to make the new dataset more diverge from the original dataset. Each $\boldsymbol{x}$ will be weighted accordingly to its class, by z, such as follows:
   - $z_1 = log_{10}(2.00001)$ for class 1.
   - $z_2 = log_{10}(2.00002)$ for class 2.
   - $z_3 = log_{10}(2.00003)$ for class 3.

   Since we only deal up to 3 classes, hence, we only specify values for three different z.
2) Support Vector Machines parameters:
   - RBF Kernel = $K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}\right)$
   - Kernel parameter, $\sigma = 0.01$
   - Regularization parameter $C = 64$

Each classifier is based on WEKA implementation [41]. Finally, Table II shows the experimentation results for each dataset, original versus synthetic data across all classifiers, based on accuracy defined in Equation 5.

$$accuracy(y, \hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1} 1(\hat{y}_i = y_i) \times 100 \qquad (5)$$

In Equation 5, y is the original target value, $\hat{y}$ is the predicted target value and n is the number of samples. We only count the number of $\hat{y}_i = y_i$, and divide it by the number of samples before being multiplied by 100 to retrieve the percentage value.

TABLE II
CLASSIFICATION RESULTS BETWEEN ORIGINAL DATA AND SYNTHETIC DATA THROUGH NAIVE BAYES (NB), SUPPORT VECTOR MACHINES (SVM) AND DECISION TREE (DT) CLASSIFIERS IN TERM OF ACCURACY PERCENTAGE.

Dataset	Type	NB	SVM	DT
Fertility	Original	88.00	88.00	85.00
	Synthetic	88.00	88.00	86.00
Iris	Original	96.00	92.67	96.00
	Synthetic	91.95	93.96	88.59
Haberman	Original	74.51	73.20	72.88
	Synthetic	76.47	73.53	75.82
Liver Disorder	Original	55.36	57.97	68.70
	Synthetic	59.71	57.97	57.68
Breast Cancer	Original	95.99	97.13	94.56
	Synthetic	95.42	96.42	93.99
Pima	Original	76.63	77.21	73.83
	Synthetic	77.34	73.82	76.04
Thyroid	Original	88.46	93.95	98.86
	Synthetic	99.01	93.87	93.18
Bank Marketing	Original	85.81	89.34	89.81
	Synthetic	99.78	97.50	99.78

From the experimentation, we obtain mixed results, some of the synthetic dataset has lower accuracy values, while other maintain or achieve higher compared to the original dataset. Regardless, based on statistical test using Wilcoxon Signed-Rank Test [42], irregardless of the classifiers, the z-value is -0.9211. The p-value is 0.35758. Hence, the result is not significant at $p \leq 0.05$. Thus, fail to reject null hypothesis. Therefore, the data utility is maintained between original and synthetic dataset.

Bank Marketing dataset provides the most appreciation of accuracy values from original dataset to synthetic dataset. There are two complementary possibilities in this result. The first one, our weighting mechanism could introduce significant bias between two classes, in which classifier can identify easily which one belong to which class. The second one is the dataset attributes could be heavily independent to each

other, hence one small change in the attribute can affect the overall classification performance, while in the same time, has no effect on other attributes.

This is also true for otherwise, such that the classification performance in synthetic data is degraded compared to the original dataset. This is observed in classification performance of **Iris**, **Liver Disorder**, **Breast Cancer**, **Pima** and **Thyroid**. The weighting mechanism might not applicable to the dataset as there probably a possibility of existing joint distribution between the dataset attributes. Therefore, when we perform the shifting process, the variance of the overall dataset is significantly increased. High variance can cause a learning algorithm to model the random noise in the training data, rather than the intended outputs. As a result, the synthetic datasets gives lower classification results compared to when doing classification task on original dataset.

Support Vector Machine provides the most consistent classification accuracy performance between original and synthetic data compared to other classifiers. Briefly, this is because Support Vector Machine task is to find the classification model that has good generalization property, compared to Naive Bayes and Decision Tree where these classifiers only try to find model that has less misclassification results.

V. Privacy Properties Analysis

In this paper, we consider 2 main privacy properties, which are *Unlinkability*, and *Deniability* [43], [19]. These properties are defined as follows:

1) *Deniability* provides a mechanism for users to deny their involvement in using certain resources.
2) *Unlinkability* ensures that a user may make multiple uses of resources or services without others being able to link these uses together.

In this paper, we assume that a randomized algorithm A satisfies unbounded ϵ-differential privacy if the relation $P(A(D_1) \in S) \leq e^{\epsilon} P(A(D_2) \in S)$ for any set S and any pairs of databases D_1, D_2, where D_1 can be obtained from D_2 by changing the value of exactly one tuple. Therefore:

1) Since we are producing similar number of records as the original dataset, then $P(A(D_1) \in S) = 1$ and D_1 is the whole S.
2) If D_1 is the original dataset and D_2 is the synthetic dataset, then, it should be $P(A(D_2) \in S) \leq e^{\epsilon} P(A(D_1) \in S)$. In this case, $e^{\epsilon} = 1$.

Hence, not only the shifted synthetic dataset is totally different compared to the original dataset, but also the composition of independent and dependent features are also changed.

VI. Discussion and Limitations

Table II provides interesting results, where classification results were increasing accurately when using synthetic data. This is because, our weighting scheme can sometime provide bias to some of the dataset attributes, which could help classifiers in discriminating each class properly, or otherwise.

Overall, the classification results tabulated in Table II shows that our privacy-preservation model can be used to preserve the data utility. However, there are several limitations, in which can be addressed properly. Currently, our model only work on numerical data. Hence, in order to apply in on non-numeric data, some sort of data transformation, from non-numeric to numeric data need to be performed.

Furthermore, as previously stated, the weighting mechanism could introduce unnecessary bias to the dataset attributes. Our suggestion is to properly normalize the data, so the bias effect can be minimized. For example, in datasets **Pima**, **Haberman**, and **Bank Marketing**, some of the attributes have very large mean and standard deviation values, while other attributes have very low mean and standard deviation values.

Moreover, regarding to the weighting issue, our framework could pose a problem if we have large number of classes. For example, if we have 10 classes, the weight of class 1 and class 10 will result a total bias to the whole dataset. Thus, adversary can directly identify that the data might somehow systematically modified, though it is hard to relink back to the original data. Regardless, one can just use similar weight value across all the classes to avoid this problem.

This paper also assumes that each data attribute is independent to each other. This is hardly the case in real world, as most of the time there must be some correlation between two or more attributes. Hence, even though we managed to preserve the data utility, it still makes no sense if two random variables with synthetic values, for example, such as height and gender has negative correlation.

Another drawback of this method is the increasingly computational complexity, at $\mathcal{O}(m^n)$, especially during the training stage. In this case, m is the total number of records in the dataset while n is the number of records from the dataset that will be used to generate new synthetic records. Hence, when the data size is increasing, the number of records from the dataset that need to be sampled is also increasing. Thus, the time to complete the generation process will also increase significantly, at the power of n.

VII. Conclusion and Future Works

In this paper, we proposed a new shifted induced dataset mechanism for preserving the data privacy, based on three different procedures: estimating distribution function of multi-dimensional data, generating new synthetic data from shifted distribution function, in which both of them are combined together and evaluated through task. Through our experimentation, we demonstrated that our method is capable in maintaining data utility while preserving the privacy of data in classification task.

There are many things that can be achieved for future work. The first thing is to improve the data generation mechanism with more sophisticated techniques. For example, properly designing weighting model to reduce the bias across dataset attributes as well to each record in the dataset. More importantly, a proper weighting mechanism must solve the problem for dataset that has large number of classes. Care also must be taken when dealing with various kind of data that is not numeric, or if the data should not be in decimal values.

Secondly, a distribution model that capture the joint distribution between the attributes itself must be designed properly, in

order to make more sensible synthetic data. If not, the variance of the data might be increased, hence reducing the overall data utility, either it is for classification task, regression task or clustering task.

Finally, instead of doing retraining, it will be better to use the existing training model that is modeled based on the synthetic data, and use it on the original data. Designing flexible learning mechanism, such that it can be generalized over two similar structured datasets, with each dataset comes from different probability distribution function, will significantly decrease the time to perform any analytic task.

ACKNOWLEDGEMENT

The work was supported by TNB Seed Fund U-TD-RD-17-02.

REFERENCES

[1] "Act 709: Personal data protection act 2010," *Laws of Malaysia*, pp. 1–95, 2010.

[2] C. Parliament, "Personal information protection and electronic documents act," *Consolidated Acts, SC 2000, c*, vol. 5, p. 13, 2000.

[3] P. Carey, *Data protection: a practical guide to UK and EU law*. Oxford University Press, Inc., 2009.

[4] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. IEEE, 2008, pp. 111–125.

[5] P. Golle and K. Partridge, "On the anonymity of home/work location pairs," in *International Conference on Pervasive Computing*. Springer, 2009, pp. 390–397.

[6] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *Security and Privacy, 2009 30th IEEE Symposium on*. IEEE, 2009, pp. 173–187.

[7] Z. Zhang, Z. Qin, L. Zhu, J. Weng, and K. Ren, "Cost-friendly differential privacy for smart meters: exploiting the dual roles of the noise," *IEEE Transactions on Smart Grid*, vol. 8, no. 2, pp. 619–626, 2017.

[8] Z. Fu, F. Huang, K. Ren, J. Weng, and C. Wang, "Privacy-preserving smart semantic search based on conceptual graphs over encrypted outsourced data," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 8, pp. 1874–1884, 2017.

[9] H. Li, G. Dán, and K. Nahrstedt, "Portunes+: Privacy-preserving fast authentication for dynamic electric vehicle charging," *IEEE Transactions on Smart Grid*, 2017.

[10] A. Abdallah and X. Shen, "Lightweight security and privacy preserving scheme for smart grid customer-side networks," *IEEE Transactions on Smart Grid*, vol. 8, no. 3, pp. 1064–1074, 2017.

[11] A. Mohammadali, M. S. Haghighi, M. H. Tadayon, and A. M. Nodooshan, "A novel identity-based key establishment method for advanced metering infrastructure in smart grid," *IEEE Transactions on Smart Grid*, 2016.

[12] W. Yang, N. Li, Y. Qi, W. Qardaji, S. McLaughlin, and P. McDaniel, "Minimizing private data disclosures in the smart grid," in *Proceedings of the 2012 ACM conference on Computer and communications security*. ACM, 2012, pp. 415–427.

[13] E. Liu and P. Cheng, "Achieving privacy protection using distributed load scheduling: A randomized approach," *IEEE Transactions on Smart Grid*, 2017.

[14] S. Bhela, V. Kekatos, and S. Veeramachaneni, "Enhancing observability in distribution grids using smart meter data," *IEEE Transactions on Smart Grid*, 2017.

[15] G. Giaconi, D. Gunduz, and H. V. Poor, "Smart meter privacy with renewable energy and a storage device," *arXiv preprint arXiv:1703.08390*, 2017.

[16] G. Giaconi, D. Gündüz, and H. V. Poor, "Optimal demand-side management for joint privacy-cost optimization with energy storage," *CoRR*, vol. abs/1704.07615, 2017. [Online]. Available: http://arxiv.org/abs/1704.07615

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

[18] C. Dwork, "Differential privacy: A survey of results," in *International Conference on Theory and Applications of Models of Computation*. Springer, 2008, pp. 1–19.

[19] V. Bindschaedler, R. Shokri, and C. A. Gunter, "Plausible deniability for privacy-preserving data synthesis," *Proceedings of the VLDB Endowment*, vol. 10, no. 5, pp. 481–492, 2017.

[20] S. K. Kinney, J. P. Reiter, A. P. Reznek, J. Miranda, R. S. Jarmin, and J. M. Abowd, "Towards unrestricted public use business microdata: The synthetic longitudinal business database," *International Statistical Review*, vol. 79, no. 3, pp. 362–384, 2011.

[21] B. Loong, A. M. Zaslavsky, Y. He, and D. P. Harrington, "Disclosure control using partially synthetic data for large-scale health surveys, with applications to cancors," *Statistics in medicine*, vol. 32, no. 24, pp. 4139–4161, 2013.

[22] M. Zanetti, E. Jamhour, M. Pellenz, M. Penna, V. Zambenedetti, and I. Chueiri, "A tunable fraud detection system for advanced metering infrastructure using short-lived patterns," *IEEE Transactions on Smart Grid*, 2017.

[23] P. H. Nguyen, N. Blaauwbroek, C. Nguyen, X. Zhang, A. Flueck, and X. Wang, "Interfacing applications for uncertainty reduction in smart energy systems utilizing distributed intelligence," *Renewable and Sustainable Energy Reviews*, vol. 80, pp. 1312–1320, 2017.

[24] Z. Wang and R. Paranjape, "Optimal residential demand response for multiple heterogeneous homes with real-time price prediction in a multiagent framework," *IEEE transactions on smart grid*, vol. 8, no. 3, pp. 1173–1184, 2017.

[25] M. M. Albu, M. Sănduleac, and C. Stănescu, "Syncretic use of smart meters for power quality monitoring in emerging networks," *IEEE Transactions on Smart Grid*, vol. 8, no. 1, pp. 485–492, 2017.

[26] F. C. Trindade and W. Freitas, "Low voltage zones to support fault location in distribution systems with smart meters," *IEEE Transactions on Smart Grid*, 2017.

[27] P. Jokar, N. Arianpoo, and V. C. Leung, "Electricity theft detection in ami using customers consumption patterns," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 216–226, 2016.

[28] 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.

[29] X. Wang, J.-K. Chou, W. Chen, H. Guan, W. Chen, T. Lao, and K.-L. Ma, "A utility-aware visual approach for anonymizing multi-attribute tabular data," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 351–360, 2018.

[30] Y. Hong, W. M. Liu, and L. Wang, "Privacy preserving smart meter streaming against information leakage of appliance status," *IEEE Transactions on Information Forensics and Security*, 2017.

[31] G. Giaconi, D. Gündüz, and H. V. Poor, "Smart meter privacy with renewable energy and an energy storage device," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 1, pp. 129–142, 2018.

[32] E. Hou, K. Sricharan, and A. O. Hero, "Latent laplacian maximum entropy discrimination for detection of high-utility anomalies," *IEEE Transactions on Information Forensics and Security*, 2018.

[33] M. A. Rahman, M. H. Manshaei, E. Al-Shaer, and M. Shehab, "Secure and private data aggregation for energy consumption scheduling in smart grids," *IEEE Transactions on Dependable and Secure Computing*, vol. 14, no. 2, pp. 221–234, 2017.

[34] X. Zhang, S. Ji, and T. Wang, "Differentially private releasing via deep generative model," *arXiv preprint arXiv:1801.01594*, 2018.

[35] K. Bache and M. Lichman, "Uci machine learning repository," 2013.

[36] D. Gil, J. L. Girela, J. De Juan, M. J. Gomez-Torres, and M. Johnsson, "Predicting seminal quality with artificial intelligence methods," *Expert Systems with Applications*, vol. 39, no. 16, pp. 12 564–12 573, 2012.

[37] S. Moro, R. Laureano, and P. Cortez, "Using data mining for bank direct marketing: An application of the crisp-dm methodology," in *Proceedings of European Simulation and Modelling Conference-ESM'2011*. Eurosis, 2011, pp. 117–121.

[38] D. D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *European conference on machine learning*. Springer, 1998, pp. 4–15.

[39] M. M. Adankon and M. Cheriet, "Support vector machine," in *Encyclopedia of biometrics*. Springer, 2009, pp. 1303–1308.

[40] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[41] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[42] R. Woolson, "Wilcoxon signed-rank test," *Wiley encyclopedia of clinical trials*, 2008.

[43] A. Pfitzmann and M. Hansen, "A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management," 2010.