On the Generation of Privatized Synthetic Data Using Distance Transforms

Kato Mivule
Bowie State University
Bowie, MD, USA
kmivule@gmail.com

Abstract—**Organizations have interest in research collaboration efforts that involve data sharing with peers. However, such partnerships often come with confidentiality risks that could involve insider attacks and untrustworthy collaborators who might leak sensitive information. To mitigate such data sharing vulnerabilities, entities share privatized data with retracted sensitive information. However, while such data sets might offer some assurances of privacy, maintaining the statistical traits of the original data, is often problematic, leading to poor data usability. Therefore, in this paper, a confidential synthetic data generation heuristic, that employs a combination of data privacy and distance transforms techniques, is presented. The heuristic is used for the generation of privatized numeric synthetic data, while preserving the statistical traits of the original data. Empirical results from applying unsupervised learning, using k-means, to test the usability of the privatized synthetic data set, are presented. Preliminary results from this implementation show that it might be possible to generate privatized synthetic data sets, with the same statistical morphological structure as the original, using data privacy and distance transforms methods**.

Keywords-Privatized synthetic data generation; Data privacy; Distance transforms; k-means clustering

I. INTRODUCTION

Research collaboration among organizations often involves the sharing of data, however, the issue of data confidentiality is often an impediment in such partnerships. To safely engage in joint research ventures, entities often retract sensitive and private information from the shared data, which reduces usability, despite confidentiality assurances. Yet still, another method used to address such data sharing vulnerabilities, is to generate privatized synthetic data sets that retain the statistical traits of the original data while at the same time ensuring privacy. In this paper, we present a confidential synthetic data generation heuristic, that employs data privacy and distance transforms methods, for the generation of privatized synthetic data while maintaining some of the statistical traits of the original data. In the initial stage, we apply distance transforms to extract the coefficients with the needed traits, from the original data (noisy data in this case), we then add the coefficients to a noisy data set, generating a privatized synthetic data set. Filtering is then applied to the privatized synthetic data set, to reduce noise and enhance usability. We then apply unsupervised learning, using k-means clustering, to test the usability of the privatized synthetic data set. We present preliminary results showing that it might be possible to generate privatized synthetic data sets, with the same statistical skeletal structure as the original, using distance transforms. Therefore, the main goal of this investigation is to employ data privacy, distance transforms, and k-means clustering approaches in the production of privatized synthetic data with similar statistical traits as the original data. The rest of the paper is organized as follows, in Section II, background and related work is given, while Section III talks about the methodology. In Section IV, a discussion of the experiment and results is done, and lastly, in Section V, the conclusion is given.

II. BACKGROUND AND RELATED WORK

Not much work exists on the application of distance transforms for privatized synthetic data generation. The technique of distance transforms has largely been used for applications in the image-processing domain. However, a look at works by researchers in the signal processing domain shows that techniques, such as, discrete cosine transforms, have been proposed for privacy preservation applications [12][13][14][15]. For instance, Mukherjee, Chen, and Gangopadhyay (2006) suggested using Fourier-related transforms, to enhance Euclidean distance-based algorithms for privacy preservation in data mining applications [1]. Of the privacy preservations problems that Mukherjee et al., (2006) observed, was that while data allocations of the original data could be maintained in the confidential data set, the space between each data point in the confidential data set is not kept, which often leads to diminished cluster outcomes [1]. Moreover, Mukherjee et al., (2006) noted that one of the benefits of using signal processing techniques, such as discrete cosine, is that the Euclidean distance among points in the confidential data set could be maintained, resulting in improved clustering results [1]. *Distance transforms:* Distance transforms, a skeletonization process, proposed by Rosenfeld and Pfaltz (1968), and mainly used in the image processing domain, is a morphological technique that alters a binary image made up of object O foreground and object O' background pixels into a resulting skeletal figure in which each object pixel has an analogous value to the smallest amount of space from the background object O'. Distance transforms can be expressed using the following formula [2][3][4]:

$$D(p) = \text{minimum}\{distance(p,q),\ q \in O\} \qquad (1)$$

The symbols O and O' represent the object foreground and background; $distance(p,q)$ represents the space between pixel p and q . The least distance $minimum\{distance(p,q)\}$ is often required; and $D(p)$ symbolized the distance point at pixel p [3]. Euclidean distance is used for the morphological process [5][6][7].

III. Methodology

In this investigation, rather than apply discrete cosine transforms, as in [1], we apply distance transforms on a noisy data set with very similar traits to the original data. We use the following implementation phases in the generation of the privatized synthetic data as illustrated in Figure 1:

- *Phase 1:* In the first step of the process, a noisy data set, instead of the original, is generated so as to add an extra layer of privacy and make it difficult for reconstruction attacks. In case of a successful reconstruction attack, what the adversary gets is the noisy data, assuming no prior insider knowledge. Data privacy, in this first step, is achieved using noise addition [9], with a distribution $\varepsilon \sim N(\mu = 1, \sigma = 0.2)$ – generating a noisy data set with similar statistical traits to the original [10].

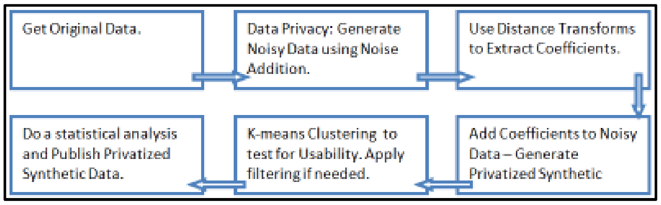
Get Original Data.
Data Privacy: Generate Noisy Data using Noise Addition.
Use Distance Transforms to Extract Coefficients.
Add Coefficients to Noisy Data – Generate Privatized Synthetic
K-means Clustering to test for Usability. Apply filtering if needed.
Do a statistical analysis and Publish Privatized Synthetic Data.

Figure 1: The Privatized Synthetic Data Generation Process

- *Phase 2:* In the second step, Distance transforms is applied on the noisy data set to extract coefficients.
- *Phase 3:* During the third step, the extracted coefficients are then added to back to the noisy data, generating the privatized synthetic data.
- *Phase 4:* In the fourth step, in order to reduce any excess noise, the moving average filtering is applied on the privatized synthetic data.
- *Phase 5:* In the fifth step of the process, we apply k-means clustering using Euclidean distance, to test the usability of the privatized synthetic data set, in comparison with the original data.
- *Phase 6:* In the final step, statistical analysis of both the original and privatized data sets is done, and the privatized synthetic data is published.

IV. Results and Discussion

The Fisher Iris data set used in this experiment, comprised of 150 data points, five attributes, namely, sepal length, sepal width, petal length, petal width, and class attribute, with three classes, namely, Setosa, Versicolar, and Virginica [8]. The plots in Figure 2 illustrate series for the Sepal length, Sepal width, Petal length, and Petal width correspondingly, before and after application of distance transforms. For each plot, the upper series symbolizes the privatized synthetic Fisher-Iris data, the middle series symbolizes the noisy Fisher-Iris data used to generate the privatized synthetic data, and the lower series in the graph symbolizes the coefficients extracted using the distance transforms method. As can be seen in Figure 2 from an anecdotal viewpoint, the privatized synthetic data series is an augmented outline of the noisy Fisher-Iris series used in the generation of the privatized synthetic data. The statistical analysis will further give more details that the statistical skeletal structure of the original data was maintained in the privatized synthetic data set. In Figure 3 the left sub-plot symbolizes the descriptive statistics of the original data, while the center sub-plot illustrates the statistical characteristics of noisy Fisher-Iris data, and the right sub-plot demonstrates the statistical traits of the generated privatized synthetic data. As shown in Figure 3 the statistical skeletal structural of the noisy data is maintained in the privatized synthetic data. For instance, the mean and median in the privatized synthetic data could be viewed as an augmentation of the same values in the noisy data, thus a statistical morphologic and skeletal structure of the original data. Since we derived the noisy data set by perturbing the original data, and likewise used distance transforms to extract coefficients from the noisy data and then generate the privatized synthetic data, the statistical skeletal structure of the original data is preserved, in this case, some level of augmentation has take place.

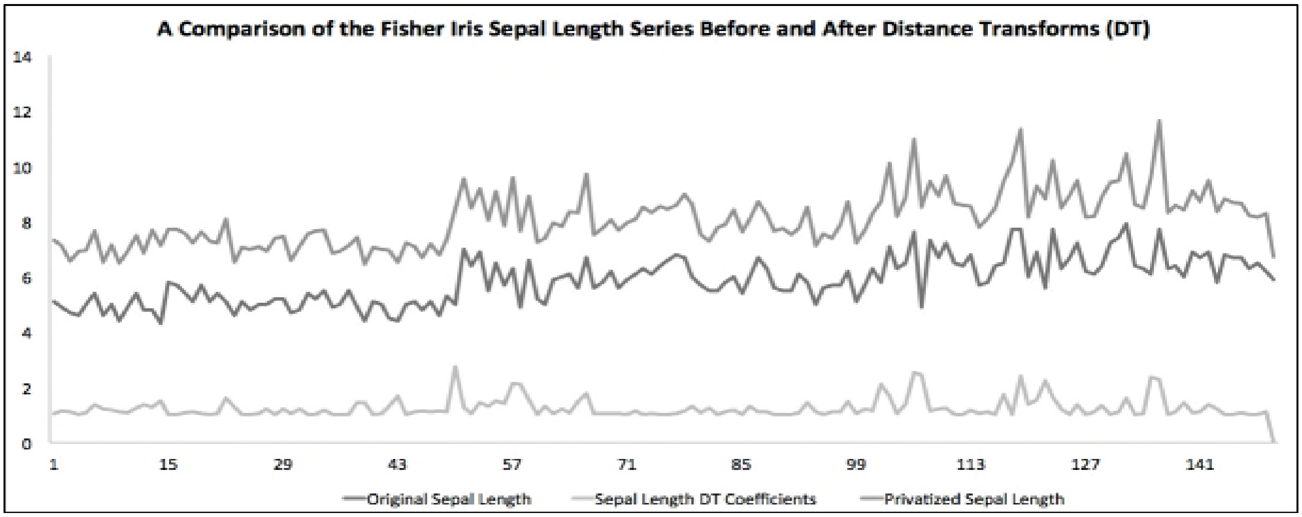

Figure 2: Original and Privatized Synthetic Fisher-Iris Sepal data series.

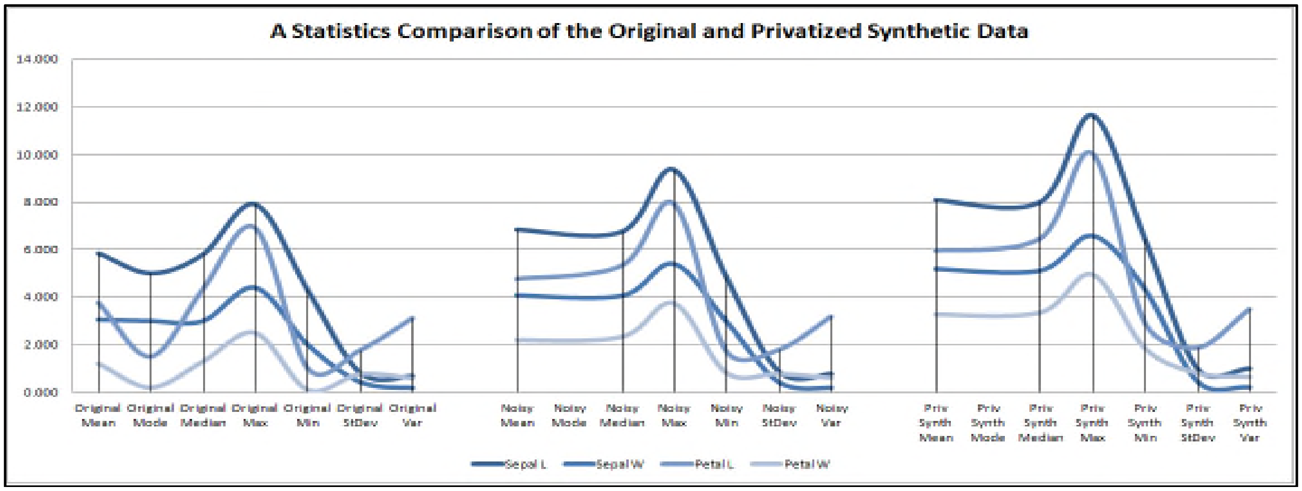

Figure 3: Original and Privatized Synthetic data – Descriptive statistics.

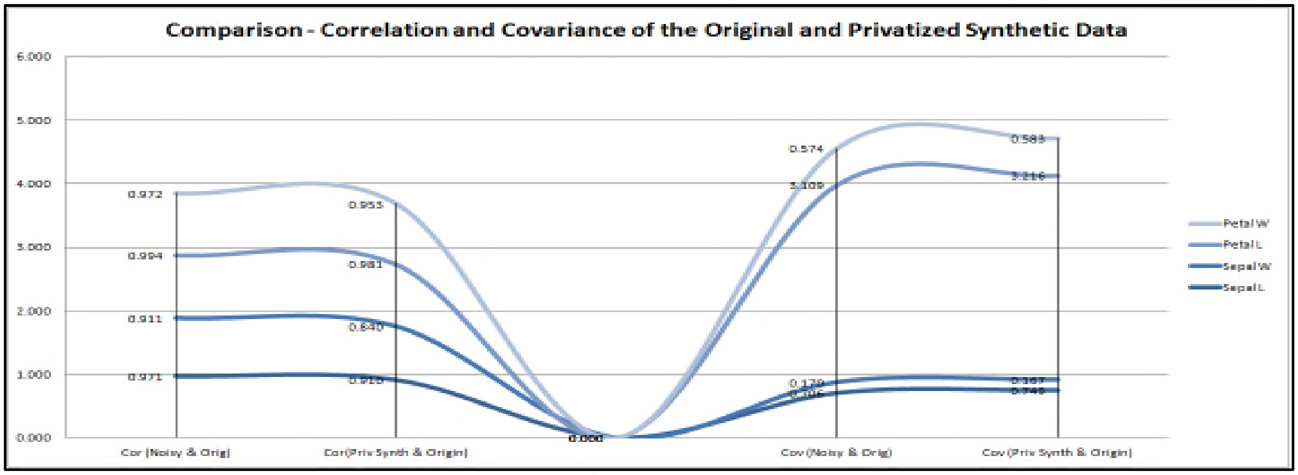

Figure 4. Correlation for Original and Privatized Synthetic data

For example, the mean values for the Sepal length is 5.834 in the original data, and 6.744 for the noisy data, with a dissimilarity of about 0.91. On the other hand, the mean value of 8.078 was registered for the privatized synthetic data, a dissimilarity of about 1.33 and 2.44, when compared with the noisy and original data respectively. The same statistical skeletal structure of the original, noisy, and privatized synthetic data sets is shown in Table 1 describing the descriptive statistics of the data sets. The same outcome is repeated for the other descriptive statistics, when comparing the original, noisy, and privatized synthetic data sets. It could then be argued that the privatized synthetic data could offer some level of data usability to researchers since it preserves some of the statistical characteristics of the original data – in this case, a statistical morphological and skeletal structure of the original data is preserved. The covariance results as shown in Figure 4 and Table 2 for the original, noisy, and privatized synthetic, vary between 0 and 3. For example, the Sepal length, Sepal width, and petal width, covariance varies between 0 and 1, showing a small proclivity for the data sets to grow together, despite the covariance being positive, in this case.

TABLE 1. ORIGIN AND PRIVATE SYNTHETIC DATA – DESCRIPTIVE STATISTICS

Statistics	Sepal L	Sepal W	Petal L	Petal W
Original Mean	5.843	3.054	3.759	1.199
Original Mode	5.000	3.000	1.500	0.200
Original Median	5.800	3.000	4.350	1.300
Original Max	7.900	4.400	6.900	2.500
Original Min	4.300	2.000	1.000	0.100
Original StDev	0.828	0.434	1.764	0.763
Original Variance	0.686	0.188	3.113	0.582
Noisy Mean	6.841	4.077	4.766	2.200
Noisy Median	6.744	4.060	5.323	2.333
Noisy Max	9.353	5.398	7.921	3.747
Noisy Min	4.846	2.978	1.716	0.819
Noisy StDev	0.883	0.433	1.784	0.779
Noisy Variance	0.780	0.188	3.183	0.607
Priv Synthetic Mean	8.078	5.185	5.959	3.279
Priv Synthetic Mode	#N/A	#N/A	#N/A	#N/A
Priv Synthetic Median	8.001	5.119	6.473	3.364
Priv Synthetic Max	11.631	6.576	10.028	4.962
Priv Synthetic Min	6.444	4.328	2.907	1.855
Priv Synthetic StDev	1.001	0.463	1.870	0.807
Priv Synthetic Var	1.001	0.214	3.497	0.651

However, for the Petal length, covariance registers a value of 3, indicating a possibility for the Petal length attributes in the data sets might grow together [9]. Additionally, as shown in Figure 4 and Table 2 the correlation values between the original, noisy, and privatized data sets vary from 0.840 to 0.994, approximately near +1, an indication of a better linear association and therefore a strong relationship [9]. For that reason, it could be argued that for the sake of data usability, the generated privatized synthetic data set might retain the statistical traits of the original data.

TABLE 2: CORRELATION FOR ORIGINAL AND PRIVATIZED SYNTHETIC DATA

Statistics	Sepal L	Sepal W	Petal L	Petal W
Cor (Noisy & Orig)	0.971	0.911	0.994	0.972
Cor (Priv Synth & Orig)	0.910	0.840	0.981	0.953
Cov (Noisy & Orig)	0.706	0.170	3.109	0.574
Cov (Priv Synth & Orig)	0.749	0.167	3.216	0.583

Clustering performance: Additionally, clustering analysis was done for the original and privatized synthetic data sets to further test for usability. Since the Euclidean distance was used in computing the morphological transforms of the privatized synthetic data set, Euclidean distance-based unsupervised learning methods, such as k-means clustering, could be used in testing for data usability of the privatized synthetic data sets.

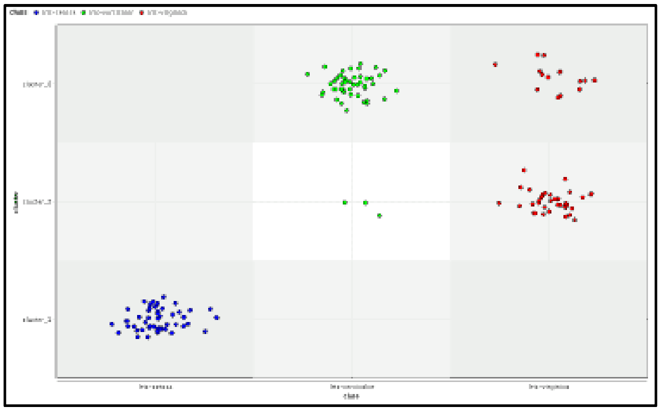

Figure 5. Original data clustering results

Furthermore, we put to test the suggestion by Mukherjee et al (2006), that one of the compensations of employing signal processing methods, such as discrete cosine, is that the Euclidean distance among points in the privatized data set could be preserved, with enhanced clustering outcomes [1][5][6]. In this case, we test to see if this proposal could hold when using image processing technique of distance transforms, as per our implementation. Clustering outcomes of the original Fisher-Iris data are shown in Figure 5 with the *x*-axis symbolizing the three classes – Iris Setosa, Iris Versicolar, and Iris Virginica; the *y*-axis symbolizes the number of clusters generated. K-means with Euclidean distance algorithm was employed for the clustering test, with *k* =3. An anecdotal view point of Figure 5 shows that Iris Virginica category did not cluster well for the original data. Yet still, from an anecdotal view, there seems to be an observable improvement in clustering results for the privatized synthetic data as shown in Figure 6, with the exception of the Iris Virginica attribute. However, after application of the Davis-Bouldin Index metric [10][11], to test the clustering performance, there was an actual degradation in the clustering performance, as shown in Figure 8 and Table 3. The Davis Bouldin Index for the original data was reported at 0.668 and 0.765 for the privatized synthetic data. The lower the Davis Bouldin Index, the greater the clustering performance. Therefore the suggestion that signal processing techniques, such as, discrete cosine transforms, could improve Euclidean distance-based clustering results, did not hold for the image processing technique of Distance Transforms, in this experiment [1].

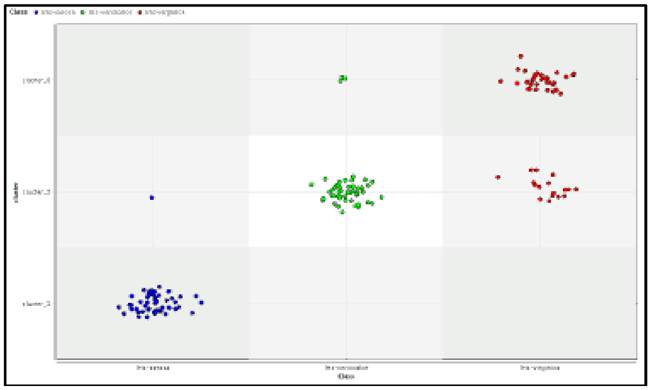

Figure 6. Privatized synthetic data clustering results

To mitigate this problem, we applied the Moving Average Filtering technique [11] on the privatized synthetic data set and then applied k-means clustering on the filtered data set again.

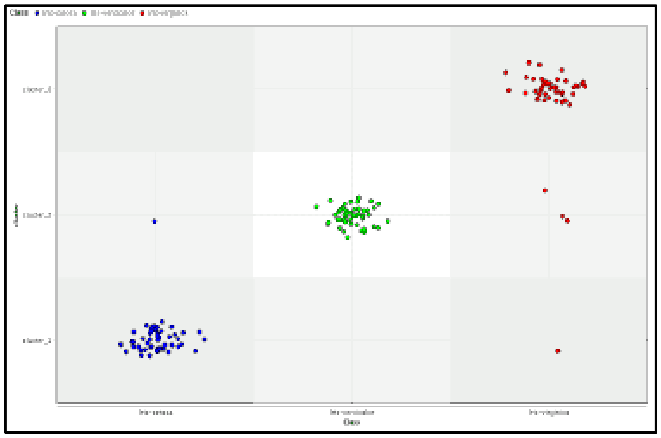

Figure 7. Filtered privatized synthetic data clustering results

Following the application of the moving average filter on the privatized synthetic data set, we clustered using k-means, with *k*=3, and as illustrated by the clustering outcome in Figure 7 there was an improvement with the Iris Virginica cluster. In fact, as illustrated in Figure 8, the Davis Bouldin index returned a value of 0.419 for the filtered privatized synthetic data, compared to the 0.668 for the original data, signifying an enhanced improvement in the clustering results for the privatized synthetic data, after filtering. Furthermore, using the distance between clusters metric – in this case, the average within centroid distance, to measure how well the clustering performed, Table 3 shows that the average within centroid distance of data points in the original data is approximately 0.5, while that for the non-filtered privatized synthetic data, is at 0.934. However, the average within centroid distance of data points in the filtered privatized synthetic data is about 0.477, an improvement that surpasses both the original and non-filtered privatized synthetic data sets.

We further tested for data usability by analyzing clustering performance, quantifying the number of items in each cluster, as a metric – the motivation was that the number of items in each cluster, in the privatized synthetic data should be similar to the number of items in each cluster, in the original data. For instance, as illustrated in Table 4 for the original data, there are 61, 50, 39, number of items in clusters 0, 1, and 2, in that order. However, for the non-filtered privatized synthetic data, there are 36, 49, 65, number of items in clusters 0, 1, and 2, respectively. Finally, for the filtered privatized synthetic data, we have the number of items as 46, 50, and 54, in clusters, 0, 1, and 2 respectively.

TABLE 3. CLUSTERING PERFORMANCE METRICS

Cluster Distance Performance	Original Data	Priv Synth Data	Filtered Priv Synth Data
Avg. within centroid distance	0.547	0.934	0.477
Avg. within centroid distance_cluster_0	0.562	0.657	0.635
Avg. within centroid distance_cluster_1	0.527	1.09	0.502
Avg. within centroid distance_cluster_2	0.492	0.961	0.268
Davies Bouldin Criterion	0.668	0.765	0.419

While the number of items in each of the clusters in the privatized synthetic data might not be close to that of the original data, we interpret this as a good indication of confidentiality, making it difficult for an adversary to know exactly how many items appeared in the clusters of the original data. Therefore, we could add to the argument that it might be possible to generate privatized synthetic data sets with acceptable levels of both confidentiality and usability.

TABLE 4. NUMBER OF ITEMS IN EACH CLUSTER

Cluster	Original Data	Synthetic Fisher Iris (DT) Data	Filtered Synthetic Fisher Iris (DT) Data
Cluster 0	61	36	46
Cluster 1	50	49	50
Cluster 2	39	65	54
Total	150	150	150

I. CONCLUSION

We have presented a confidential synthetic data generation heuristic, that employs a combination of data privacy and distance transforms techniques, for the generation of privatized synthetic data with similar statistical traits of the original data. We have also presented empirical results from applying unsupervised learning, using k-means, to test the usability of the privatized synthetic data sets. We applied average moving filtering on the privatized synthetic data and showed that filtering might help improve clustering results. Based on our empirical results from this study and implementation, we argue that it might be possible to generate privatized synthetic data sets, with acceptable levels of both privacy and data usability, while preserving the same statistical morphological and skeletal structure of the original, using a combination of data privacy, distance transforms, and filtering techniques. On the limitations of this study and future work, we focused on implementing the generation of privatized synthetic data sets using data privacy, distance transforms, and filtering techniques. While much effort could have been given to the testing of the privatized synthetic data sets to various adversarial attacks, our efforts were largely spent on the generation of the privatized synthetic data sets, leaving the study of attacks on privatized synthetic data sets for future work. Finally, generation of privatized synthetic data sets that retain the statistical structure of the original data, remains a challenge, and is in the early stages of research. More investigations on theoretical studies, practical implementations, and gathering of empirical results, is highly necessary for the advancement of privatized synthetic data set generation with enhanced levels of usability.

ACKNOWLEDGEMENT

Portion of this work was presented as part of the dissertation by the author in fulfillment of the requirements for the D.Sc. degree, in the Computer Science Department, at Bowie State University [10]. With great gratitude, I extend my sincere thanks Dr. Claude Turner and Dr. Soo-Yeon Ji, in the Computer Science Department, at Bowie State University, for the untiring assistance during this study. I would like to acknowledge with great appreciation, the HBGI Grant from the United States Department of Education for the facilitation of this work.

REFERENCES

[1] S. Mukherjee, Z. Chen, and A. Gangopadhyay, “A privacy-preserving technique for Euclidean distance-based mining algorithms using Fourier-related transforms,” VLDB J., vol. 15, no. 4, pp. 293–315, Aug. 2006.

[2] F. Y.-C. Shih and O. R. Mitchell, “A mathematical morphology approach to Euclidean distance transformation.,” IEEE Trans. Image Process., vol. 1, no. 2, pp. 197–204, Jan. 1992.

[3] O. Cuisenaire and B. Macq, “Fast Euclidean Distance Transformation by Propagation Using Multiple Neighborhoods,” Comput. Vis. Image Underst., vol. 76, no. 2, pp. 163–172, Nov. 1999.

[4] A. Rosenfeld and J. L. Pfaltz, “Distance functions on digital pictures,” Pattern Recognit., vol. 1, no. 1, pp. 33–61, 1968

[5] C. T. Huang and 0. Robert Mitchell, “A Euclidean distance transform using grayscale morphology decomposition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 16, no. 4, pp. 443–448, 1994.

[6] D. G. Bailey, “An efficient euclidean distance transform,” in LNCS Combinatorial Image Analysis, 2004, vol. 2, no. 3, pp. 394–408.

[7] D. G. Bailey, “Accelerating the distance transform,” in Proceedings of the 27th Conference on Image and Vision Computing New Zealand - IVCNZ ’12, 2012, pp. 162–167.

[8] K. Bache and M. Lichman, “Iris Fisher Dataset - UCI Machine Learning Repository.” University of California, School of Information and Computer Science., Irvine, CA, 2013.

[9] K. Mivule, “Utilizing Noise Addition for Data Privacy , an Overview,” in Proceedings of the International Conference on Information and Knowledge Engineering (IKE 2012), 2012, pp. 65–71.

[10] K. Mivule, "An Investigation of Data Privacy and Utility Using Machine Learning as a Gauge", D.Sc. Dissertation, Computer Science Dept., Bowie State University. 2014: 262 pages; ProQuest: 3619387.

[11] K. Mivule and C. Turner, "Applying Moving Average Filtering for Non-interactive Differential Privacy Settings", Procedia Computer Science, Volume 36, ISSN: 1877-0509, 2014, Pages 409-415. DOI: 10.1016/j.procs.2014.09.013

[12] S. Hajian and M. A. Azgomi, “A privacy preserving clustering technique using Haar wavelet transform and scaling data perturbation,” in 2008 International Conference on Innovations in Information Technology, 2008, pp. 218–222.

[13] O. Chertov and D. Tavrov, “Providing Group Anonymity Using Wavelet Transform,” in Lecture Notes in Computer Science - BNCOD 27, 2012, vol. 6121, pp. 25–36.

[14] N. V. Lalitha, G. Suresh, and P. Telagarapu, “Audio authentication using Arnold and Discrete Cosine Transform,” in 2012 International Conference on Computing, Electronics and Electrical Technologies (ICCEET), 2012, pp. 530–532.

[15] M. Niimi, F. Masutani, and H. Noda, “Protection of privacy in JPEG files using reversible information hiding,” in 2012 International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS), 2012, no. Ispacs, pp. 441–446.