Exploring the Tradeoff between Utility and Privacy for Analyzing Stock Trend.

Zamil S. Alzamil.

College of Computer and Information Sciences, Majmaah University, Al-Majmaah 11952, Saudi Arabia, z.alzamil@mu.edu

Abstract

In this study of exploring the tradeoff between utility and privacy, we have used the S&P 500 data as a proof of concept, for Association Rule Data Mining using privacy-preserved data. The purpose of choosing S&P 500 was that unlike synthetically generated data, S&P 500 index though public, was real-world data nonetheless and hence incorporated in it all the political and social events that occurred over the period of time under consideration. Secondly, we could use stock price as a quasi-identifier. Any seasoned stockbroker would be able to identify the organization simply by looking at the stock price fluctuations over a period even if the ticker information is removed. And stock price is the most important variable used in any data mining related with stock market. Hence, we had to perturb this one variable in order to test for anonymity and utility. Finally, the data being numerical in nature helped us utilize the statistical data perturbation techniques. Association rule mining techniques were first applied to the original data. The data was then perturbed and, once again, using binary similarity algorithms, trend matching was performed on the identified stocks from the original data. The three tests of different random data generations were evaluated for data utility and privacy. The results re-emphasized on the tradeoff between the data utility and privacy where greater perturbation meant greater privacy but higher utility loss and vice versa.

Keywords:

Association Rule Mining; tradeoff; Privacy-preserved data; Utility.

1. Introduction

As the amount of data and the ability to store it grows exponentially in recent times; legal, social and political issues surrounding the relationship between collection and dissemination of this information are becoming ever more complex. Data privacy deals with many of these issues in relation with the personally identifiable information. The primary focus of data privacy is to share data while keeping personally identifiable information discrete. The collection of this data can be from a wide range of sources including but not limited to healthcare, financial and location.

Most privacy protection techniques rely on the suppression or generalization of the "quasi-identifier" attribute such as location, ethnicity, age, etc. There are various ways to achieve this including techniques like k-anonymity, l-diversity and data perturbation.

We looked at the data privacy problem from a data mining perspective. With this astronomical growth in data there is now an increased effort from organizations to utilize this data for insights which can lead

to business profitability. Association rule data mining is frequently used for this purpose. The challenge, however, with data mining is that few analysts have an exact idea of what they are looking for beforehand. In other words, they are unaware of any significant relationships that they might discover in data. Therefore, privacy preserving techniques can have a significant impact on the results which we might get using association rule data mining. Consequently, in a paper that appeared in KDD 2008, Brickell and Shmatikov [1] proposed an evaluation methodology by comparing privacy gain with utility gain resulting from anonymizing data, and concluded that "even modest privacy gains require almost complete destruction of the data-mining utility". We however, are not as skeptic.

Thus, the primary objective of this paper is to quantify the impact of data perturbation techniques on association rule mining. We performed this by first performing association rule mining on S&P 500 data to identify relationships that met a user-defined support function. We then performed data perturbation on the same data and executed the same association rule mining functions to identify whether the original patterns still existed and to what degree of significance.

A secondary purpose was to utilize stock market data to search for trends in stock price movement. Data mining is being used extensively of late, with varying degrees of success, to extrapolate for future stock price movement. Many individuals and organizations are out to make a profit out of any pattern identification through data mining.

This paper is thus a practical implementation of data privacy and association rule data mining techniques on real-world data in order to discover interesting trends and to quantify information loss using associative rule data mining.

2. Related Work

The tradeoff between privacy and utility in data processing and usage is indeed a crucial and complex topic in today's digital age. It revolves around the challenges of balancing the benefits of data utilization for various purposes such as innovation, research, and personalization with the need to protect personal and sensitive data and privacy rights. The tradeoff has gained significant prominence due to the growing concerns about data breaches, surveillance and misuse of personal information. There exist many ways to protect personal and sensitive data and one of them is data perturbation which is one of the effective solutions to deal with such a problem. In [8], the authors explore the privacy-preserving aspects of random data perturbation techniques, especially by focusing on adding random noise or making random modifications to the data to protect privacy in data mining tasks.

In [1], the paper introduces the concept of anonymization which is a technique used to protect individual privacy when publishing datasets for research and analysis. It highlights the importance of balancing privacy protection with data utility. The

authors also discuss the inherent tension between privacy preservation and the usefulness of data for data mining and analysis. It delves into the challenges and costs associated with achieving privacy through data anonymization, particularly in the context of data mining and analysis. It underscores the need for a nuanced approach to balancing privacy concerns with the usefulness of data for research purposes.

Kargupta et al. [9] explore the concept of privacy-preserving data mining and focus on random-data perturbation techniques as a means to protect individual privacy while still enabling valuable data analysis. It highlights the challenges and tradeoffs involved in this field and calls for continued research to improve the effectiveness and efficiency of privacy-preserving methods. This work is part of the broader effort to reconcile the benefits of data mining with the need to safeguard privacy in an era of increasing data availability.

The paper in [10] presents a fuzzy-based data perturbation technique aimed at preserving privacy while allowing for meaningful data mining. This technique leverages fuzzy logic to introduce controlled perturbations into the data, adding a layer of uncertainty that protects individual records. Liu et al. [11] present a privacy-preserving data mining technique that combines random projection and multiplicative perturbation to enable distributed data mining while protecting individual privacy. It includes an experimental evaluation of the approach and discusses its applications in various domains. This work contributes to the ongoing efforts to develop effective privacy-preserving techniques for data mining in distributed and collaborative settings.

In [12], the authors propose a privacy-preserving technique based on geometric data perturbation. Geometric perturbation involves transforming the data points in a way that preserves the overall data distribution and statistical properties but obscures individual data values. The paper also compares its proposed approach to other privacy-preserving techniques, highlighting its advantages and limitations in the context of outsourced data mining.

The literature also discusses the security of databases, in [13] the authors discuss the importance of ensuring the security of sensitive information stored in databases, especially in the context of potential unauthorized access or breaches. They propose an additive data perturbation technique as a means of enhancing database security. The additive perturbation used involves introducing controlled noise or random values to the original data, making it more challenging for malicious actors to extract meaningful information.

Kiran and Shirisha [14] present a privacy preservation technique in the context of data mining, specifically focusing on k-anonymization. K-anonymization is a method for ensuring that each record in a dataset is indistinguishable from at least k-1 other records, making it more challenging to identify individual data points. The algorithm involves selecting suitable attributes for generalization, deciding on a suitable

value of k, and applying perturbation techniques accordingly. Their work contributes to the ongoing efforts to protect individual privacy while extracting valuable insights from data.

In [15], the authors propose a data perturbation technique based on min-max normalization. Min-max normalization is a commonly used data scaling technique that transforms data values into a specified range (usually [0, 1]). The normalization process may be used to obscure the original data values while preserving the relative relationships between data points. The paper concludes by summarizing the contributions and significance of their perturbation technique in the context of data mining and privacy preservation. It also suggests directions for future research, such as optimizing the perturbation process or exploring its applicability in specific domains.

In [16], authors proposed a new index, tuple-relation, which reflects the association strength between POI sets in an indoor environment. This index considers the potential association information such as spatial and semantic information between indoor POI sets.

In [17], authors have proposed a formal approach for feature extraction by first applying feature selection heuristics based on the statistical impurity measures, the Gini Index, Max Minority, and the Twoing Rule for obtaining the top 100-400 genes. Then the associative dependencies between the genes were analyzed and assigned weights to the genes based on their degree of participation in the rules. The authors introduced a weighted Jaccard and vector cosine similarity measure to compute the similarity between the discovered rules. Finally, the rules were grouped by with hierarchical clustering.

In the context of exploring data mining techniques used in finding causal relationships in the stock market, there exist many published papers in the literature such as in [2], [3], [4], [5], [6] and [7].

3. Data Preprocessing: Why Process the Data?

We used S&P 500 data for the year 2012 for association rule data mining. We utilized this data because it encompassed in itself all the economic, social and political distortions of the real world. The data was acquired from the Wharton Research Data Services (WRDS) database, owned by The Wharton Business School, University of Pennsylvania.

We started the data cleaning process by removing incomplete values and null data. There were a few companies which had joined the S&P index during 2012 and did not have data for the complete year. Hence companies like ABBV, ADT Corp, DLPH, DG to name a few were removed from the final dataset. In addition, we removed the day of October 29th, 2012 from the dataset since this was the day hurricane Sandy struck New York and stock market was closed for that day. Hence once our data cleaning tasks were complete, our list was reduced to four hundred and eighty eight companies from the original 500 once we removed the companies with incomplete data for the year of 2012. We now had 250

days of trading data for these companies.

Data preprocessing is the most important stage and one of the essential parts while doing data mining tasks, it is said that the more time you spend on your data preprocessing the better and more accurate results you will get from your data mining analysis.

In this project, we used several tools to facilitate the data preprocessing process. The tools employed were Weka 3, MS Excel, MS Access and IBM SPSS.

a.Weka

Weka (Waikato Environment for Knowledge Analysis) is a free software written in Java under the GNU General Public License, developed at the University of Waikato, New Zealand. It is a data explorer utility which can be used for data understanding and exploration in order to better understand data for preprocessing. Weka gains its popularity due to ease of use, and its visualization tools and algorithms for data analysis. It contains tools for preprocessing, classifying, clustering, and association data mining.

b. Excel

Excel is a powerful tool for data processing. We used MS Excel primarily for data reduction as well as normalization and transformation. Excel was used for transforming the numbers into binary attributes, also, transforming some large values and to reduce them to a smaller scale, helping to better compare different data sets. Furthermore, we added the binary stock price trend in addition to the binary transformation using excel.

c. Access

We utilized MS Access to extract data from Excel files for each company independently. Hence, by using MS Access and SQL (Structured Query Language), we retrieve our data for the format that we want and put it in separate files to get the maximum use of it in further analysis in SPSS. For instance, we executed queries that extract the binary price movement and the trend attribute from the original table and rearranged it again by date in a new spreadsheet.

4. Approach and Analysis

After preprocessing the S&P 500 dataset for the year of 2012, we used two approaches to achieve our goals; converting stock prices into binary values and changing them into trends. The objective was to perform association rule mining in order to identify similar trends in individual stock price movements over the year. Hence the stock price was converted into binary value and then to convert that binary value into a weighted moving average to reflect a trend. Finally, once the trends have been generated for the companies, we used Simple Matching Coefficient (SMC) to mine for similar trends among different stocks.

Hence the stock price was converted into the binary value using the equation $V(d) = p(d) - p(d-1)$. Here, $V(d)$ is the difference in stock value on day d, $p(d)$ is the price of stock on day d, while $p(d-1)$ is the price of same stock on the previous day. So, if V is greater than zero then the binary value is 1 else is 0. Once we had the binary stock price values, we then convert-

ed them into binary trend values using a five-day moving average with a "Recency Bias". The following equation was used to convert the binary stock values into trend: T (d) = ((V (d) * 1) + (V (d-1) * 0.8) + (V (d-2) *0.6) + (V (d-3) *0.4) + (V (d-4) *0.2)) / 3, where T is the trend value for the day d, and V is binary daily movement value for day d. Hence, if the value of T is greater than 0.5, then binary trend value for day d is 1 else it is 0.

Once stock price trends for selected companies were calculated, the data was then uploaded to IBM SPSS Statistics software for data analysis. SPSS contains various data mining techniques including correlation and association rule mining. We used the Simple Matching Coefficient algorithm for similarity analysis between various stock price trends. SMC was preferred over other similarity measures because it takes into consideration both 1s and 0s matches, meaning that in our case it took both up and down movement in stock prices. The following Table shows the similarity matrix generated by SPSS:

Table 1: Proximity matrix using SPSS.

Proximity Matrix															
Simple matching Measure															
	A	AA	AAPL	ABC	ABT	ACE	CAN	ADBE	ADI	ADM	ADP	ADSK	AEE	AEP	AES
A	1	0.62	0.518	0.584	0.535	0.563	0.669	0.653	0.669	0.547	0.62	0.682	0.58	0.539	0.604
AA	0.62	1	0.531	0.58	0.547	0.624	0.584	0.633	0.6	0.559	0.6	0.58	0.527	0.6	0.608
AAPL	0.518	0.531	1	0.478	0.576	0.49	0.522	0.571	0.555	0.531	0.58	0.576	0.49	0.514	0.588
ABC	0.584	0.58	0.478	1	0.535	0.62	0.571	0.588	0.547	0.62	0.596	0.584	0.62	0.62	0.596
ABT	0.535	0.547	0.576	0.535	1	0.637	0.629	0.596	0.604	0.571	0.653	0.559	0.604	0.604	0.645
ACE	0.563	0.624	0.49	0.62	0.637	1	0.576	0.6	0.576	0.6	0.641	0.62	0.633	0.641	0.616
CAN	0.669	0.584	0.522	0.571	0.629	0.576	1	0.682	0.665	0.657	0.616	0.629	0.559	0.608	0.608
ADBE	0.653	0.633	0.571	0.588	0.596	0.6	0.682	1	0.722	0.641	0.649	0.678	0.551	0.641	0.616
ADI	0.669	0.6	0.555	0.547	0.604	0.576	0.665	0.722	1	0.649	0.624	0.653	0.551	0.6	0.665
ADM	0.547	0.559	0.531	0.62	0.571	0.6	0.657	0.641	0.649	1	0.649	0.596	0.641	0.673	0.633
ADP	0.62	0.6	0.58	0.596	0.653	0.641	0.616	0.649	0.624	0.649	1	0.645	0.649	0.665	0.649
ADSK	0.682	0.58	0.576	0.584	0.559	0.62	0.629	0.678	0.653	0.596	0.645	1	0.473	0.588	0.62
AEE	0.58	0.527	0.49	0.62	0.604	0.633	0.559	0.551	0.551	0.641	0.649	0.473	1	0.682	0.624
AEP	0.539	0.6	0.514	0.62	0.604	0.641	0.608	0.641	0.6	0.673	0.665	0.588	0.682	1	0.673
AES	0.604	0.608	0.588	0.596	0.645	0.616	0.608	0.616	0.665	0.633	0.649	0.62	0.624	0.673	1

From the stock prices of 150 companies analyzed, the following 2 groups of companies (ATI, CLF) and (CMA, BBT) had a similarity coefficient of greater than 0.8. The next table illustrates that:

Table 2: Similarity table.

Proximity Matrix				
	Simple matching Measure			
	ATI	CLF	CMA	BBT
ATI	1.000	.824	.678	.637
CLF	.824	1.000	.624	.665
CMA	.678	.624	1.000	.812
BBT	.637	.665	.812	1.000

Figure 1 and Figure 2 below show the actual stock price movement for the 2 groups of companies for the year of 2012. We can see from the diagrams that the stock prices of the companies grouped together followed a similar trend over the year.

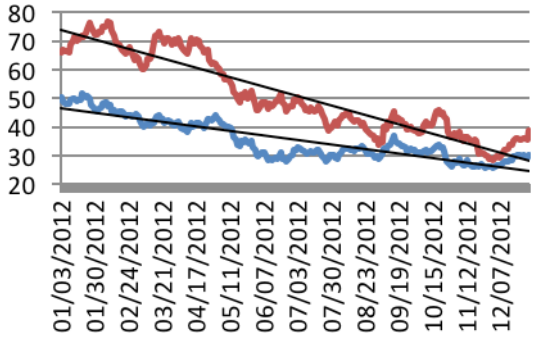

Figure 1: ATI and CLF stock price trends.

5. Data Obfuscation

Once we had identified the two trends from associative analysis that were above our confidence threshold, we proceeded to apply the data privacy techniques on this data and perform the similar association rule mining using SMC.

Many application domains nowadays require having data shared from multiple sources, which can range from passing data to a single party or having data shared among many entities for the need of getting use of it to do some analysis, in our case, data mining.

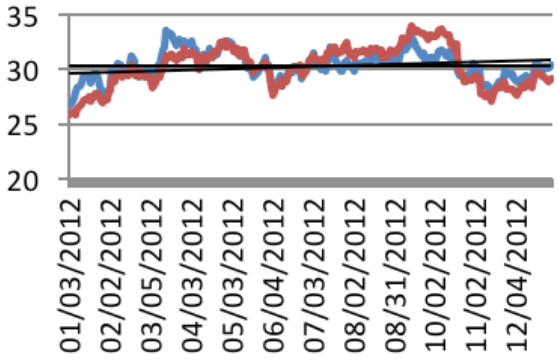

Figure 2: CMA and BBT stock price trends.

Data perturbation is one of the effective solutions to deal with such problem. Our data obfuscation technique perturbs the price of each stock by using a randomization approach. So, the stock price data of the four companies which exhibited high similarity using SMC were perturbed using randomized normal distribution.

To illustrate this, we assumed that original stock values are X1, X2, …, Xn. In order to hide these values, we generate a random normal probability distribution Y1, Y2, …, Yn. The new perturbed stock values will be generated by adding these two values X1+Y1, X2+Y2, …, Xn+Yn. Three different normal random distributions were used to analyze the impact on data utility. We kept the mean of the random normal probability distribution at 43 (this is an arbitrary number; here we took it simply because the average value of the four stock prices over the year was 43). We took the standard deviation values of 10, 3, and 1 to the random probability distribution to generate 3 varying distributions in order to compare the loss in data utility for different levels of data perturbation.

6. Results

The following Figure exhibits the new proximity matrices generated by IBM SPSS using perturbed data generated from the three different random normal probability distributions. The three proximity matrices exhibit the similarity coefficients from each random normal probability distribution. We also calculated a data loss matrix for each random distribution where the original matrix was used as the base reference. Data loss was calculated using the following equation: |(OV–PV)/OV|. Here, OV is the original similarity coefficient while PV is the perturbed coefficient value.

Proximity Matrix

	Simple matching Measure			
	ATI	CLF	CMA	BBT
ATI	1.000	.539	.469	.506
CLF	.539	1.000	.449	.486
CMA	.469	.449	1.000	.498
BBT	.506	.486	.498	1.000

Mean 43
St Dev 10

Data Loss Matrix

	ATI	CLF	CMA	BBT
ATI	0%	35%	31%	21%
CLF	35%	0%	28%	27%
CMA	31%	28%	0%	39%
BBT	21%	27%	39%	0%

Avg Data Loss 22%

	Simple matching Measure			
	ATI	CLF	CMA	BBT
ATI	1.000	.551	.522	.535
CLF	.551	1.000	.514	.567
CMA	.522	.514	1.000	.457
BBT	.535	.567	.457	1.000

Mean 43
St Dev 3

	ATI	CLF	CMA	BBT
ATI	0%	33%	23%	16%
CLF	33%	0%	18%	15%
CMA	23%	18%	0%	44%
BBT	16%	15%	44%	0%

Avg Data Loss 19%

	Simple matching Measure			
	ATI	CLF	CMA	BBT
ATI	1.000	.600	.563	.518
CLF	.600	1.000	.563	.518
CMA	.563	.563	1.000	.588
BBT	.518	.518	.588	1.000

Mean 43
St Dev 1

	ATI	CLF	CMA	BBT
ATI	0%	27%	17%	19%
CLF	27%	0%	10%	22%
CMA	17%	10%	0%	28%
BBT	19%	22%	28%	0%

Avg Data Loss 15%

Figure 3: Proximity matrices after applying the perturbation method used.

The following three charts help visualize the changes due to the different standard deviations used:

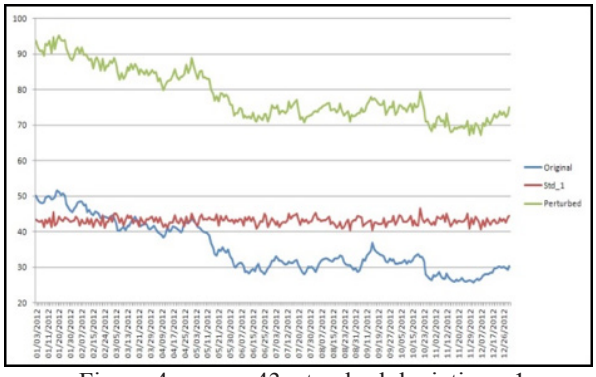

Figure 4: mean=43, standard deviation =1

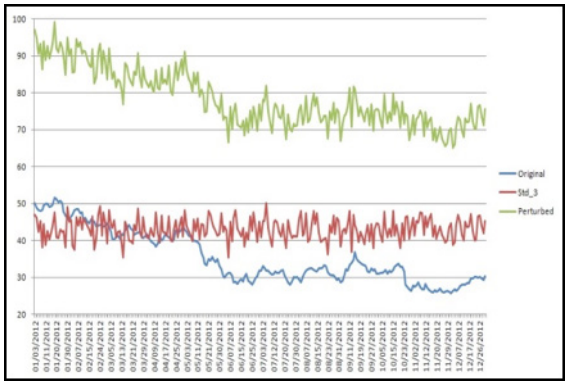

Figure 5: mean=43, standard deviation=3

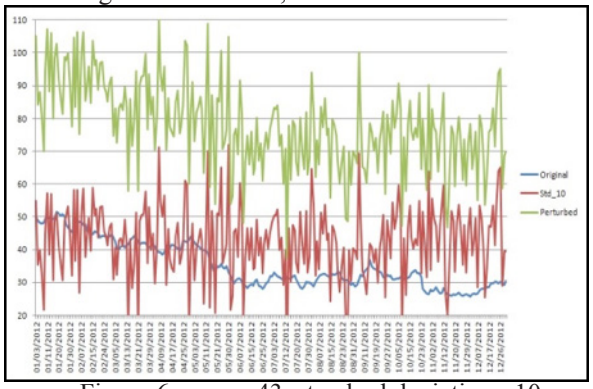

Figure 6: mean=43, standard deviation =10

7. Conclusion

Our results concluded that the similarity patterns were still identifiable in the perturbed data with varying degrees. However, the patterns were not as strong as the original data and utility depended on the amount of perturbation. This was expected as increasing data anonymity would lead

to data loss, but the data was not completely rendered useless. Hence we believe that degree of privacy and utility loss will depend on the external variables like regulations and business requirements. Not all businesses, data types or regulations will require the same amount of privacy and hence varying degrees of data perturbations can be used to fulfill the maximum constraints.

Acknowledgements

I would like to express my heartfelt gratitude to Dr. Ussama Yaqub, The Lahore University of Management Sciences, whose unwavering support and invaluable assistance were instrumental in the successful completion and publication of this research paper. His guidance, expertise, and dedication were indispensable throughout every stage of this endeavor.

References

[1] Brickell, Justin, and Vitaly Shmatikov. "The cost of privacy: destruction of data-mining utility in anonymized data publishing." In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 70-78. ACM, 2008.

[2] Hsieh, Y. L., Don-Lin Yang, and Jungpin Wu. "Using data mining to study upstream and downstream causal relationship in stock market." Computer 1 (2005): F02.

[3] Cicil Fonseka and Liwan Liyanage (2008), A Data mining algorithm to analyze stock market data using lagged correlation, IEEE, 978-1-4244-2900-4/08/2008, 4.

[4] Ou, Phichhang, and Hengshan Wang. "Prediction of stock market index movement by ten data mining techniques." Modern Applied Science 3, no. 12 (2009): P28.

[5] Soon, Lay-Ki, and Sang Ho Lee. "Explorative Data Mining on Stock Data–Experimental Results and Findings." In Advanced Data Mining and Applications, pp. 562-569. Springer Berlin Heidelberg, 2007.

[6] Mahajan, Mrs Keerti S., and R. V. Kulkarni. "A REVIEW: APPLICATION OF DATAMINING TOOLS FOR STOCK MARKET."

[7] Khedkar, Mr Amit, and R. V. Argiddi. "To Study and Analyze to foresee market Using Data Mining Technique.", International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 9- Sep 2013

[8] Kargupta, Hillol, Souptik Datta, Qi Wang, and Krishnamoorthy Sivakumar. "On the privacy preserving properties of random data perturbation techniques." In Third IEEE international conference on data mining, pp. 99-106. IEEE, 2003.

[9] Kargupta, Hillol, Souptik Datta, Qi Wang, and Krishnamoorthy Sivakumar. "Random-data perturbation techniques and privacy-preserving data mining." Knowledge and Information Systems 7 (2005): 387-414.

[10] Shynu, P. G., H. Md Shayan, and Chiranji Lal Chowdhary. "A fuzzy based data perturbation technique for privacy preserved data mining." In 2020 International Conference on Emerging Trends in

Information Technology and Engineering (ic-ETITE), pp. 1-4. IEEE, 2020.

[11] Liu, Kun, Hillol Kargupta, and Jessica Ryan. "Random projection-based multiplicative data perturbation for privacy preserving distributed data mining." IEEE Transactions on knowledge and Data Engineering 18, no. 1 (2005): 92-106.

[12] Chen, Keke, and Ling Liu. "Geometric data perturbation for privacy preserving outsourced data mining." Knowledge and information systems 29 (2011): 657-695.

[13] Muralidhar, Krishnamurty, Rahul Parsa, and Rathindra Sarathy. "A general additive data perturbation method for database security." management science 45, no. 10 (1999): 1399-1415.

[14] Kiran, Ajmeera, and N. Shirisha. "K-Anonymization approach for privacy preservation using data perturbation techniques in data mining." Materials Today: Proceedings 64 (2022): 578-584.

[15] Kiran, Ajmeera, and D. Vasumathi. "Data mining: min–max normalization based data perturbation technique for privacy preservation." In Proceedings of the Third International Conference on Computational Intelligence and Informatics: ICCII 2018, pp. 723-734. Singapore: Springer Singapore, 2020.

[16] N. Mou, H. Wang, H. Zhang and X. Fu, "Association Rule Mining Method Based on the Similarity Metric of Tuple-Relation in Indoor Environment," in IEEE Access, vol. 8, pp. 52041-52051, 2020, doi: 10.1109/ACCESS.2020.2980952.

[17] Sethi P, Alagiriswamy S. Association rule based similarity measures for the clustering of gene expression data. Open Med Inform J. 2010;4:63-73. doi: 10.2174/1874431101004010063, Epub 2010 May 28. PMID: 21603179; PMCID: PMC3096052.