Wavelet Transforms through Differential Privacy

N SriDevi[1], V Sunitha[2]
M.Tech Student, Avanthi Institute of Engineering and Technology, Narsipatnam
Assistant Professor, CSE Dept, Avanthi Institute of Engineering and Technology, Narsipatnam

Abstract:

Privacy preservation has become a major issue in many data analysis applications. When a data set is released to other parties for data analysis, privacy-preserving techniques are often required to reduce the possibility of identifying sensitive information about individuals. However, many solutions exist for privacy preserving data; differential privacy has emerged as a new paradigm for privacy protection with very conservative assumptions and guarantees the strongest privacy. In particular, for a count query answered by output data set, the noise in the result makes it vain as the result set could be equivalent to the number of Tuples in the data. This paper proposes a data publishing technique that not only ensures differential privacy, but also provides accurate results for all range-count queries, i.e., count queries where the predicate on each attribute is a range. The main aim of the solution provides a frame work called privelet that applies wavelet transforms s on the data before adding noise to it. This paper also outlines the instantiations of the Privelet for both ordinal and nominal data and theoretical Analysis is provided to prove the privacy guarantee of Privelet with the nominal wavelet transform.

Keywords – Privacy, Wavelet, Tuples, Dworks, Privelet.

Introduction

A Large number of data are maintained like census bureaus and hospitals, by number of organizations. Such data collections are of significant research value, and there is much benefit in making them publicly available. Nevertheless, as the data are sensitive in nature, proper measures must be taken to ensure that its publication does not endanger the privacy of the individuals that contributed the data. A canonical solution to this problem is to modify the data before releasing them to the public, such that the modification prevents inference of private information while retaining statistical characteristics of the data.

A plethora of techniques have been proposed for privacy-preserving data publishing. Existing solutions make different assumptions about the background knowledge of an adversary who would like to attack the data i.e., to learn the private information about some individuals. Assumptions about the background knowledge of the adversary determine what types of attacks are possible. A solution that makes very conservative assumptions about the adversary's background knowledge is differential privacy. Informally, differential privacy requires that the data to be published should be generated using a randomized algorithm G, such that the output of G is not very sensitive to any particular tuple in the input.

The simplest method to enforce differential privacy, as proposed by Dwork, is to first compute the frequency distribution of the tuples in the input data and then publish a noisy version of the distribution. We introduce privacy preserving wavelet (Privelet), a data publishing technique that not only ensures differential privacy, but also provides accurate results for all range-count queries, i.e., count queries where the predicate on each attribute is a range. Specifically, Privelet guarantees that any range-count query can be answered with a noise variance that is polylogarithmic in m. This significantly improves over the O (m) noise variance bound provided by Dwork et al.'s method.

The effectiveness of Privelet results from a novel application of wavelet transforms a type of linear transformations that has been widely adopted for image processing and approximate query processing.

Background Work

A survey of recent developments, we got to find the below methods

- Privacy-Preserving Data Publishing: A Survey of Recent Developments
- Approximate Query Processing Using Wavelets
- Wavelet Synopses with Error Guarantees
- Privacy-preserving logistic regression
- Adaptive Wavelet Thresholding for Image Denoising and Compression

a) Privacy-Preserving Data Publishing

The collection of digital information by governments, corporations, and individuals has created tremendous opportunities for knowledge- and information-based decision making. Driven by mutual benefits, or by regulations that require certain data to be published, there is a demand for the exchange and publication of data among various parties. Data in its original form, however, typically contains sensitive information about individuals, and publishing such data will violate individual privacy. The current practice in data publishing relies mainly on policies and guidelines as to what types of data can be published and on agreements on the use of published data. This approach alone may lead to excessive data distortion or insufficient protection. Privacy-preserving data publishing (PPDP) provides methods and tools for publishing useful information while preserving data privacy. Recently, PPDP has received considerable attention in research communities, and many approaches have been proposed for different data publishing scenarios. In this survey, we will systematically summarize and evaluate different approaches to PPDP, study the challenges in practical data publishing, clarify the differences and requirements that distinguish PPDP from other related problems, and propose future research directions.

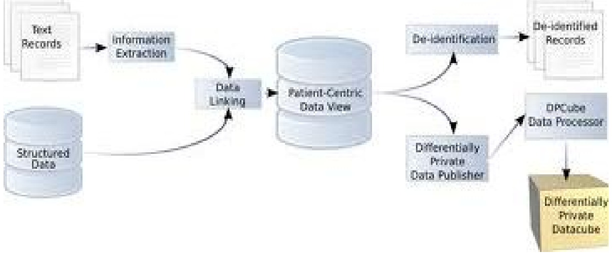


Fig1: Privacy-preserving data publishing

b) Approximate Query Processing Using Wavelets

Approximate query processing has emerged as a cost-effective approach for dealing with the huge data volumes and stringent response-time requirements of today’s decision-support systems. Most work in this area, however, has so far been limited in its applicability and query processing scope. In this paper, we propose the use of multi-dimensional wavelets as an effective tool for general-purpose approximate query processing in modern, high-dimensional applications. Our approach is based on building wavelet-coefficient synopses of the data and using these synopses to provide approximate answers to queries. We develop novel query processing algorithms that operate directly on the wavelet-coefficient synopses of relational tables, allowing us to process arbitrarily complex queries entirely in the wavelet-coefficient domain. This, of course, guarantees extremely fast response times since our approximate query execution engine can do the bulk of its processing over compact sets of wavelet coefficients, essentially postponing the expansion into relational tuples until the end-result of the query. An extensive experimental study with synthetic as well as real-life data sets establishes the effectiveness of our wavelet-based approach compared to sampling and histograms.Approximate query processing has recently emerged as a viable solution for dealing with the huge amounts of data, the high query complexities, and the increasingly stringent response-time requirements that characterize today’s Decision-Support Systems (DSS) applications.

c) Wavelet Synopses with Error Guarantees

Recent work has demonstrated the effectiveness of the wavelet decomposition in reducing large amounts of data to compact sets of wavelet coefficients (termed “wavelet synopses”) that can be used to provide fast and reasonably accurate approximate answers to queries. A major criticism of such techniques is that unlike, for example, random sampling, conventional wavelet synopses do not provide informative error guarantees on the accuracy of individual approximate answers. In fact, as this paper demonstrates, errors can vary widely (without bound) and unpredictably, even for identical queries on nearly-identical values in distinct parts of the data. This lack of error guarantees severely limits the practicality of traditional wavelets as an approximate query-processing tool, because users have no idea of the quality of any particular approximate answer. In this paper, we introduce Probabilistic Wavelet Synopses, the first wavelet-based data reduction technique with guarantees on the accuracy of individual approximate answers. Whereas earlier approaches rely on deterministic thresholding for selecting a set of “good” wavelet coefficients, our technique is based on a novel, probabilistic thresholding scheme that assigns each coefficient a probability of being retained based on its importance to the reconstruction of individual data values, and then flips coins to select the synopsis. We show how our scheme avoids the above pitfalls of deterministic thresholding, providing highly-accurate answers for individual data values in a data vector. We propose several novel optimization algorithms for tuning our probabilistic thresholding scheme to minimize desired error metrics. Experimental results on real-world and synthetic data sets evaluate these algorithms, and demonstrate the effectiveness of our probabilistic wavelet synopses in providing fast, highly-accurate answers with error guarantees.

d) Privacy-preserving logistic regression

This paper addresses the important tradeoff between privacy and learnability, when designing algorithms for learning from private databases. We focus on privacy-preserving logistic regression. First we apply an idea of Dwork to design a privacy-preserving logistic regression algorithm. This involves bounding the sensitivity of regularized logistic regression, and perturbing the learned classifier with noise proportional to the sensitivity. We then provide a privacy-preserving regularized logistic regression algorithm based on a new privacy-preserving technique: solving a perturbed optimization problem. We prove that our algorithm preserves privacy in the model We provide learning guarantees for both algorithms, which are tighter for our new algorithm, in cases in which one would typically apply logistic regression. Experiments demonstrate improved learning performance of our method, versus the sensitivity method. Our privacy-preserving technique does not depend on the sensitivity of the function, and extends easily to a class of convex loss functions. Our work also reveals an interesting connection between regularization and privacy. Privacy-preserving machine learning is an emerging problem, due in part to the increased reliance on the internet for day-to-day tasks such as banking, shopping, and social networking. Moreover, private data such as medical and financial records are increasingly being digitized, stored, and managed by independent companies.

d) Adaptive Wavelet Thresholding for Image Denoising and Compression

The first part of this paper proposes an adaptive, data-driven threshold for image denoising via wavelet soft-thresholding. The threshold is derived in a Bayesian framework, and the prior used on the wavelet coefficients is the generalized Gaussian distribution (GGD) widely used in image processing applications. The proposed threshold is simple and closed-form, and it is adaptive to each subband because it depends on data-driven estimates of the parameters. Experimental results show that the proposed method, called BayesShrink, is typically within 5% of the MSE of the best soft-thresholding benchmark with the image assumed known. It also outperforms Donoho and Johnstone's SureShrink most of the time. The second part of the paper attempts to further validate recent claims that lossy compression can be used for denoising. The BayesShrink threshold can aid in the parameter selection of a coder designed with the intention of denoising, and thus achieving simultaneous denoising and compression. Specifically, the zero-zone in the quantization step of compression is analogous to the threshold value in the thresholding function. The remaining coder design parameters are chosen based on a criterion derived from Rissanen's minimum description length (MDL) principle. Experiments show that this compression method does indeed remove noise significantly, especially for large noise power. However, it introduces quantization noise and should be used only if bitrate were an additional concern to denoising. IMAGE is often corrupted by noise in its acquisition or transmission. The goal of denoising is to remove the noise while retaining as much as possible the important signal features. Traditionally, this is achieved by linear processing such as Wiener filtering.

System Overwiew

a) System Structure

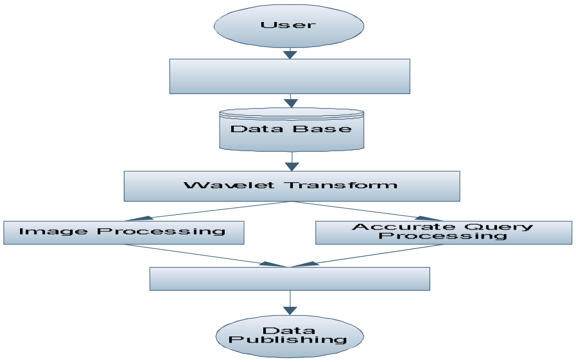


Fig.2: System Architecture of the Framework

b) System Functionality

The framework is designed for five basis entities they are User/Admin Login, User Register, Wavelet Transforms, Approximate Query Processing and Data Publishing.

User/Admin Login:

While the released table gives useful information to researchers, it presents disclosure risk to the individuals whose data are in the table. Therefore, our objective is to limit the disclosure risk to an acceptable level while maximizing the benefit. This is achieved by anonym zing the data before release.

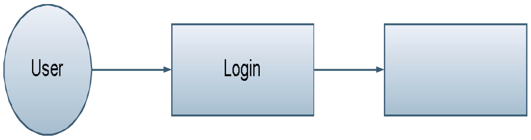


Fig.3: User-Admin Login interface

User Register:

User has Register their original information. Typically, such data are stored in a table, and each record (row) corresponds to one individual. Each record has a number of attributes.

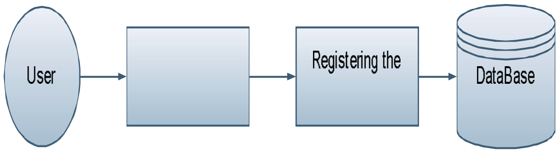


Fig.4: User Registration Interface

Wavelet Transforms:

The core of our solution is a framework that applies wavelet transforms on the data before adding noise to it. The effectiveness of Privelet results from a novel application of wavelet transforms a type of linear transformations that has been widely adopted for image processing and approximate query processing.

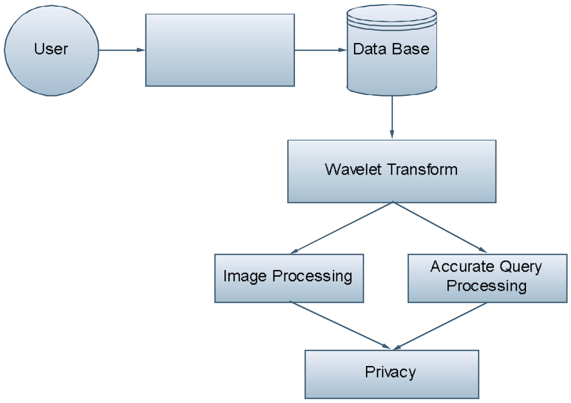


Fig.6: Wavelet Transforms interface

Approximate Query Processing:

The simplest method to enforce differential privacy, as proposed by Dwork ,For example, given the medical records in Table 1, Dwork method first maps the records to the frequency matrix in Table 2, where each entry in the first (second) column stores the number of diabetes (nondiabetes) patients in Table 1 that belong to a specific age group. After that, Dwork method adds independent noise1 with variance to each entry in Table 2and then publishes the noisy frequency matrix.

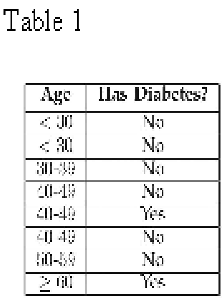


Table 1

Age	Has Diabetes?
< 30	No
< 30	No
30-39	No
40-49	No
40-49	Yes
40-49	No
50-59	No
≥ 60	Yes

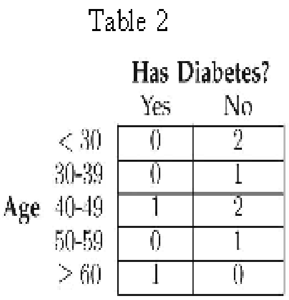


Table 2

		Has Diabetes?	
		Yes	No
Age	< 30	0	2
	30-39	0	1
	40-49	1	2
	50-59	0	1
	> 60	1	0

Data Publishing:

Privacy-preserving data publishing has attracted considerable research interest in recent years. We develop a data publishing technique that ensures differential privacy while providing accurate answers for range-count queries, i.e., count queries where the predicate on each attribute is a range. A plethora of techniques have been proposed for privacy-preserving data publishing. Existing solutions make different assumptions about the background knowledge of an adversary who would like to attack the data i.e., to learn the private information about some individuals.

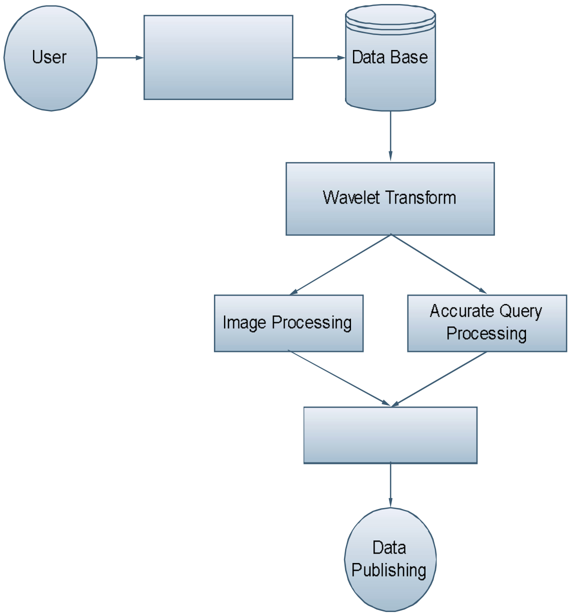


Fig.7: Data Publishing Interface

Design & Implementation

This Component design diagram helps to model the physical aspects of an object oriented software system i.e., for the proposed framework it illustrates the architecture of the dependencies between service provider and consumer.

The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted

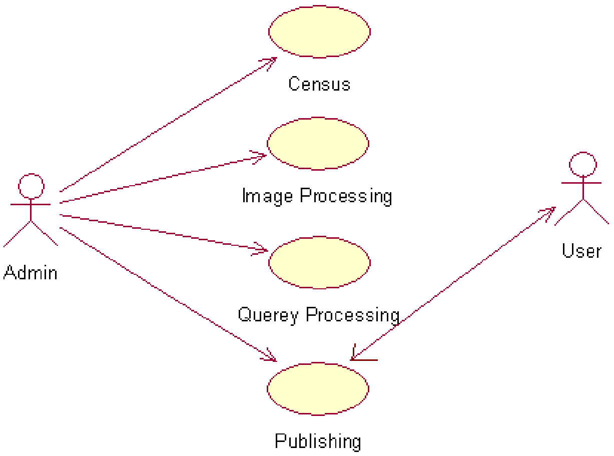


Fig.8:Inter-Operational Use case Diagram for Framework

A sequence diagram shows, as parallel vertical lines (lifelines), different processes or objects that live simultaneously, and, as horizontal arrows, the messages exchanged between them, in the order in which they occur. This allows the specification of simple runtime scenarios in a graphical manner

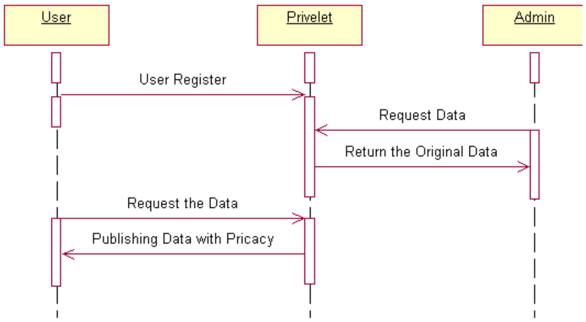
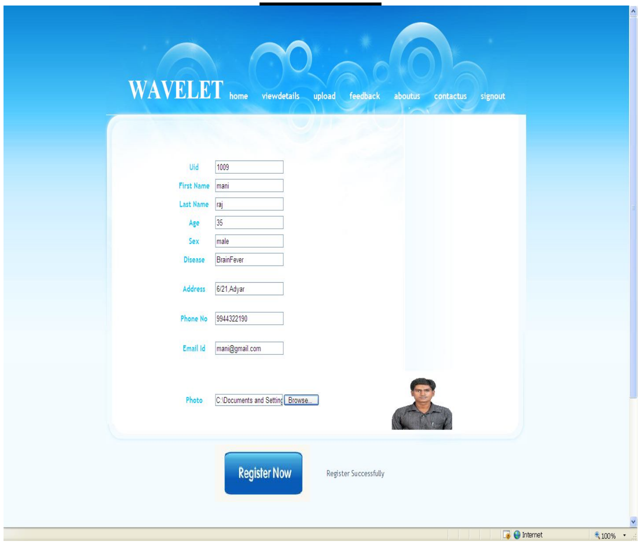


Fig.9:Inter-Operational Sequence Diagram for Framework

RESULTS



Uid 1009

First Name mani

Last Name raj

Age 35

Sex male

Disease BrainFever

Address 6/21,Adyar

Phone No 9944322190

Email Id mani@gmail.com

Photo C:\Documents and Setting

Register Successfully

Fig.10: User Registration

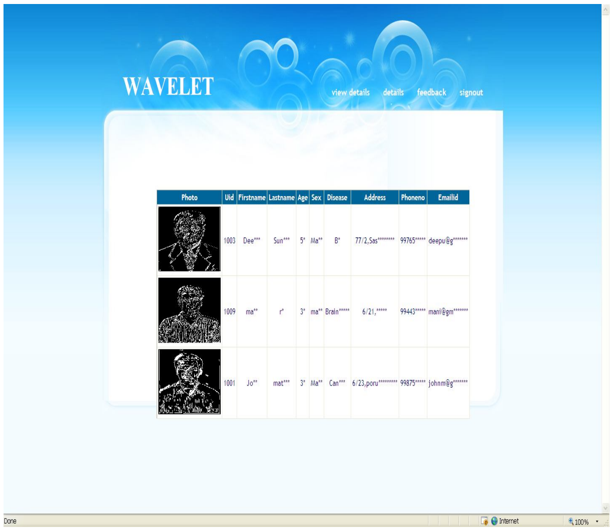
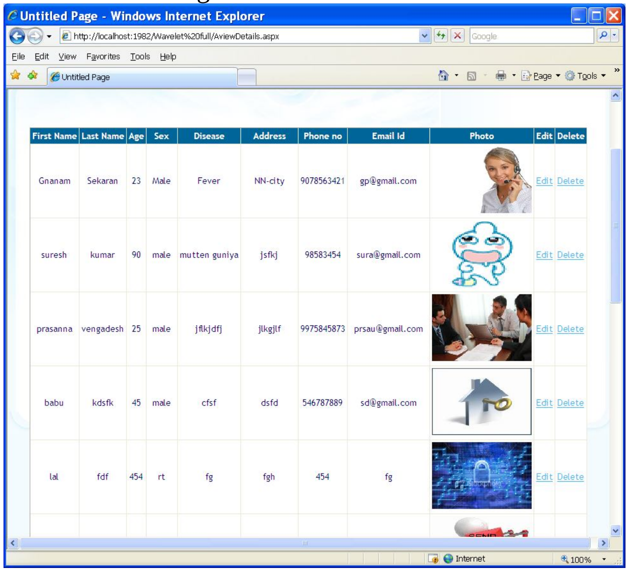


Photo	Uid	Firstname	Lastname	Age	Sex	Disease	Address	Phoneno	Emailid
	1003	Dee***	Sun***	5*	Ma**	B*	77/2,Sas********	99765*****	deepu@g*******
	1009	ma**	r*	3*	ma**	Brain*****	6/21,*****	99443*****	mani@gm*******
	1001	Jo**	mat***	3*	Ma**	Can***	6/23,poru*********	99875*****	johnm@g*******

Fig.1: Secure Data Details


Untitled Page - Windows Internet Explorer
http://localhost:1982/Wavelet%20full/AviewDetails.aspx
Google
File Edit View Favorites Tools Help
Untitled Page
Page
Tools
Internet
100%

First Name	Last Name	Age	Sex	Disease	Address	Phone no	Email Id	Photo	Edit	Delete
Gnanam	Sekaran	23	Male	Fever	NN-city	9078563421	gp@gmail.com		Edit	Delete
suresh	kumar	90	male	mutten guniya	jsfkj	98583454	sura@gmail.com		Edit	Delete
prasanna	vengadesh	25	male	jflkjdfj	jlkgjlf	9975845873	prsau@gmail.com		Edit	Delete
babu	kdsfk	45	male	cfsf	dsfd	546787889	sd@gmail.com		Edit	Delete
lal	fdf	454	rt	fg	fgh	454	fg		Edit	Delete

Fig.13: Original Data

CONCLUSION

We present the privelet framework for incorporating wavelet transforms in data publishing, and we establish a sufficient condition for achieving differential privacy better than the existing solutions .This paper also demonstrates the effectiveness and efficiency of Privelet, a technique that heuristically reduces the amount of noise in the data. Our experimental result shows that Privelet outperforms in terms of the accuracy of range count queries and it incurs a smaller query error, whenever the query coverage is larger. Currently Privelet only provides bounds on the noise variance in the query results, for future work we want to investigate what guarantees Privelet may offer for other.

Reference

[1] A. Blum, K. Ligett, and A. Roth, “A Learning Theory Approach to Non-Interactive Database Privacy,” Proc. 40th Ann. ACM Symp. Theory of Computing (STOC), pp. 609-618, 2008.

[2] S.P. Kasiviswanathan, H.K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, “What Can We Learn Privately?” Proc. 49th Ann. IEEE Symp. Foundations of Computer Science (FOCS), pp. 531-540, 2008.

[3] K. Nissim, S. Raskhodnikova, and A. Smith, “Smooth Sensitivity and Sampling in Private Data Analysis,” Proc. 39th Ann. ACM Symp. Theory of Computing (STOC), pp. 75-84, 2007.

[4] D. Donoho and I. Johnstone, “Ideal Spatial Adaptation via Wavelet Shrinkage,” Biometrika, vol. 81, pp. 425-455, 1994.

[5] M. Elad, “Why Simple Shrinkage is Still Relevant for Redundant Representations?” IEEE Trans. Information Theory, vol. 52, no. 12, pp. 5559-5569, Dec. 2006.

[6] . Li,M.Hay,V.Rastogi,G.Miklau,andA.McGregor, “Optimizing Linear Counting Queries under Differential Privacy,” Proc. 29th ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems (PODS), pp. 123-134, 2010.

[7] A. Ghosh, T. Roughgarden, and M. Sundararajan, “Universally Utility-Maximizing Privacy Mechanisms,” Proc. Ann. ACM Symp. Theory of Computing (STOC), pp. 351-360, 2009.

[8] S.R. Ganta, S.P. Kasiviswanathan, and A. Smith, “Composition Attacks and Auxiliary Information in Data Privacy,” Proc. 14th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD), pp. 265-273, 2008.

[9] D. Kifer, “Attacks on Privacy and de Finetti’s Theorem,” Proc. ACM SIGMOD Int’l Conf. Management of Data, pp. 127-138, 2009.

[10] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating Noise to Sensitivity in Private Data Analysis,” Proc. Third Theory of Cryptography Conf. (TCC), pp. 265-284, 2006.

[11] A. Machanavajjhala, D. Kifer, J.M. Abowd, J. Gehrke, and L. Vilhuber, “Privacy: Theory Meets Practice on the Map,” Proc. 24th IEEE Int’l Conf. Data Eng. (ICDE), pp. 277-286, 2008.

AUTHOR LIST

Nakka SriDevi received B.Tech in Computer science and Information Technology from Vignan Institute of Information Technology affiliated

To JNTU, Hyderabad, and Pursuing M.Tech in Information Technology from Avanthi institute of Engineering & Technology Affiliated to JNTUK. My area of interests is Data Mining.

V Sunitha. Received M.Tech from JNTU and presently working as Assistant Professor in the department of CSE at Avanthi institute of Engineering and Technology, Narsipatnam. Her area of Interest is Data Mining.