ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.4. No.5, Pages : 19-24 (2015)
Special Issue of ICETEM 2015 - Held on October 27, 2015 in The Solitaire Hotel, Bangalore, India
http://www.warse.org/IJATCSE/static/pdf/Issue/icetem2015sp06.pdf

Privacy Preserving Data Mining: Survey of Approaches

SANJANA
RNSIT Bangalore,INDIA
sanjana.kiran1994@gmail.com

ABSTRACT
Privacy is one of the most important properties of an information system must satisfy, in which systems the need to share information among different, not trusted entities, the protection of sensible information has a relevant role. Thus privacy is becoming an increasingly important issue in many data mining applications. For that privacy secure distributed computation, which was done as part of a larger body of research in the suppression, cryptography, randomization, sumarization has achieved remarkable results. These results were shown using generic constructions that can be applied to any function that has an efficient representation as a circuit. A relatively new trend shows that classical access control techniques are not sufficient to guarantee privacy when data mining techniques are used in a malicious way. Privacy preserving data mining algorithms have been recently introduced with the aim of preventing the discovery of sensible information. In this paper we will describe the implementation of suppression, cryptography, randomization, sumarization in that data mining for privacy preserving.

KEYWORDS
Datamining,suppression,randomization,cryptography,summarization.

INTRODUCTION

Privacy preserving data mining (PPDM) refers to the area of data mining that seeks to safeguard sensitive information from unsanctioned disclosure. Mosttraditional data mining techniques analyze and model the dataset statistically, inaggregation, while privacy preservation is primarily concerned with protecting againstdisclosure of individual data records. This domain separation points to the technicalfeasibility of PPDM.

Historically, issues related to PPDM were first studied by the national statistical agenciesinterested in collecting private social and economical data, such as census and tax records,

and making it available for analysis by public servants, companies, and researchers.

Building accurate socio-economical models is vital for business planning and publicpolicy. Yet, there is no way of knowing in advance what models may be needed, nor is itfeasible for the statistical agency to perform all data processing for everyone, playing therole of a "trusted third party." Instead, the agency provides the data in a sanitized form

that allows statistical processing and protects the privacy of individual records, solving aproblem known as *privacy preserving data publishing*.

The term "privacy preserving data mining" was introduced in papers (Agrawal &Srikant,2000) and (Lindell &Pinkas, 2000). These papers considered two fundamental problemsof PPDM, privacy preserving data collection and mining a dataset partitioned acrossseveral private enterprises. Agrawal and Srikant (2000) devised a randomizationalgorithm that allows a large number of users to contribute their private records for

efficient centralized data mining while limiting the disclosure of their values; Lindell andPinkas (2000) invented a cryptographic protocol for decision tree construction over adataset horizontally partitioned between two parties. These methods were subsequentlyrefined and extended by many researchers worldwide.

SURVEY OF APPROACHES

The naïve approach to PPDM is "security by obscurity", where algorithms have noproven privacy guarantees. By its nature, privacy preservation is claimed *for all* datasetsand attacks of a certain class, a claim that cannot be proven by examples or informalconsiderations. We will avoid further discussion of this approach inthis forum. Recently, however, a number of principled approaches have been developedto enable PPDM, some listed below according to their method of defining and enforcingprivacy.

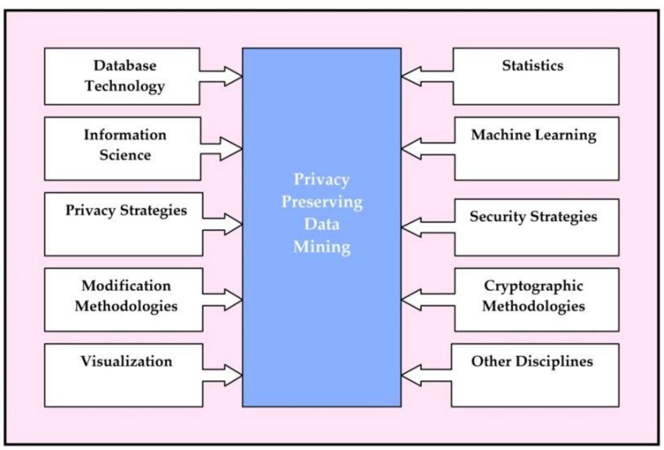

Figure 1:Methodologies

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.4. No.5, Pages : 19-24 (2015)
Special Issue of ICETEM 2015 - Held on October 27, 2015 in The Solitaire Hotel, Bangalore, India
http://www.warse.org/IJATCSE/static/pdf/Issue/icetem2015sp06.pdf

1.SUPPRESSION

Privacy can be preserved by simply suppressing all sensitive data before any disclosureor computation occurs. Given a database, we can suppress specific attributes in particularrecords as dictated by our privacy policy. For a partial suppression, an exact attributevalue can be replaced with a less informative value by rounding, top-coding, generalization (e.g. address to zip code), byusing intervals etc. Often the privacy guarantee trivially follows from the suppression policy. However, the analysis may bedifficult if the choice of alternative suppressions depends on the data being suppressed, orif there is dependency between disclosed and suppressed data. Suppression cannot beused if data mining requires full access to the sensitive values.

Rather than protecting the sensitive values of individual records, we may be interested in suppressing the identity (of a person) linked to a specific record. The process of alteringthe dataset to limit identity linkage is called *de-identification*. One popular definition forde-identification privacy is *k*-anonymity, formulated in (Samarati& Sweeney, 1998). Aset of personal records is said to be *k-anonymous* if every record is indistinguishable fromat least *k* _ 1 other records over given "quasi-identifier" subsets of attributes. A subset of attributes is a *quasi-identifier* if its value combination may help link some record to other personal information available to an attacker, e.g. the combination of age, sex andaddress.

To achieve *k*-anonymity, quasi-identifier attributes are completely or partially suppressed.

A particular suppression policy is chosen to maximize the utility of the *k*-anonymized dataset. The attributes that are not amongquasi-identifiers, even if sensitive (e.g. diagnosis), are not suppressed and may get linkedto an identity (Machanavajjhala et al. 2006). Utility maximization may create anexploitable dependence between the suppressed data and the suppression policy. Finally,*k*-anonymity is difficult to enforce before all data is collected in one trusted place;however, a cryptographic solution is proposed in (Zhong et al. 2005) based on Shamir's secret sharing scheme.

Suppression can also be used to protect from the discovery of certain statistical characteristics, such as sensitive association rules, while minimizing the distortion of other data mining results. Many related optimization problems are computationally intractable, but some heuristic algorithms were studied (Atallah et al. 1999) (Oliveira & Zaïane, 2003).

2.RANDOMIZATION

Suppose there is one central server, e.g. of a company, and many customers, each having a small piece of information. The server collects the information and performs datamining to build an aggregate data model. The randomization approachprotects the customers data by letting them randomly perturb their records beforesending them to the server, taking away some true information and introducing somenoise. At the server's side, statistical estimation over noisy data is employed to recoverthe aggregates needed for data mining. Noise can be introduced e.g. by adding ormultiplying random values to numerical attributesor bydeleting real items and adding "bogus" items to set-valued records. Given the right choice of the method and the amount of randomization, it is sometimes possible to protect individual values while estimating theaggregate model with relatively high accuracy.

Privacy protection by data perturbation has been extensively studied in the statisticaldatabases community. Incontrast to the above scenario, this research focuses mainly on the protection of publishedviews once all original data is collected in a single trusted repository. Many moreperturbation techniques are available in this case, including attribute swapping acrossrecords and data re-sampling by imputation.

A popular privacy definition to characterize randomization has its roots in the classicalsecrecy framework and in the work on disclosure risk and harmmeasures for statistical databases, but received its current formulationonly recently. Todeal with the uncertainty arising from randomization, the data miner's knowledge (belief)is modeled as a probability distribution. A simplified version of the definition is given inthe next paragraphs.

Suppose Alice is a customer and Bob is a company employee interested in miningcustomers' data. Alice has a private record *x* and a randomization algorithm *R*. To allowBob to do the mining while protecting her own privacy, Alice sends Bob a randomizedrecord *x*_ □□*R*(*x*). Let us denote by *pR*(*x*_ | *x*) the probability that algorithm *R* outputs *x*_ oninput *x*. We say that algorithm *R* achieves □-*leakage* (also called □-*privacy* or *at most* _-*amplification*) at output *x*_ if for every pair of private records *x*1 and *x*2 we have:

pR(*x*_ | *x*1) / *pR*(*x*_ | *x*2) □□_, where _ = exp(□)

We assume that Bob has some *a priori* belief about Alice's record, defined as the probability distribution *p*(*x*) over all possible private records. Once Bob receives arandomized record, his belief changes to some *a posteriori* distribution. If randomization*R* achieves □-leakage at output *x*_, then randomized record *x*_ gives Bob only a boundedamount of knowledge of Alice's unknown private record *x*. In fact, for every question *Q*about Alice's record, Bob's *a posteriori* belief *p*(*Q* | *x*_) that the answer to *Q* is "yes" is

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.4. No.5, Pages : 19-24 (2015)
Special Issue of ICETEM 2015 - Held on October 27, 2015 in The Solitaire Hotel, Bangalore, India
http://www.warse.org/IJATCSE/static/pdf/Issue/icetem2015sp06.pdf

bounded with respect to his *a priori* belief $p(Q)$ as follows:
$p(Q \mid x\_)\ p(Q)$
$1 - p(Q \mid x\_)$
□□_
$1 - p(Q)$
If R achieves □-leakage at every output, Bob's knowledge gain about Alice's record isalways bounded; if R achieves □-leakage at some outputs but not others, Bob'sknowledge gain is bounded only with a certain probability.
The above definition assumes that Bob cannot gain any knowledge of Alice's record bycollecting data from other customers, i.e. that all customers are independent. Theparameter □□is chosen to attain the right balance between privacy and the accuracy of theaggregate estimators used by the data miner. One advantage ofrandomization is that privacy guarantees can be proven by just studying the randomization algorithm, not the data mining operations. One disadvantage is that theresults are always approximate; high enough accuracy often requires a lot of randomizeddata.

3.CRYPTOGRAPHY

The cryptographic approach to PPDM assumes that the data is stored at several privateparties, who agree to disclose the result of a certain data mining computation performedjointly over their data. The parties engage in a cryptographic protocol, i.e. they exchangemessages encrypted to make some operations efficient while others computationally intractable. In effect, they "blindly" run their data mining algorithm. Classical works insecure multiparty computation such as Yao (1986) and Goldreich et al. (1987) show thatany function $F(x1, x2, \ldots, xn)$ computable in polynomial time is also securely computablein polynomial time by n parties, each holding one argument, under quite broadassumptions regarding how much the parties trust each other. However, this genericmethodology can only be scaled to database-sized arguments with significant additionalresearch effort.
The first adaptation of cryptographic techniques to data mining is done by Lindell &Pinkas (2000), for the problem of decision tree construction over horizontally partitioneddata; it was followed by many papers covering different data mining techniques andassumptions. The assumptions include restrictions on the input data and permitted disclosure, the computational hardness of certain mathematical operations such asfactoring a large integer, and the adversarial potential of the parties involved: the partiesmay be *passive* (*honest-but-curious*, running the protocol correctly but taking advantageof all incoming messages) or *malicious* (running a different protocol), some parties may be allowed to *collude* (represent a single adversary) etc. In addition to the genericmethodology such as oblivious transfer and secure Boolean circuit evaluation, the keycryptographic constructs often used in PPDM include homomorphic and commutativeencryption functions, secure multiparty scalar product and polynomial computation. Theuse of randomness is essential for all protocols.The privacy guarantee used in this approach is based on the notion of computational indistinguishability between random variables. Let $Xk$and Ykbe two random variablesthat output Boolean vectors of length polynomial in k; they are called *computationallyindistinguishable* if for all polynomial algorithms Ak(alternatively, for any sequence ofcircuits of size polynomial in k), for all $c > 0$ and for all sufficiently large integers k :
$| \text{Prob}\ [Ak(Xk) = 1] - \text{Prob}\ [Ak(Yk) = 1] \mid < 1 / kc.$
The above essentially says that no polynomial algorithm can tell apart $Xk$from Yk. Toprove that a cryptographic protocol is secure, we show that each party's view of theprotocol (all its incoming messages and random choices) is computationally indistinguishable from a simulation of this view by this party alone. When simulating theview of the protocol, the party is given everything it is allowed to learn, including thefinal data mining output. The exact formulation of the privacy guarantee depends on theadversarial assumptions. Goldreich (2004) and Stinson (2006) provide a thoroughintroduction into the cryptographic framework.
Scalability is the main stumbling block for the cryptographic PPDM; the approach isespecially difficult to scale when more than a few parties are involved. Also, it does notaddress the question of whether the disclosure of the final data mining result may breachthe privacy of individual records.

4.SUMARIZATION

This approach to PPDM consists of releasing the data in the form of a "summary" that allows the (approximate) evaluation of certain classes of aggregate queries while hidingthe individual records. In a sense, summarization extends randomization, but a summaryis often expected to be much shorter, ideally of sub-linear size with respect to the originaldataset. The idea goes back to statistical databases, where two summarization techniqueswere studied and widely applied: sampling and tabular data representation. *Sampling* corresponds to replacing theprivate dataset with a small sample of its records, often combined with suppression orperturbation of their values to prevent re-identification. *Tabular representation* summarizes data in a collection of aggregate quantities such as sums, averages or counts, aggregated over the range of some attributes while other attributes are fixed, similarly toOLAP (On Line Analytical Processing) cubes. Verifying privacy guarantees for tabulardata is challenging because of the potential for disclosure by inference.
Some of more recent summarization methods are based on pseudorandom sketches, aconcept

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.4. No.5, Pages : 19-24 (2015)
Special Issue of ICETEM 2015 - Held on October 27, 2015 in The Solitaire Hotel, Bangalore, India
http://www.warse.org/IJATCSE/static/pdf/Issue/icetem2015sp06.pdf

borrowed from limited-memory data stream processing.

Here is an illustration ofone such method. Suppose Alice has a small private set *S* of her favorite book titles, andwants to send to Bob a randomized version of this set. Alice splits *S* into two disjointsubsets, *S* = *S*0 □□*S*1, then constructs her randomized record *SR* by including *S*1, excluding*S*0, and for every book not in *S* including it into *SR* at random with probability 1/2. Ifthere are 1,000,000 possible book titles, *SR* will contain around 500,000 items, most ofthem purely random. Luckily, however, *SR* can be shortened. Let *G*(□, *i*) be apseudorandom generator that takes a short random seed □□and a book number *i*and computes a bit *bi*. Now Alice has a better strategy: once she selects *S*0 and *S*1 as before,she sends to Bob a randomly chosen seed □□such that *G*(□, #book) = 0 for all books in *S*0and *G*(□, #book) = 1 for all books in *S*1. Bob can use *G* and □□to reconstruct the entirerandomized record; and if *G* is sufficiently "well-mixing," every book not in *S* stillsatisfies *G*(□, #book) = 1 with probability 1/2. Thus, the short seed □□serves as thesummary of a randomized record. For complete analysis, see (Evfimievski et al. 2003)and (Mishra & Sandler, 2006).

The summarization approach is still in its infancy, more results are likely to come in thefuture. There has also been some work on combining sketches and approximationtechniques with the cryptographic approach,observe that the disclosure of an approximate function*f*appr(*x*) □□*f* (*x*) over private data *x* may be unacceptable even if the exact result *f* (*x*) ispermitted to disclose; indeed, just by learning whether *f*appr(*x*) □□*f* (*x*) or *f*appr(*x*) >*f* (*x*) theadversary may already get an extra bit of information about *x*. This issue is important tokeep in mind when designing sketch-based PPDM protocols.

APPLICATION SCENARIOS

Surveys and data collection. Companies collect personal preferences of their customersfor targeted product recommendations, or conduct surveys for business planning; politicalparties conduct opinion polls to adjust their strategy. The coverage of such datacollection may significantly increase if all respondents are aware that their privacy isprovably protected, also eliminating the bias associated with evasive answers. Therandomization approach has been considered as a solution in this domain.

Monitoring for emergencies. Early detection of large-scale abnormalities with potentialimplications for public safety or national security is important in protecting our wellbeing.Disease outbreaks, environmental disasters, terrorist acts, manufacturing accidentscan often be detected and contained before they endanger a large population. The firstindication of an impending disaster can be difficult to notice by looking at any individual case, but easy to see using data mining: an unusual increase in certain health symptoms ornon-prescription drug purchases, a surge in car accidents, a change in on-line trafficpattern, etc. To be effective, an early-detection system would have to collect personal,commercial, and sensor data from a variety of sources, making privacy issues paramount.

Product traceability. Before a product (e.g. a car or a drug) reaches its end-user, itusually passes through a long chain of processing steps, such as manufacturing,packaging, transportation, storage, and sale. In the near future, many products and package units will carry a radio-frequency identification (RFID) tag and will beautomatically registered at every processing step (Finkenzeller, 2003), (Garfinkel&Rosenberg, 2005). This will create a vast distributed collection of RFID traces, whichcan be mined to detect business patterns, market trends, inefficiencies and bottlenecks, criminal activity such as theft and counterfeiting, etc. However, such extremely detailedbusiness process data is a highly valuable and sensitive asset to the companies involved.Privacy safeguards will be very important to enable cooperative RFID data mining.

Medical research. Personal health records are one of the most sensitive types of privatedata; their privacy standards have been codified into law in many countries, e.g. HIPAA(Health Insurance Portability and Accountability Act) in the U.S. (OCR Privacy Brief,2003). On the other hand, data mining over health records is vital for medical, pharmaceutical, and environmental research. For example, a researcher may want tostudy the effect of a certain gene *A* on an adverse reaction to drug *B* (Agrawal et al. 2003).

But due to privacy concerns, the DNA sequences and the medical histories are stored atdifferent data repositories and cannot be brought together. Then, PPDM over verticallypartitioned data can be used to compute the aggregate counts while preserving theprivacy of records.

Social networks. In business as well as in life, the right connections make a huge difference. Whether it is expertise location, job search, or romance matchmaking, finding new connections is notoriously difficult, not least because the publicly available data is often very scarce and of low quality. Most of the relevant information is personal, copyrighted, or confidential, and therefore kept away from the Web. It is possible that

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.4. No.5, Pages : 19-24 (2015)
Special Issue of ICETEM 2015 - Held on October 27, 2015 in The Solitaire Hotel, Bangalore, India
http://www.warse.org/IJATCSE/static/pdf/Issue/icetem2015sp06.pdf

PPDM techniques can be utilized to allow limited disclosure options, prompting more people to engage in productive social networking, and guarding against abuse.

FUTURE TRENDS

The main technical challenge for PPDM is to make its algorithms scale and achievehigher accuracy while keeping the privacy guarantees. The known proof techniques andprivacy definitions are not yet flexible enough to take full advantage of existing PPDMapproaches. Adding a minor assumption (from the practical viewpoint) may slash the computation cost or allow much better accuracy, if the PPDM methodology is augmentedto leverage this assumption. On the other hand, proving complexity lower bounds andaccuracy upper bounds will expose the theoretical limits of PPDM.

One particularly interesting "minor assumption" is the existence of a computationallylimited trusted third party. Computer manufacturers such as IBM produce specialdevices called *secure coprocessors* (Dyer et al. 2001) that contain an entire computer within a sealed tamper-proof box. Secure coprocessors are able to withstand mosthardware and software attacks, or destroy all data if opened. For practical purposes,these devices can be considered trusted parties, albeit very restricted in the speed ofcomputation, in the volume of storage, and in communication with the untrusted components. It is known that secure coprocessors can be leveraged to enable privacypreserving operations over datasets much larger than their storage capacity (Smith &Safford, 2000) (Agrawal et al. 2006). Thus, applying them to PPDM looks natural.

If a data mining party cannot get accurate results because of privacy constraints enforcedby the data contributors, it may be willing to *pay* for more data. Kleinberg, et al. (2001)suggests measuring the amount of private information in terms of its monetary value, as aform of intellectual property. The cost of each piece of data must be determined in a"fair" way, so as to reflect the contribution of this piece in the overall profit. The paperborrows the notions of fairness from the theory of coalitional games: the core and theShapley value. Bridging game theory and PPDM could lay the theoretical foundation fora market of private data, where all participants receive appropriate compensations fortheir business contribution.

Among potential future applications for PPDM, we would like to emphasize data miningin healthcare and medical research. During the last few years the attention of the U. S.government has focused on transitioning the national healthcare system to an infrastructure based upon information technology (PITAC Report, 2004); a similar trendoccurs or is expected in countries around the world. Within a short time, millions ofmedical records will be available for mining, and their privacy protection will be requiredby law, potentially creating an urgent demand for PPDM. In addition to the traditionaldata mining tasks, new healthcare-specific tasks will likely become important, such as record linkage or mining over ontology-based and semistructured data, e.g. annotatedimages.

CONCLUSION

Privacy-preserving data mining emerged in response to two equally important (andseemingly disparate) needs: data analysis in order to deliver better services and ensuringthe privacy rights of the data owners. Difficult as the task of addressing these needs mayseem, several tangible efforts have been accomplished. In this paper, an overview of thepopular approaches for doing PPDM was presented, namely: suppression, randomization, cryptography and summarization. The privacy guarantees, advantages and disadvantagesof each approach were stated in order to provide a balanced view of the state of the art.Finally, the scenarios where PPDM may be used and some directions for future workwere outlined.

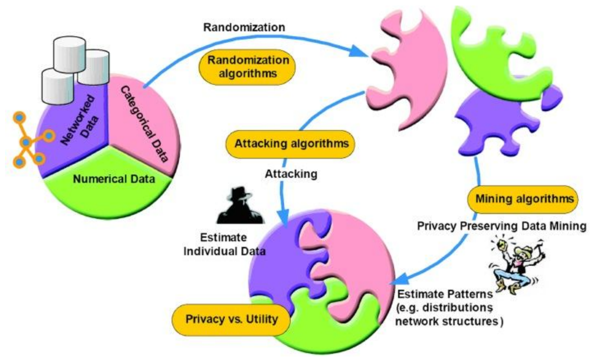

Figure 2;privacy and secured flow in process

REFERENCES

[1].Adam, N. R. &Wortmann, J. C. (1989). Security-Control Methods for StatisticalDatabases: A Comparative Study. ACM Computing Surveys, Vol. 21, N. 4, pp. 515–556.Aggarwal, G., Bawa, M., Ganesan, P., Garcia-Molina, H., Kenthapadi, K., Mishra, N.,Motwani, R., Srivastava, U., Thomas, D., Widom, J., & Xu, Y. (2004). Vision Paper:Agrawal, R. &Srikant, R. (2000). Privacy Preserving Data Mining. In Proc. of ACM SIGMOD Conference on Management of Data (SIGMOD'00), Dallas, TX.

[2].Agrawal, R., Kiernan, J., Srikant, R., and Xu, Y. (2002). Hippocratic Databases.In Proc. 28th International Conference on Very Large Data Bases (VLDB'02), Hong Kong, China.

[3].Agrawal, R., Evfimievski, A., &Srikant, R. (2003). Information Sharing Across Private Databases. In Proc. of ACM SIGMOD International Conference on Management of Data

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.4. No.5, Pages : 19-24 (2015)
Special Issue of ICETEM 2015 - Held on October 27, 2015 in The Solitaire Hotel, Bangalore, India
http://www.warse.org/IJATCSE/static/pdf/Issue/icetem2015sp06.pdf

(SIGMOD'03), San Diego, CA. pp. 86–97.

[4]Agrawal, R., Srikant, R., & Thomas, D. (2005).Privacy Preserving OLAP.In Proc. ACM SIGMOD International Conference on Management of Data (SIGMOD'05), pp. 251–262.Agrawal, R., Asonov, D., Kantarcioglu, M., & Li, Y. (2006).Sovereign Joins.In Proc. of the 22nd International Conference on Data Engineering (ICDE'06).

[5]Blum, A., Dwork, C., McSherry, F., &Nissim, K. (Chawla, S., Dwork, C., McSherry, F., Smith, A., and Wee, H. (2005). Towards Privacy in Public Databases.In Proc. 2nd Theory of Cryptography Conference, pp. 363–385.

DeRosa, M. (2004).Data Mining and Data Analysis for Counterterrorism.Center for Strategic and International Studies Report, March 2004, 32 pages.

[6].Evfimievski, A., Srikant, R., Agrawal, R., and Gehrke, J. (2002). Privacy PreservingMining of Association Rules. In Proc. 8th ACM SIGKDD International Conference onKnowledge Discovery and Data Mining (KDD'02), Edmonton, Canada, pp. 217–228.

[7].Samarati, P. & Sweeney, L. (1998). Protecting Privacy when Disclosing Information: *k*–Anonymity and Its Enforcement through Generalization and Suppression. In Proc.Ofthe IEEE Symposium on Research in Security and Privacy, Oakland, CA.

[8].Wang, L., Jajodia, S., &Wijesekera, D. (2004). Securing OLAP Data Cubes AgainstPrivacy Breaches. In Proc. 2004 IEEE Symposium on Security and Privacy, pp. 161–175

[9].Zhong, S., Yang, Z., & Wright, R. N. (2005). Privacy-Enhancing *k*-Anonymization of Customer Data. In Proc. of the 24th ACM Symposium on Principles of Database Systems (PODS'05), June 13-15, 2005, Baltimore, MD, pp. 139-147.