TRI-Patternization on Generic Visualized Time Series Data

Sathiya.M
PG Scholar
Tamilnadu
India

Kumaresan.A
Head of the Department
Tamilnadu
India

Vijayakumar.K
Associate Professor
Tamilnadu
India

ABSTRACT

In the recent years, Privacy preserving techniques have been actively studied in the time-series data in various fields like financial, medical, and weather analysis. We are focusing towards preserving the data through Anonymity and Generalization technique. We first investigate, what's the privacy to be incorporated at the time-series data and after finding the data which need to be preserved various perturbation terminologies were identified and worked out towards secure multi-party computation (SMC) and encryption techniques in the distributed computing. Our project focused towards Generalized technique in which the data will be filtered or generalized in a grouped structure based on time series grouping algorithm and the data will be shown in the approximation format. So that, the data won't get disclosed. The second technique involves the display of data in the graphical format providing no clue about the exact data and approximation technique incorporates an exact preserving of data. The third technique involves the arrangement of data in the binary tree pattern and this provides an efficient way of ordering the data on the performance basis. The proposed system incorporates all the necessary features, In addition we are trying to incorporate security by adding a deformable/detectable noise to this time series data.

Keywords

K- anonymity; Generalization; Secure Multi-party Computation; Binary Tree Pattern; Graphical.

1. INTRODUCTION

In the recent years, Privacy preserving techniques have been actively studied in the time-series data in various fields like financial, medical, and weather analysis. We are focusing towards preserving the data through Anonymity and Generalization technique. Privacy protection of time series is a challenging study commonly used for the complex queries. In specifically, the most frequently used "complex" queries on time series not only covers the range queries in the attribute values at specified time instantaneous but also pattern similarity queries which treat each sequence more completely. Unfortunately, it is no insignificant task to support such variety of queries without disclosing the response to the information of individuals. The time series is one of the most important types of data stored in human society databases. The exposure of these data on the internet has taken care of the most creative applications range from product analysis of social network for tracking and pattern matching the information. However, such kind of larger data also imply huge amount of privacy but if not protected properly because it may benefit from the source for misuse and crimes. The most capable similarity search methods are techniques that perform dimensionality reduction on the data, then it uses a multidimensional index structure to index the data in the transformed liberty. The discrete Fourier transform (DFT) [12] perform dimensionality reduction method but the other techniques include singular vector decomposition (SVD) [13], discrete wavelet transforms [11] and the main idea is explained in more details the perturbation method. Perturbation reveals the true value leaks at a particular time. If we [10] desire to guarantee that such information doesn't help to deduce anything about the perturbation of additional time instant, then each time instantaneous must be perturbed separately of others. Conversely, if the time series reveal certain pattern, such perturbation independently each value in the time domain can be illustrated from the original data and filtered out. On the other hand, original series completely interchangeable by ensuring protection against any filtering method [14] requires. This can be achieved by perturbation and accurate copy of the data. In this case, yet a particular true value can disclose how all other values have been perturbed. Anonymization [10] of data points involve transforming the authentic data point into unrecognizable terms of the exact data values, by using the data generalization and suppression. The performance of an anonymization is a consequential loss in the quality of datasets. Most of the studies to minimize the information loss for a predetermined value of k have focused [14] in the algorithms.

2. RELATED WORK

In this section, we will discuss time series related information are hiding partially in the existing system. The existing partial hiding information can be divided into two approaches [14] the perturbation based approach and partition based approaches. Data can be protected by adding noise to the original information with the help of perturbation based approach[15]. Partition based approach is used to form a group in the case of disjoint set of data and each group were release some general informations. There are two approaches to form a cluster of similar information they are K-anonymity[16] and condensation[9] approaches. There exists number of defense mechanisms for protecting data for preserving privacy on time series. We explain 5 existing systems and their disadvantages.

A.1: A Condensation Approach to Privacy Preserving[9] Data Mining

Abstract

Here the issues raised by proposing a framework for Time Series Data Management (TSDM) is discussed.The inner idea of the anticipated domain specific framework is the view of Business Sections, Group of Time Series, and Time Series itself.The framework integrates least specification on the topic of structural and efficient aspects of time series data management.

Merits

- We pledge to the analysis that objects in temporal databases can be classified into three different categories

1. Time-invariant objects
2. Time-varying objects
3. Time series object

Demerits

- The framework has to have adequate clarity to attain the supplies of time series data and the management of that data.

A.2 On the complexity of optimal K-Anonymity[10]

Abstract

An valuable process of MAP transform calculation for time series with fixed time step is projected.The methods of claim of association measure to construction of association network of time series are discussed.

Merits

- Association network is represented by graph which is denoted as time series data.

Demerits

- Association measure can be used in clustering of time series and in solving different tasks in time series data mining which use similarity between time series.

A.3:Time Series Compressibility and Privacy[11]

Abstract

In this paper, we proposed an proficient introduced an competent technique to ascertain hybrid-multidimensional associative rules in synchronous, where each time series is associated with quantitative elements in multiple time series database. We recommend a two-step loom for this purpose: in the first step, our prominence is on ascertaining frequent outlines in different time series by doing chronological mining across time pieces.And in the next step, we focus on the quantitative characteristics of only those time series that are present in the patterns discovered in the first stage.

Merits

- Mining multiple time series data will be easier and it will provide an interesting outlines can be discovered. Our system provide an common and pertinent to any multiple time series dataset layout.

Demerits

- Using EMG database in multiple time series data is very difficult to mining the data.
- It is applicable for real data set that can be involving multiple sequence or time series outline that are associate with other detached attributes.
- This type of mining on EMG tributaries is in the bio-medical fields such as prosthetic designs, substantial medicines, and rehabilitations.

3. IMPLEMENTATION

Implementation is the stage of the project when the theoretical design is turned out into a working system.

3.1 Time Series Input module
3.2 Individual group Clustering module
3.3 Group Reweightage module
3.4 Graphical mapping integration module
3.5 Hierarchical time series classification module
3.6 Security evaluation metrics modules

3.1 Time Series Input module

In this module, The input from the user are gathered in the form of time series in order to proceed it with further process.

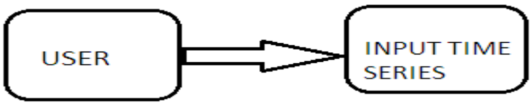


Fig:1.1 Input Dataset

3.2 Group Reweightage module

In this module, grouped data from previous modules are revamped with the addition of noise in the series as well as the similarity measures between them are checked individually.

Grouped data
Revamped Data
Check Similar Measures

Fig:1.2 Clusters the Similar Values

3.3 Individual group Clustering module

In this module, Similar set of input data's are grouped up using the clustering method individually. Cluster is mainly used to form a similar set of data grouped in this module. That is defined by three types of data are stored in database i.e., graphical representation, statistical graph and hierarchy. In these data to form a cluster.fig shows that the similar data are grouped in the cluster.

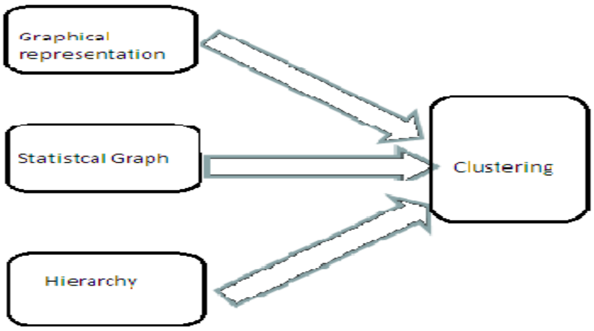


Fig: 1.3 Individual Clusters

3.4 Graphical mapping integration module

In this module, Added tuples are visualized graphically in every aspects that makes the data unpredictable thereby adding noise with it.

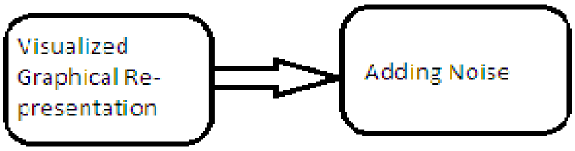


Fig: 1.4 Graphical data with noise

3.5 Hierarchical Time Series Classification module

In this module, fig represented time series are represented hierarchical with noise. In this, the hierarchical partition is made by considering the good and bad leaf segregation. The hierarchical structure of the VAR-HME models residential here allows the structure of extremely bendable forms to illustrate the non-stationary and non- linearities often present in multiple time series.

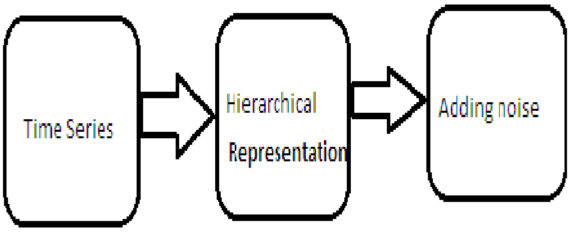


Fig:1.5 Hierarchical Representation with noise

3.6 Security Evalution Metrics module

In this module, the represented time series are analyzed for its security under metrics to check its preserving ability against prediction of time series.

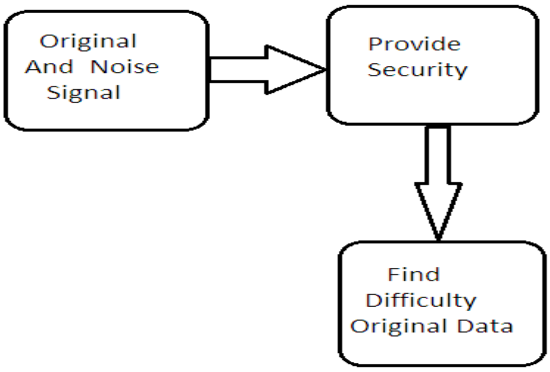


Fig:1.6 Security evaluation measures

4.Privacy Preserving Technique

4.1 Secure Multi-Party Computation

The goal of secure multiparty computation[17] job is for the contribution parties to *securely* compute some role of their disseminated and private inputs. Secure multi-party computation protocols are one of the first techniques used in preserving privacy for time series data in a distributed database.

While doing so, the protocol does not reveal. Multiparty computation allows N festivities to distribute a multiplication, from their own input and output can be inferred only for each and every learing parties.

For example, the parties can compute review information on their collective transaction logs, as well as cross-checking of the logs against counterparties to a transaction, without enlightening those logs. Our focus is the SMC protocol for distributed k-anonymity previously studied.

K-anonymity is a well-known privacy preservation technique proposed is to prevent linking attacks on shared databases. Linkage attacks are performed by adversaries who know some attributes (qi) of an individual to identity him/her in the dataset. A database is said to be k-anonymous if every tuple projected over the quasi identifier attributes appears at least k times in the databases.

Let [15][17]us Assume that there is a function that all parties wish to compute the secure computation. It shows how to compute the secure computation using the possible function in the safest way. In particular, the guarantees to produce the minimum information leakage of our result.

We focus[15] on the problem of secure multiparty computation

1. description pattern
2. probability of outcome
3. effectiveness concern

SMC is a collection of person with private inputs. Parties wish to combined their inputs to compute secure function. Properties must be ensured even if some of the parties maliciously attack the protocol. Certain security properties (like privacy and accuracy) are preserved

4.2 Graphical Representation

Graphical technique[18] makes use of produce the data and to summarize the descriptive measures. Descriptive information involves organizing, shortening, and present a group of data in such a way that constructively information is formed. The only acceptable computation on ostensible data is to count the occurrence of each value of the variable.

We can recapitulate the data in a table that presents the groups and their tally called a frequency sharing. A qualified occurrency distribution lists the categories and the quantity with which each occurs.

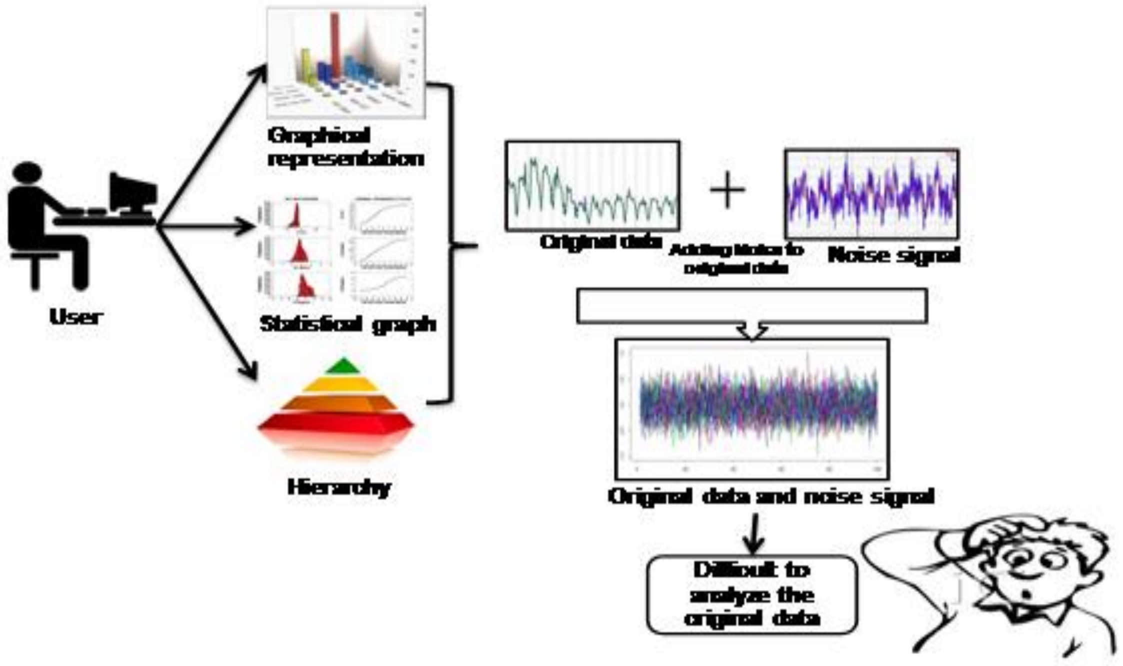

Graphical representation
User
Statistical graph
Hierarchy
Original data
Adding Noise to original data
Noise signal
Original data and noise signal
Difficult to analyze the original data

Fig:2 Overall Structure for Privacy Preserving Time Series Data

4.2.1 Graphical Techniques for Quantitative Data

There are several graphical methods that are used when the data are quantitative (i.e. Numeric, non-categorical). The most important of these graphic methods is the histogram. The histogram is not only a powerful graphical technique used to summarize interval data, but it is also used to explain the probabilities. Cross-sectional data are defined as the annotations considered at the identical point in time series data. Time series data are regarded as the successive points in time in the observed measure. Time-series data are displayed on a line chart, which design the assessment of the inconsistency on the perpendicular axis aligned with the time periods on the flat axis.

4.3 Binary Tree Pattern

Secure file sharing[19][20][21] method are efficiently developed for presenting a secure file cluster identity(id) .This each clusters are maintained by the user and then using the binary tree patterns are maintained each participants associated with the help of cluster id. Using the binary tree patterns with the file clustering id for providing privacy on time series data. In this binary tree pattern mainly used to operate the file distributing techniques.

The first phase of binary tree is used to allocate the secure file cluster id using the file protection package. The second scheme for providing privacy preserving on file distributing technique. Secure file sharing cluster id method used in the first phase of file sharing evaluation then its maintained the file by users from a file protection package(fpp).

The first phase of file sharing evaluates the process of secure file sharing id method for every file maintained by the users by a file protection package(FPP). It maintains the security related data for files stored in the form of FPP structure. A file block id is used for processing file sharing techniques and this id is assigned to file sharing methods. The FPP user data can be inspected rather than the other security result. File and block id should be recorded by FPP, so others do not create or modify the FPP and it does not make own security decision.

The second phase illustrates the procedure of a confidentiality perpetuation scheme for file sharing approaches using binary trees with a protected file and building block id for each file. The confidentiality perpetuation scheme of file sharing approaches is complete with secured file and building block id related to each member Id using binary trees.

4.3.1 Transfering Files And Cluster Id

The file protection [19] [21] package is a storage liberty which is assigned cluster id for each file and each participant can modify the data in the fpp without the information of the vendor of the particular file. The participant in the clusters that contain one or more files which is prepared to be transformed with the other users of the cluster. It is essential to assign the cluster id for each participant by outline the exploits of that contributor in the statement involved before the actions. File directory point(fdp) has been attached to [19] two examines which is substituted by fdp. Fdp is reduced xml tallying elements are described the size, name, shared data. Participants have permitted in a meticulous for other metadata stored in files. Meta data shared for each confidentially encompasses a possible data that is used as a symmetric encryption method for implement transfering the data in fdp. As an alternative of all data visible to share with an vigorous set of members, file protection package visibly classify the correct level of a determined collection of members and Respective cluster id of their files and the secondly information about which users have which files and then place separated data resources by streaming object search for the overlie. Finally,as a substitute of file cluster identities and participants by means of sources transferring data openly to receivers and data transfers are completed.

4.3.2 Binary Tree Pattern Designed For Privacy Preservation Scheme:

After allocating[21] each files and participants are maintained in the file and participant id's. in this section we are departing to perceive concerning how the files shared in a safe mode without releasing the private data of the participant by using binary trees. For each file maintained from the binary tree pattern framework which is followed in the communication path involved for each participant.

Participant A and participant B are [21][19]share their information by using their cluster id and then file id. The cluster id used to refer the exact data stored in the location in the record. The file id used to provides information about the file location. Participants sharing their file id and cluster id that uses a smaller amount of time to accomplish the file sharing techniques since it scrupulous the correct position of the data substance to be shared.

5. ALGORITHM

In our approach,We proposed the data fly algorithm for privacy preserving time series data.

DataFly

Sweeney regard as that the finest clarifications are the attributed that are accomplished after generalizing the elements with the largest part of individual values. The whole lattice is again search for a space because to find the solution of this approach is a very small number of nodes are having the lattice. The data fly approach very efficient from a time perspective.

Here is a summary of the Datafly algorithm:

Input: Private Table **PT**, quasi-identifier set **QI** = {A1,….,An}, anonymity parameter **k,** and hierarchies **DGHAi,** where i=1,…,n.

Output: **MT** a K-anonymization of **PT[QI].**

Method:

```
1. MT PT[QI]; freq Frequency list of MT

2.while there exists frequencies in freq less than k which together represent at least k rows do

    2.1. let Aj be attribute in MT having the most number of distinct values

    2.2. MT[Aj] generalized values of Aj according to DGHAj

    2.3. recalculate freq Frequency List of MT

3. Suppress rows in MT occurring less than k times.

4. enforce k requirement on suppressed rows in MT
```

5. EXPERIMENTAL RESULT

Step 1: user must transfer the data it should be revealed to the network, where the data should be in the following manner.

It may be,

(i) Graphical Representation
(ii) Statistical Graph
(iii) Hierarchy

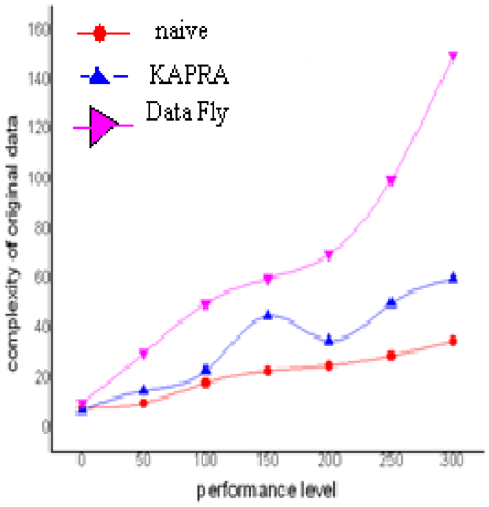
Our aim to reveal the original data only to the authorized person that means we should provide privacy for time series data. Eventhough, it should available for authorized persons only.

For that we are using three techniques which are already existing, but our system proposing a new idea to it.i.e, we are adding noise to the original data.

Step2: Addition of noise signal to the original data for making the hide the exact data.

Step3: Final step for provide authorized persons are only access the original data in the more secured manner. That is the data revealed in the network for more secured manner.

Fig.1 specifies the overall structure of our system. For understanding we have divided it into 4 steps.



6. CONCLUSION

In this paper, we have described and proposed a work of fiction anonymity representation called (k,P)- anonymity for time-series data. Relying on a nonspecific description to pattern representations, our reproduction may perhaps put off three types of connection attacks and in point of fact support the the majority extensively used queries on the anonymized data. We proposed a naive explanation and a more highly residential method called Data Fly to enforce (k,P)-anonymity on timeseries data. Our loom allowed for personalized data publishing and provided estimation methods to support queries on such data. The extensive experiments demonstrated the efficiency of (k,P)-anonymity in resisting linkage attacks while preserving the outline information of time series.

7. FUTURE ENHANCEMENT

The time series dat over the outperformed naïve elucidation in terms of conformity in pattern protection and runtime performance. Our results also illustrated the effectiveness and efficiency of the proposed inference methods for modified data publishing. (k,Pinscrutability may lead to a few tempting directions for future study. Our contemporary solution imposes a very severe constraint on PR correspondence and this may effect serious prototype loss. In the future work, we will consider loosing the PR equality circumstance on the assertion of ensuring privacy defense ability. This policy may greatly reduce the in sequence loss.

REFERENCES:

[1] G. Bhargava, P. Goel, and B.R. Iyer. Hypergraph based reorderings of outer join queries with complex predicates. In ACM SIGMOD Conference, pages 304–315, 1995.

[2] J.A. Blakeley, V. Rao, I. Kunen, A. Prout, M. Henaire, and C. Kleinerman. .NET database programmability and extensibility in Microsoft SQL Server. In Proc. ACM SIGMOD Conference, pages 1087–1098, 2008.

[3] J. Clear, D. Dunn, B. Harvey, M.L. Heytens, and P. Lohman. Non-stop SQL/MX primitives for knowledge discovery. In ACM KDD Conference, pages 425–429, 1999.

[4] E.F. Codd. Extending the database relational model to capture more meaning. ACM TODS, 4(4):397–434, 1979.

[5] C. Cunningham, G. Graefe, and C.A. Galindo-Legaria. PIVOT and UNPIVOT: Optimization and execution strategies in an RDBMS. In Proc. VLDB Conference, pages 998–1009, 2004.

[6] C. Galindo-Legaria and A. Rosenthal. Outer join simplification and reordering for query optimization. ACM TODS, 22(1):43–73, 1997.

[7] H. Garcia-Molina, J.D. Ullman, and J. Widom. Database Systems: The Complete Book. Prentice Hall, 1st edition, 2001.

[8] G. Graefe, U. Fayyad, and S. Chaudhuri. On the efficient gathering of sufficient statistics for classification from large SQL databases. In Proc. ACM KDD Conference, pages 204–208, 1998.

[9] C.C. Aggarwal and P.S. Yu, "A Condensation Approach to Privacy Preserving Data Mining," Proc. Ninth Int'l Conf. Extending Database Technology (EDBT), pp. 183-199, 2004.

[10] R. Dewri, I. Ray, and D. Whitley, "On the Optimal Selection of k in the k-Anonymity Problem," Proc. IEEE 24th Int'l Conf. Data Eng.(ICDE), pp. 1364-1366, 2008.

[11] D. Gunopulos and G. Das, "Time Series Similarity Measures,"Proc. Tutorial Notes of the Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (Tutorial PM-2), pp. 243-307, 2000.

[12] O. Abul, M. Atzori, F. Bonchi, and F. Giannotti, "Hiding Sequences," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE) Workshops, pp. 147-156, 2007.

[13] S. Papadimitriou, F. Li, G. Kollios, and P.S. Yu, "Time Series Compressibility and Privacy," Proc. 33rd Int'l Conf. Very Large Data Bases (VLDB), pp. 459-470, 2007.

[14] Lidan Shou, Xuan Shang, Ke Chen, And Chao Zhang, " Supporting Pattern-Preserving Anonymization For Time-Series Data Proc. Ieee 25th Computer Society,Pp.1041-4347,2013.

[15] Yehuda Lindell,and Benny Pinkas, "Secure Multiparty Computation for Privacy-Preserving Data Mining" The Journal of Privacy and Confidentiality , pp. 59-98, 2009.

[16] L. Sweeney, "k-Anonymity: Privacy Protection Using Generalization and Suppression," Int'l J. Uncertainty Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 571-588, 2002.

[17]Tutorial on Secure Multi-Party Computation. Yehuda Lindell. IBM T.J.Watson. Outline. Part 1: A rigorous approach to security; Defining security.

[18] "Graphical and Tabular Summarization of Data", www.utdallas.edu/~scniu/OPRE6301../Graphical_and_Tabular.pd.

[19] M. Balamurugan, S. Chenthur Pandian and J. Bhuvana, "Privacy Preservation for File Sharing Scheme Using Secured File Block id with Binary Trees", American Journal of Applied Sciences, ISSN: 1546-9239, 2013 (http://www.thescipub.com/ajas.toc).

[20] Likun Liua, liang Hub, DiWangc, Yanmei Huod, Lei Yange, Kexin Yang, " Two Noise Addition Methods For Privacy-Preserving Data Mining", I.J. Wireless and Microwave Technologies, pp. 28-33,2012.

[21] Michael Dairyko,Lara Pudwell,Samantha Tyner,Casey Wynn," Non-Contiguous Pattern Avoidance in Binary Trees", Published: Aug 23, 2012,Mathematics Subject Classifications: 05C30,05.