Effective and Efficient Data Aggregation Technique for Common Web Log XML Format for Digital Forensics Investigation (DFI)

Amit Pratap Singh
Research Scholar
Department of Computer Applications,
Samrat Ashok Technical Institute,
Vidisha, M.P.,India

Dr. R. C. Jain
Ex. Director
Samrat Ashok Technical Institute,
Vidisha, M.P., India

***Abstract---* In this paper talks about the process of data aggregation. Data aggregation is one of the method which is mostly used in reducing the size of the records set. Reduced data set could be helpful at many places. This reduction of record set can be done through data aggregation method. This data aggregation ca n be of two types as lossy and lossless. Here the proposed work is best on lossy technique. This lossy technique reduces the data for very large extends but it is useful only when original data regeneration is not required. The proposed work would not require any regeneration of original record set. This feature allows to use lossy aggregation technique. The proposed work is implemented in MATLAB.**

Keywords:- Data Aggregation, Weblog, XML, Lossy, Lossless.

I. INTRODUCTION

The primary concern of the EFF (The Electronic Frontier Foundation) is the use of data aggregation by the government to assemble personal profiles of American citizens and identify potential terrorists. In projects like Total Information Awareness, the government has proposed to use combinations of medical, financial, educational, employment, travel, and telephone records to search for patterns of transactions that are suggestive of terrorist activity. While the EFF understands the need for national security, TIA and similar projects are unlikely to be effective in practice and will raise too many security and privacy concerns in the process.

Although the most dangerous use of data aggregation is that of the government, data aggregation by private companies is also cause for concern. This is because companies can sell to the government information that it could not legally obtain on its own. Even companies that refuse to do business with the government can be subpoenaed. For these reasons, data aggregation by private companies poses the same risks that data aggregation by the government poses.

II. DATA AGGREGATION

Advancement in computing technology has led to the production of wireless sensors capable of observing and reporting various real world phenomena in a time sensitive manner. However such systems suffer from bandwidth, energy and throughput constraints which limit the amount of information transferred from end-to-end. Data aggregation is a known technique addressed to alleviate these problems.

The phenomenal growth in distributed wireless communication technology has led a novel paradigm known as sensor networks. They have been proposed for use in various applications including military and civilian applications. Many dynamically changing scenarios such as battlefield, commercial inventory must be monitored using adaptive methods that utilize critical, real-time information gathered from integrated low-powered sensors. With large number of sensor devices being quickly and flexibly deployed in these networks, each sensor device must be autonomous and capable of organizing itself in the overall community of sensors to perform coordinated activities with global objectives. The sensors are programmed to listen for events. When an event occurs, the sensors inform the end-point by generating wireless traffic. As the number of nodes in the sensor network increases the probability of congestion near events increases. This localized congestion leads to sub-optimal routing performance. Additionally, lot of packets gets dropped and the overall response time increases. Further, sensors around the event spend considerable amount of energy to transmit packets which finally do not reach the end point. [1, 3].

Data aggregation [2] is a technique which tries to alleviate the localized congestion problem. It attempts to collect useful information from the sensors surrounding the event. It then transmits only the useful information to the end point thereby reducing congestion and its associated problems.

A. Type of data aggregation

We propose three data aggregation techniques [2, 4, 5]:

- In-Network
- Grid-based
- Hybrid schemes to perform data aggregation.

Data in aggregation, without personally identifiable information, is normally sufficient to track usage patterns for administering online services. Any online service should be able to define a short period of time after which individual logs will no longer be needed for troubleshooting. Large databases are targets for malicious users, criminals, and terrorists. It is unlikely that even the most sophisticated database security technology with infinite funding will be immune to attack, especially as the data inside the databases are aggregated from more sources (and therefore of more value).

III. RELATED WORK

A. P-function set method for security conserving aggregation of information in WSN

In-network data aggregation presents a critical challenge for data privacy in resource constraint wireless sensor networks. Existing schemes based on local collaboration have unfavorable communication cost, and some other schemes based on secret sharing with the sink are low resistant to data loss. To address these issues, we propose a PAPF scheme, in which a novel p-function set taking advantage of the algebraic properties of modular operation is constructed. Thanks to the p-functions, nodes can perturb their privacy data without extra data exchange, and the aggregation result can be recovered from the perturbed data in the cluster head. Extensive analysis and simulations show that PAPF scheme is able to preserve privacy more efficiently while consuming less communication overhead, and has a good resistance to data loss [4].

B. PIP: Privacy and Integrity Preserving Data Aggregation in Wireless Sensor Networks

With the exponential rise of pervasive computing applications, data privacy has become much more of an important issue than before. When data are aggregated at each hop in a sensor network, it becomes harder to protect its privacy. A number of privacy preserving data aggregation algorithms have recently appeared for wireless sensor networks (WSNs), very few of them however also address the issue of data integrity along with privacy. Data privacy and integrity are two contrasting objectives to achieve in general. In a privacy preserved data aggregation, it becomes easier for an attacker to inject false data, hence we suggest that both privacy and integrity of data should be treated together. In this paper, we present an energy efficient, privacy preserving data aggregation algorithm which also preserves data integrity in WSNs. We analyze the security of the algorithm and provide proofs for confidentiality and integrity. We enhance this algorithm further to localize, to a certain degree, the corrupt aggregator. We provide the results of our implementation of the algorithm on TelosB motes, illustrating that both the computational overhead and the energy consumption are very low. Finally, we compare our algorithm with other schemes having similar objectives, demonstrating that our algorithm performs better in terms of band with usage and energy consumption in a WSN environment [3].

C. Data Aggregation ensuring a different level of privacy preserving

Data Aggregation ensuring a different level of privacy preserving: (DADPP) [6] this method offers a different level of aggregation of information along with the privacy preserving on different nodes for prior treatment of data. The idea of this method is stimulated by the work of Shao and the CDPA for different level of privacy and in accomplishing such issues respectively. In this approach a hierarchical structure is created and sensor is organized in the form of clusters containing cluster head below the base station node which is energy efficient. By taking into consideration the privacy level, each node belonging to same cluster is organized in different classes. The pre-treatment of data is within the same class and determination of privacy level is based on the size of the class. The privacy level of lowest rate partitioned the class having at minimum 3 sensor nodes. On the other hand privacy level of the highest rate partitioned the class having 4 sensor nodes. If a situation arises such as all sensor nodes of one cluster belonging to the same class, then this case is considered as the highest level of privacy. However, aggregation of information is similar as in CDPA. Firstly the pre-treatment of data is taking place in each individual class. After that cluster head is responsible for the aggregation of pre-treated information. Similarly the data are aggregated by cluster heads upto the base station. This hierarchical structure is presented in Figure 1. However, it reduces some amount of traffic by dividing a cluster containing n sensor node into multiple classes according to the desired privacy level, but it brings more aloft of communication and computations. Also, these aloft grow by expanding the levels of privacy.

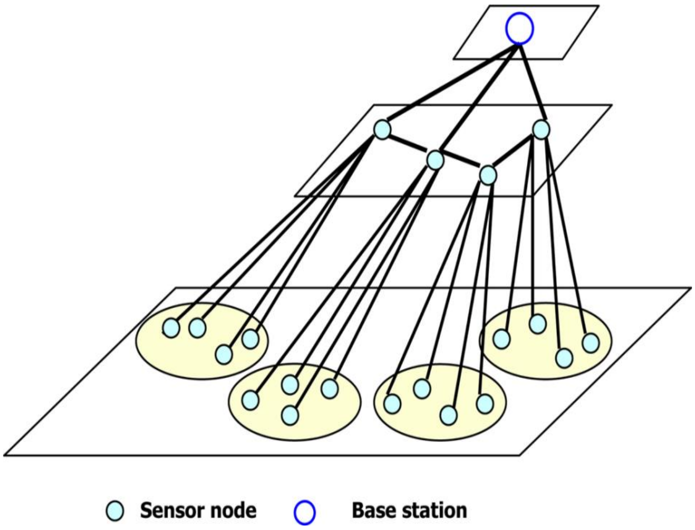


Fig 1 DADPP in a hierarchical network [6]

D. HCB aggregation of data for ensuring data integrity and privacy in WSN

In wireless sensor networks, the sensor nodes are restrained in resources and energy to address these issues data aggregation method has been proposed. An efficient data aggregation method is, when it would be able to achieve the privacy of the data. However, some efficient method for data aggregation and preserving privacy in wireless sensor network are CPDA [7], SMART [6], the Key dependent techniques, and Generic Privacy-

Preservation Solutions. But the existing disadvantage with this method is First, the associated cost of communication for network construction is considerably high.

TABLE 1: System Configuration

Model	Pentium Dual Core CPU 2.20Ghz
RAM	2GB
32 Bit Operating System	
Windows 7 Ultimate	

Second, the feature of data integrity is not backing .The two methods that back the feature of data integrity is, iCPDA and iPDA. But further the associated communication cost is high because for supporting integrity features more message is to be transmitted. To completely solve the existing problem, here a method called as Hillbert – curved based(HBC) aggregation of data for ensuring data privacy and data integrity in wireless sensor network is proposed. For decreasing the cost of communication here we used a tree based architecture for constructing the network and aggregating the data. For ensuring the privacy factor we used seed transfer discovery and Hillbert curved method for encrypting the data .For supporting the data integrity, we used integrity check discovery depend on the PIR method by having direct communication among child node and parent node. Last in term of performance our proposed method beats the other existing technique in energy consumption and preserving privacy of data. It is shows in figure 2.

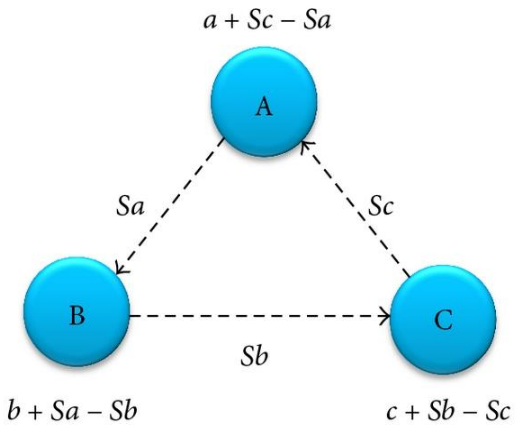


Fig 2: Original data change by seed exchange reduce communication cost [8]

IV. METHODOLOGY

In order to reduce the effort of any method, it is always better to reduce the size of the record set size. This size reduction is of two types: (i) Lossy and (ii) Lossless. Here the proposed methodology to perform size reduction is done by the means of lossy technique to reduce the effort as regeneration of original record set is not needed in the case of digital forensic detection.

The Proposed method for the reduction of record set is done based on aggregation technique. The proposed method of aggregation is discuss in algorithm mentioned bellow (fig 3).

Step 1: Open the AccessLog file

Step 2: Create the empty aggregated.xml file

Step 3: Make 2 lists one of 'normal' office hours and another of 'not normal' hours.

Step 4: Read xml file of Step 1 and Check the time and accordingly populate the 'normal' and 'not normal' lists

Step 5: Write these lists of normal and not normal into the xml file

Fig 3: Proposed Aggregation Method

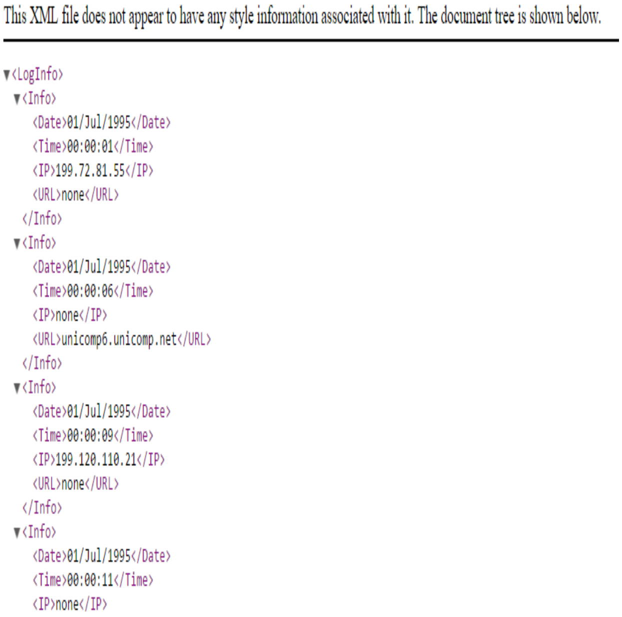
V. RESULT AND ANALYSIS

The system used for execution of the proposed method for record aggregation is as follows:

In order to evaluate the aggregation of the various data records through this propose method, two web log datasets are taken into consideration.

1. NASA weblog file[9]
2. comdotzone weblog file [10]

The input common XML file contains total 90860 records which gets aggregated by proposed method. Figure 4 show the structure of input record set which is in XML format.

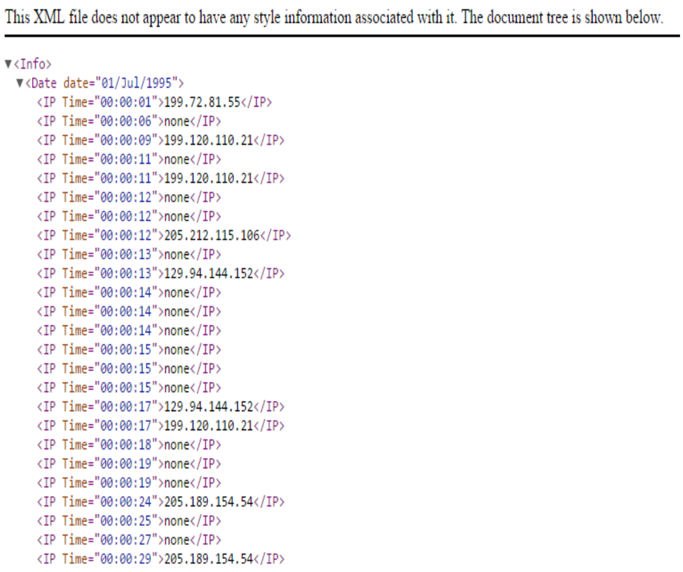


This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
▼<LogInfo>
  ▼<Info>
     <Date>01/Jul/1995</Date>
     <Time>00:00:01</Time>
     <IP>199.72.81.55</IP>
     <URL>none</URL>
   </Info>
  ▼<Info>
     <Date>01/Jul/1995</Date>
     <Time>00:00:06</Time>
     <IP>none</IP>
     <URL>unicomp6.unicomp.net</URL>
   </Info>
  ▼<Info>
     <Date>01/Jul/1995</Date>
     <Time>00:00:09</Time>
     <IP>199.120.110.21</IP>
     <URL>none</URL>
   </Info>
  ▼<Info>
     <Date>01/Jul/1995</Date>
     <Time>00:00:11</Time>
     <IP>none</IP>
```

Figure 4 show the structure of input record set which is in XML format.

The output XML file contains total only 3 records after proposed method. Figure 5 show the structure with records of output record set, which is in XML format. Here the aggregated file by Date (expanded nodes) is shown

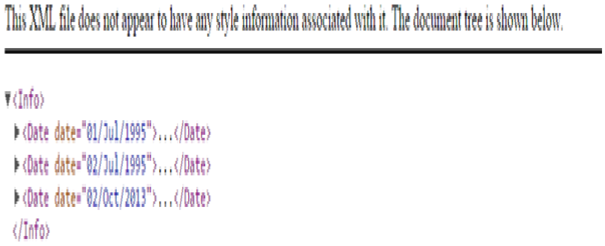


This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
▼<Info>
  ▼<Date date="01/Jul/1995">
     <IP Time="00:00:01">199.72.81.55</IP>
     <IP Time="00:00:06">none</IP>
     <IP Time="00:00:09">199.120.110.21</IP>
     <IP Time="00:00:11">none</IP>
     <IP Time="00:00:11">199.120.110.21</IP>
     <IP Time="00:00:12">none</IP>
     <IP Time="00:00:12">none</IP>
     <IP Time="00:00:12">205.212.115.106</IP>
     <IP Time="00:00:13">none</IP>
     <IP Time="00:00:13">129.94.144.152</IP>
     <IP Time="00:00:14">none</IP>
     <IP Time="00:00:14">none</IP>
     <IP Time="00:00:14">none</IP>
     <IP Time="00:00:15">none</IP>
     <IP Time="00:00:15">none</IP>
     <IP Time="00:00:15">none</IP>
     <IP Time="00:00:17">129.94.144.152</IP>
     <IP Time="00:00:17">199.120.110.21</IP>
     <IP Time="00:00:18">none</IP>
     <IP Time="00:00:19">none</IP>
     <IP Time="00:00:19">none</IP>
     <IP Time="00:00:24">205.189.154.54</IP>
     <IP Time="00:00:25">none</IP>
     <IP Time="00:00:27">none</IP>
     <IP Time="00:00:29">205.189.154.54</IP>
```

Figure 5 show the structure with records of output record set, which is in XML format, aggregated file by Date (expanded nodes).

Aggregated file by Date (collapsed nodes) – having 3 lines of 3 dates and it is shown in figure 6.



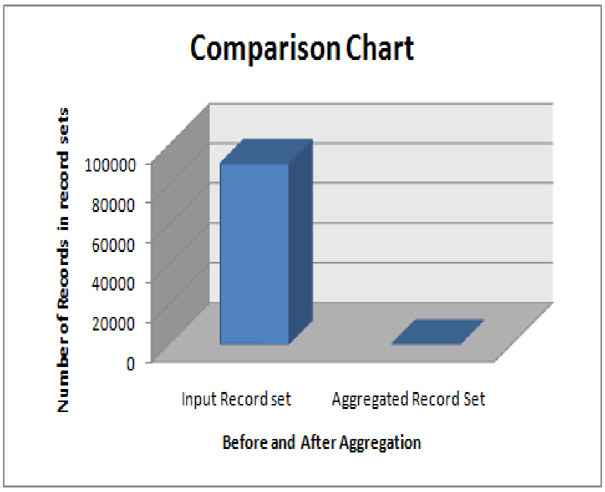
This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
▼<Info>
  ▶<Date date="01/Jul/1995">...</Date>
  ▶<Date date="02/Jul/1995">...</Date>
  ▶<Date date="02/Oct/2013">...</Date>
 </Info>
```

Figure 6: Aggregated file by Date (collapsed nodes)

TABLE 1: Number of Record

Input Record set	**Aggregated Record Set**
90860	3



Graph 1: Show the comparative number of records before and after aggregation method.

VI. CONCLUSION

This work investigates the Data aggregation problem with high level of efficiency, and presents an efficient algorithms Optimized Data Aggregation Method. From the secsion 5, researchers have shown and proved that the proposed algorithm works with great efficiency over these parameters for the taken dataset.

VII. REFERENCES

[1] Kelly Heffner, Rachel Popkin, Reem Alsweilem, Anjuli Kannan, “Report on Data Aggregation”, The Electronic Frontier Foundation Defending Freedom in the Digital World.

[2] Ramesh Rajagopalan and Pramod K. Varshney, “Data aggregation techniques in sensor networks: A survey, Electrical Engineering and Computer Science 2006.

[3] Ankit Tripathi , Sanjeev Gupta , Bharti Chourasiya, “Survey on Data Aggregation Techniques for Wireless Sensor Networks”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 7, July 2014.

[4] Report on “In-network Aggregation Techniques for Wireless Sensor Networks: A Survey”, Elena Fasolo, Michele Rossi, Jorg Widmer and Michele Zorzi.

[5] Neng-Chung Wang, Yung-Kuei Chiang, Chih-Hung Hsieh, and Young-Long Chen, “Grid-Based Data Aggregation for Wireless Sensor Networks”, Journal of Advances in Computer Networks, Vol. 1, No. 4, December 2013.

[6] Shaimaa Ezzat Salama and Mohamed I. Marie, “Web Server Logs Preprocessing for Web Intrusion Detection”, Computer and Information Science Vol. 4, No. 4; July 2011.

[7] K.R. Suneetha and R. Krihnamoorthi, “Identifying User Behavior by Analyzing Web Server Access Log” IJCSNS, 2009.

[8] David Lisak1, Lori Gardinier2, Sarah C. Nicksa2, and Ashley M. Cote2, “False Allegations of Sexual Assault: An Analysis of Ten Years of Reported Cases”, Symposium on False Allegations of Rape 2010.

[9] http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html.

[10] www.comdotzone.com.