IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 5, Ver. IX (Sep-Oct. 2014), PP 29-35
www.iosrjournals.org

Secure Transmission of Record after Record Linkage for Crime Detection Using AES

Rincy K Raj
(Dept. of Computer Science and Engineering, Mahatma Gandhi University, Kerala, India)

Abstract*:In many applications like crime detection, health sector, taxation sector etc… record linkage is used to find out the matched data items from different data sources. Finding matched records from different data sources corresponding to same entity is referred to as record linkage. It provides data integrity, data quality and also the reuse of existing data for advanced studies. The complexity of finding matching records is high due to the increased size of databases. The proposed system contain the secure information retrieval after efficient record linkage with indexing. AES algorithm is applied for the secure transmission of matched data. The indexing step generates candidate record pairs that are to be compared in record linkage process. After finding the matched records, it is sent to the user using secure AES algorithm in-order to avoid the malpractices.*
Keywords*:Record linkage, data integrity, data quality, indexing, AES algorithm.*

I. Introduction

A database is a collection of organized data so that it can be easily retrieved and updated. Database management systems (DBMSs) are specially designed applications that interact with the user, other applications, and the database itself to capture and analyse data.Database management system (DBMS) is a software system designed to allow the definition, creation, querying, update, and administration of databases. Different DBMSs can interoperate by using standards such as SQL and ODBC or JDBC to allow a single application to work with more than one database. A database is not generally portable across different DBMS. Data retrieval from different databases is a complex task since it involves the extraction of needed data records. Record linkage is a solution to this problem.

Several applications such as medical research, taxation sector, crime detection etc… need large amount of data from different databases. Mining of records from several databases that refers to same entity is (called record linkage) needed. The records to be matched refers to the entities like tax payers, customers, patients etc… Record linkage provides data quality, data integrity and efficient reuse of the data for advanced studies. Record linkage is widely used by both businesses and government agencies. Businesses might use record linkage techniques to remove duplicateentries from mailing lists, thereby reducing both printing and mailing costs, and otherwise operating moreefficiently. Businesses might also use record linkage to improve the functionality of their databases. Record linkage is an important tool in health research. In health sector matched data provides the detection of diseases, information's that improve health polices etc… It can be used to improve data holdings, data collection, quality assessment, and the dissemination of information. Data sources can be examined to eliminate duplicate records, to identify under-reporting and missing cases (e.g., census population counts), to create person-oriented health statistics, and to generate disease registries and health surveillance systems [2][3]. For example, foetal and infant mortality is a general indicator of a country's socioeconomic development, public health, and maternal and child services. If infant death records are matched to birth records, it is possible to use birth variables, such as birth weight and gestational age, along with mortality data, such as In crime detection, the investigators need files that are similar to the particular crime file under investigation. This is used to prevent crimes and terror as early as possible. Before transmitting the matched records to the investigators, it get encrypted using AES algorithm. Duplicate records returned by the search engines are to be removed.The most basic application of RL is identifying duplicates within a file or identifying duplicates across two files. In some situations, computerized record linkage can help preserve the confidentiality of information in a particular file or in a group of files.

Sophisticated linkage techniques are required in most cases since no unique identifiers are shared by all databases. The simplest kind of record linkage, called deterministic or rules-based record linkage, generates links based on the number of individual identifiers that match among the available data sets. The Probabilistic record linkage, sometimes called fuzzy matching (also probabilistic merging or fuzzy merging in the context of merging of databases), takes a different approach to the record linkage problem by taking into account a wider range of potential identifiers, computing weights for each identifier based on its estimated ability to correctly identify a match or a non-match, and using these weights to calculate the probability that two given records refer to the same entity.

The first step of indexing for record linkage is to standardise the data, which means make the data in databases into a standard form by avoiding the noises in the data and incompletely structured data into correct form. Most real world data are incomplete and it becomes a stumbling block for successful record linkage. Indexingprovides the candidate records for record linkage. It involves two phases build and retrieve. After RL the records are securely transmitted to the user using AES algorithm.

II. Related work

D.Dey, V.Mookerjee and D.Liu supported statistical record linkage for record linkage in their work [1, DebabrataDey, Vijay S. Mookerjee, and Dengpan Liu, 2011].. This paper refers statistical RL as a solution for several type of heterogeneity problem. Entity heterogeneity problem arises when same entity is represented using different names in different databases. The statistical record linkage produce some communication bottlenecks. To avoid this, a matching tree method is implemented which have sequential attribute acquisition, sequential identifier acquisition and concurrent attribute acquisition schemes for record linkage.

Peter Christen perform an experimental survey on different indexing techniques that can be performed for scalable RL. He explained RL as the process of matching records from several databases that refer to the same entities [4, Peter Christen, 2012]. By the indexing techniques discussed, the total number of record pair comparison can be reduced. He also explained deduplication as the method of RL performed on a single database. For each data set, three different blocking keys were defined using a variety of combination of record fields. String fields were phonetically encoded using Double-Metaphone algorithm. Run time results are also calculated. In his paper the six different indexing techniques are explained , also their complexity is analysed and their performance and scalability is evaluated using both synthetic and real datasets. He concluded in his paper that the q-gram based indexing is the overall slowest technique and array based sorted neighbourhood and the three suffix array based indexing technique achieve high run rate value for all data sets.

William E. Winkler provides an overview of methods and systems developed for record linkage in his paper [7, William E. Winkler, 2006]. This paper describes enhancement to a RL methodology which implements string comparator for string that do not agree with the character-by-character approach and a new assignment algorithm for forcing 1-1matching. Efficient string comparator function is used to deal with typographical variations of string and it is the first need of record linkage. The second need for RL is the estimation of matching record parameters and error rates. Third need is the 1-1 matching. Paper also describes string comparator, the parameter estimation algorithm and assignment algorithm, examples showing the empirical matches and a new method for estimating error rates.

Rohan Baxter, Peter Christen and Tim Churches explain different blocking schemes in their work [8, Rohan Baxter, Peter Christen and Tim Churches, 2003].This paper addresses the comparison between the old blocking schemes and new blocking schemes. The old blocking scheme include the standard blocking scheme and sorted neighbourhood scheme whereas the bigram indexing and canopy clustering with TFIDF (Term frequency/inverse document frequency)are considered as new blocking schemes . The new blocking technique provide scalable blocking and potential for large performance speedup and better accuracy. This paper includes the comparison of the speed and accuracy with the new blocking methods. Good blocking method reduces the number of comparisons between the records. In SB schemes blocking key is constructed from attributes of the record and is used to divide the data set into different blocks. Each record is compared with other record in the same block. Hence a record pair is constructed. In sorted neighbourhood sort the records based on the blocking keys .after that the window is moved over the sorted array and candidate record pairs are generated. In bigram based approach the blocking key is converted into bigram and based on this possible permutations will be built using the threshold (between 1.0 and 0.0). Canopy Clustering with TFIDF (Term Frequency/Inverse Document Frequency) forms blocks of records based on those records placed in the same canopy cluster. A canopy cluster is formed by choosing a record at random from a candidate set of records (initially, all records) and then putting in its cluster all the records within a certain loose threshold distance of it. The record chosen at random and any records within a certain tight threshold distance of it are then removed from the candidate set of records. We use the TFIDF distance metric, where bigrams are used as tokens.

E. Rahm and H. H. Do, in their paper "Data cleaning: Problems and current approaches" [6] addresses the data quality problems during data cleaning process and provide solutions for that. Data cleaning is required when integrating heterogeneous data sources and should be addressed with schema related data transformations. Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. Data quality problems are present in single data collections, such as files and databases, e.g., due to misspellings during data entry, missing information or other invalid data. When multiple data sources need to be integrated, e.g., in data warehouses, federated database systems or global web-based information systems, the need for data cleaning increases significantly. This is because the sources often contain redundant data in different representations. In order to provide access

to accurate and consistent data, consolidation of different data representations and elimination of duplicate information become necessary.

III. Proposed System Design

System design defines the architecture, components, modules, interfaces, and data for a system to satisfy the specified requirements. Systems design could be seen as the application of systems theory to product development. There is some overlap with the disciplines of systems analysis, systems architecture and engineering This section includes the architecture of record linkage process, its explanations and the AES algorithm applied at the end user for secure data transmission. This is implemented in the crime investigation application.

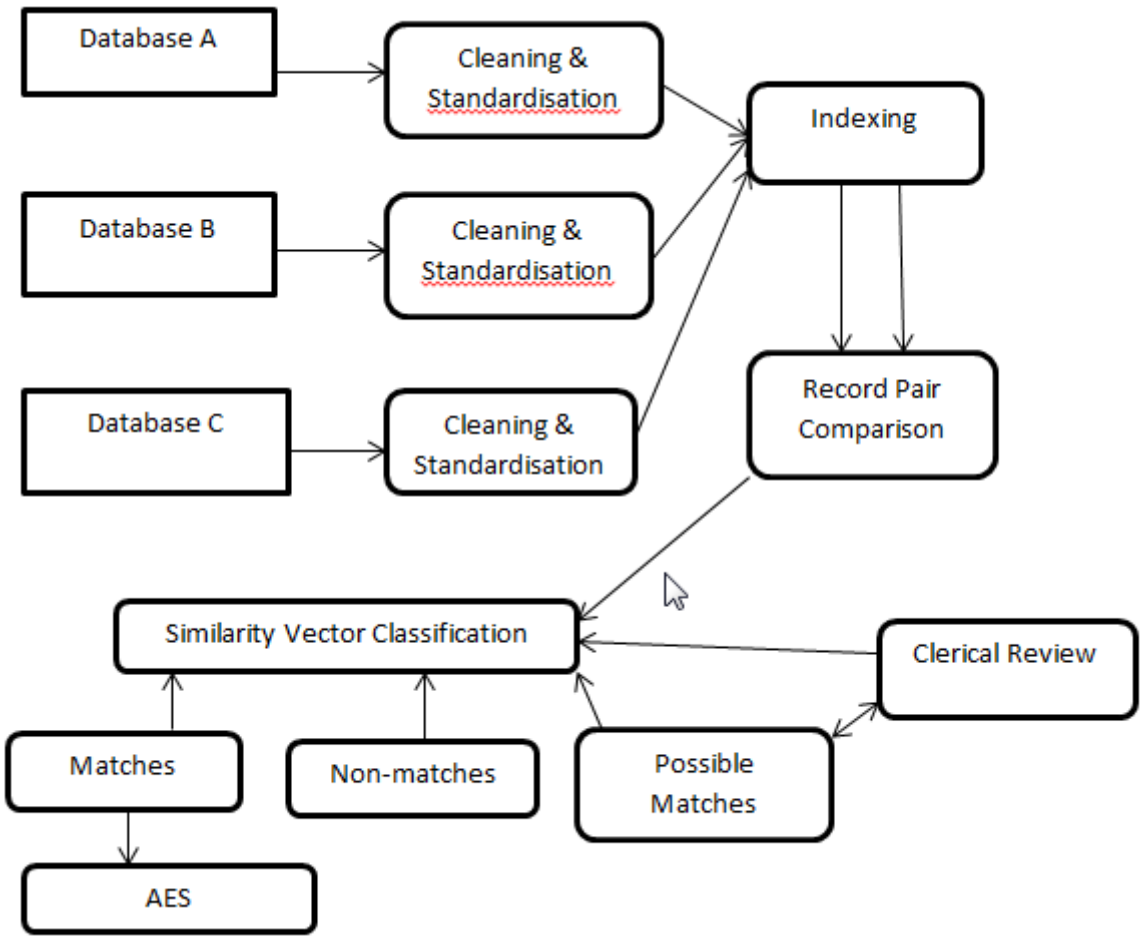

Fig 1. Architecture of Proposed System

The architecture defines the structure of the system. It includes different databases, cleaning process, indexing module, record linkage etc… The main modules of the proposed system include cleaning and standardization, indexing, record linkage, similarity vector classification, clerical review and secure transmission of data.

3.1Cleaning And Standardization

Data cleaning is the process of detecting and removing incorrect data from records, tables or databases. After cleaning, a data will become consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage. Data cleaning process include the elimination of typographical errors or validating and correcting values against a known list of entities. Standardization of data is a means of changing a reference data set to a new standard, ex, use of standard codes.The main task of data cleaning and standardization is the conversion of the raw input data into well defined, consistent forms, as well as the resolution of inconsistencies in the way information is represented and encoded.

According to E. Rahm and H.H. Do [6] data cleaning includes several phases such as data analysis, definition of transformation work flow and mapping rules, verification, transformation and backflow of clean

data. The data analysis phase include removing inconsistencies in data, detecting errors etc… The definition of transformation workflow and mapping rules include a large number of data transformation and cleaning steps based on the degree of heterogeneity and dirtiness of data. The verification phase include the testing and evaluation of the correctness and effectiveness of a transformation workflow and the transformationdefinitions. Transformation phase involves the execution of the transformation steps either by running the ETL workflow for loadingand refreshing a data warehouse or during answering queries on multiple sources. In the last phase after (single-source) errors are removed, the cleaned data should also replacethe dirty data in the original sources in order to give legacy applications the improved data too and toavoid redoing the cleaning work for future data extractions.

The cleaning process is performed as follows. All input text strings are store as lowercase letters and age is calculated from date of birth. By storing text strings as lowercase letters normalization can be performed. These text strings and age can be used for the preparation of BKV. The data from three different data bases are to be used for linkage process and that will be the input to this module. Cleaned data are given to the next module.

3.2Indexing Process

Based on the indexing scheme the records are arranged. It helps to divide the whole data set into groups or blocks. This will reduce the number of record pair comparison. There are so many indexing schemes are available. Standard blocking, sorted neighborhood indexing, canopy clustering, Q-gram based indexing, suffix array based indexing, and String map based indexing. Simplest approach array based sorted neighborhood are the fastest technique and is selected as the indexing technique in this project. SNI was proposed in 1990. In this approach, a BKV is generated from the attribute values of each record. The BKV act as a sorting key. BKVs from all databases will be inserted into one combined array and then sorted alphabetically. Sort each database according to the BKVs and sequentially move a window of a fixed number of records w (w > 1) over the sorted values. Candidate record pairs are generated in such a way that for each pair one record is selected from each of the two databases. The total number of candidate record pairs generated depends upon the window sizew, and on the harmonic mean of the sizes of the twodatabases that are linked. [Peter Christen, 2012].Fig 2 shows an example for sorted neighborhood approach.

Window positions	BKVs (Surname)	Identifiers
1	Millar	R6
2	Miller	R2
3	Miller	R8
4	Myler	R4
5	Peters	R3
6	Smith	R1
7	Smyth	R5
8	Smyth	R7

Window range	Candidate record pairs
1 – 3	(R6,R2), (R6,R8), (R2,R8)
2 – 4	(R2,R8), (R2,R4), (R8,R4)
3 – 5	(R8,R4), (R8,R3), (R4,R3)
4 – 6	(R4,R3), (R4,R1), (R3,R1)
5 – 7	(R3,R1), (R3,R5), (R1,R5)
6 – 8	(R1,R5), (R1,R7), (R5,R7)

Fig 2. Example for sorted neighbourhood technique with window size 3.

Consider the length of the sorted array is:

$$(n_A + n_B)$$

then the number of window positions equals:

$$n_A + n_B - w + 1$$

While for de-duplication the number of windows is :

$$n_A - w + 1$$

Since the window size is fixed in this approach, the number of candidate record generated depends on the window size and the size of the database and independent of the frequency distribution of blocking key value.If

$$\alpha = \frac{n_A}{n_A + n_B}$$

denotes the ratio of the number of records in database A (crime database) over the number of records in both databases and

$$\beta = \frac{n_B}{n_A + n_B} = (1 - \propto)$$

the corresponding ratio for database B (hospital database) then for a record linkage the number of unique candidate record pairs generated equals :

$$U = (\alpha w)\,(\beta w) + (nA + nB - w) \times (\alpha\,((w-1)\,\beta) + \beta\,((w-1)\,\alpha))$$
$$= \frac{n_A n_B}{(n_A + n_B)^2}(w^2 + 2(nA + nB - w)(w-1))$$

The first term in this equation equal to the number of candidate record pairs that are generated in the remaining (nA+nB+-w) window positions. The complexity of sorting array is O(n log n), where n= nA + nB.

3.3Record Pair Comparison

The indexing step generates pairs of candidate records which will be compared in detail in the comparison step using a variety of comparison functions appropriate to the content of the record fields.The compared candidate records are classifieds into matches, non matches and possible matches depending upon the decision model used. Approximate string comparisons, which take typographical variations into account, are commonly used. Record pair comparison results probability of match, possible match and miss match. The calculation is based on the calculated probability of selected attribute.

There are two kinds of record linkages are defined, Deterministic and Probabilistic [5, Ivan P. Fellegi and Alan B. Sunter, 1969]. The simplest kind of record linkage, called deterministic or rules-based record linkage, generates links based on the number of individual identifiers that match among the available data sets. Two records are said to match via a deterministic record linkage procedure if atleast some identifiers are identical. Deterministic record linkage is a good option when the entities in the data sets are identified by a common identifier, or when there are several representative identifiers e.g., name, date of birth, and sex when identifying a person, whose quality of data is relatively high. The most simple deterministic record linkage strategy would be to pick a single identifier that is assumed to be uniquely identifying, and declare that records sharing the same value identify the same person while records not sharing the same value identify different people.

The Probabilistic record linkage, otherwise called fuzzy matching (also probabilistic merging or fuzzy merging in the context of merging of databases), takes a different approach to the record linkage problem. It takes a wider range of potential identifiers into account, computing weights for each identifier based on the estimated ability to correctly identify a match or a non-match record, and using these weights calculate the probability that two given records refer to the same entity. Record pairs with probabilities above a certain threshold are considered to be matches, while pairs with probabilities below another threshold are considered to be non-matches. And the pairs that fall between these two thresholds are considered to be "possible matches" and can be dealt with accordingly.deterministic record linkage wants a series of potentially complex rules to be programmed ahead of time, probabilistic record linkage methods can be "trained" in order to perform well with much less human intervention.

3.4 Similarity Vector Classification

The matched records from record pair comparison module is submitted to this module and is classified into three blocks by this module. The task is based on the resulted probabilities of the previous module. Using these similarity values, the next step in the record linkage process is to classify the compared candidate record pairs into matches, non-matches, and possible matches, depending upon the decision model used. Non matches include record pairs that were removed in the indexing step without being compared explicitly. This section describes techniques that have been applied for matching fields with string data, in the duplicate record detection context. The techniques include Character-based similarity metrics, Token-based similarity metrics, Phonetic similarity metrics and Numeric Similarity Metrics. To handle well typographical errors character based similarity metric is used. Token based similarity metric is used for string type data. Strings may be phonetically similar but can have variations in a character or token level. For example the word Kageonne is phonetically similar to Cajun despite the fact that the string representations are very different. The phonetic similarity metrics are trying to address these issues and match such strings. Numerical similarity deals with the numerical type of data.

3.5 Clerical Review

A clerical review process is required for the record pairs in the possible matches where these pairs are manually assessed and classified into matches or no matches. This module provides more accuracy in the linkage process. The overall linkage process can be verified here. This module checks whether the linking of

partial matched records are relevant or not. Measuring and evaluating the quality and complexity of a record linkage project can be done in evaluation phase. The three searching methods used in this project are Sequential Attribute Acquisition (SAA), Sequential Identifier Acquisition (SIA) and Concurrent Attribute Acquisition (CAA). Among these CAA is the most efficient searching method.

3.6 Secure Transmission Of Data

While using applications there arise the need of secure data record transmission at the user level. AES algorithm is implemented for security. The application used is the crime detection. In this application there is two registration one is the user registration and other is the investigating officer registration. Admin have the power to add or remove the user and officer. User register in the site with relevant details and get an id and password. Using this user can register his/her complaints. Similarly officers should register in the site. Investigating officer search for some information related to some cases. Then the admin sent the details available in the databases in an encrypted form using AES algorithm in order to avoid the malpractices and third party interfere. AES approach is iterative rather than Feistel cipher. In this the key is expanded into array of 32-bit words. It has a simple structure with four different stages such as subbytes, shiftrows, mix columns andAddRoundKey. AddRoundKey is a form of vernam cipher. Decryption uses keys in reverse order and decryption recover plaintext. Final step have only three stages.

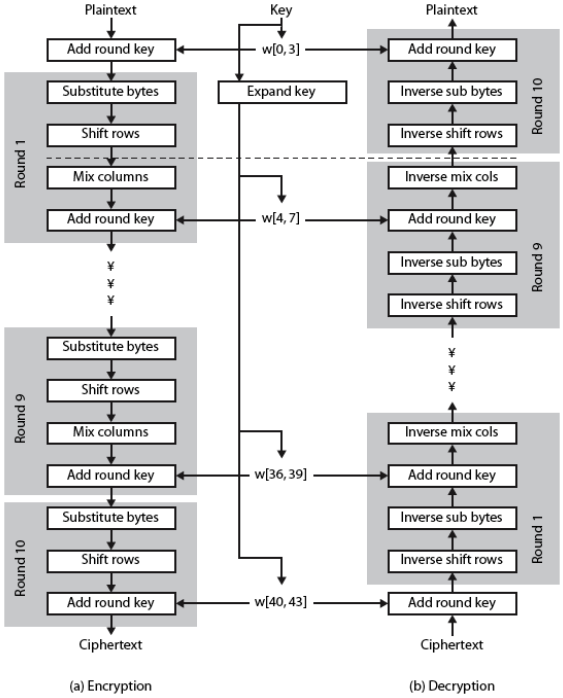

Fig 3. AES structure

IV. System Analysis

RL technique is an efficient technique to link records which are available online. Crime records can be managed by using this technique and the secure transmission of data to the end user is a new thing. It makes the criminal recordsearching more easily and prevent third party interference. The advantage of the system include, it will avoid problem associated with manual searching**.**By this work an investigating officer can collect accused details from different databases by a single click. The details will be in encrypted form which can be decrypted to get the original data. The AES algorithm is implemented for this and it provides high security and speedup the transmission of data's. Communication overhead is reduced in this approach. RL is a cheap technique compared to other studies which involve direct data collection and it also reduces the survey burden. Record linkage yields larger sample sizes at equal costs and provide high data quality. This work explains the simplicity of online record linkage in the area of searching and to differentiate the three searching methods – SIA, SAA and CAA. Using this application investigating officer can simply search and retrieve accused details from different database without knowing the difference. There is no need to understand the difference of databases by the investigating officer. The system will collect information from different department's database and combine those similar data records and provide complete details of the particular entity. This system help the user to register the complaints without going to the police station.

V. Conclusion

Crime record management using AES and RL technique is a new thing to the real world. Using this technique investigating officer can simply search and retrieve accused details from different database without knowing the difference using a single click. RL system can overcome many of the existing problems of the traditional searching retrieval system. There is no need to understand the heterogeneity of databases by the officer. The system will collect information from different department's database and combine those similar data records and provide complete details of the particular entity using AES encryption technique to the officer. Using this technique the users can register their complaints without going to the police station and the investigation officers get some assumption from the records provided as a result of search.

References

[1] D. Dey, V. Mookerjee, and D. Liu, "Efficient techniques for online record linkage," IEEE Transactions on Knowledge and Data Engineer- ing, vol. 23, no. 3, pp. 373–387, 2010.
[2] D. E. Clark, "Practical introduction to record linkage for injury research," Injury Prevention, vol. 10, pp. 186–191, 2004.
[3] C. W. Kelman, J. Bass, and D. Holman, "Research use of linked health data – A best practice protocol," Aust NZ Journal of Public Health, vol. 26, pp. 251–255, 2002.
[4] Peter Christen, "A Survey Of Indexing Techniques For Scaable Record Linkage and Deduplication," vol. 2 no.9 2012
[5] I.P. Fellegi and A.B. Sunter, "A Theory of Record Linkage," J. Am.Statistical Assoc., vol. 64, pp. 1183-1210, 1969.
[6] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," IEEE Data Engineering Bulletin, vol. 23, no. 4, 2000.
[7] W. E. Winkler, "Overview of record linkage and current research directions," US Bureau of the Census, Tech. Rep. RR2006/02, 2006.
[8] R. Baxter, P. Christen, and T. Churches, "A comparison of fast blocking methods for record linkage," in ACM IGKDD'03 work-shop on Data Cleaning, Record Linkage and Object Consolidation, Washington DC, 2003, pp. 25–27.
[9] W.E. Winkler, "Advanced Methods of Record Linkage," Proc. Section Survey Research Methods, pp. 467-472, 1994.
[10] AES Encryption and Decryption – Available: http://en.wikipedia.org/wiki/AES_cipher.