Detection of Data Breaches Using Cuckoo Filter with Skip-Gram

Shinde Monika C.[1], Chavan Preeti L.[2], Phalphale Jyoti V.[3] ,

Somavanshi Gauri B.[4],Anpat Ram B.[5]

[1-4]BE Student, Dept .Of Computer Engineering, SBPCOE, India
[5]Professor, Dept .Of Computer Engineering , SBPCOE, India

---***---

Abstract - *In many organizations there is various type of sensitive data is stored to prevent that data from data breaches is big risk. During data transformation various types of data breaches can be happen. In data breaches the data can be leaked due to lack of data encryption and error which are created by humans or the attackers. It is big risk for managing the security problem. It is very important to prevent the sensitive type of data. The data prevention is much important for various applications. So preventing sensitive type of data is challenging due to data transformation from unauthorized access. In this paper we are using the advanced sampling and parallel algorithms which can detect the data breaches. In previous researches the data breaches can be detected fast using bloom filter with n-gram technique. In this paper we are using cuckoo filter with skip gram technique. The cuckoo filter is an alternative for bloom filter. Cuckoo filter having advanced properties like space optimization and detection of elements present in data structure. The skip gram is alternative to n-gram. The skip gram technique removes the data sparsity problem. Our system results better performance in data breach detection and prevention in many organizations.*

Key Words: Data breach detection, Cuckoo filter, Cuckoo hash function, false positive rate, Skip gram.

1.INTRODUCTION

Reports show that the total number of reported data breaches reached an all time high of 3,930 in 2015 exposing over 736 million records. 64.6% of reported breaches due to hacking 288 incidents involved third parties. Three 2015 incidents have secured a place on the top 20 All Time Breach List. The number of reported incidents tracked by Risk Based Security has exceeded 18,900 exposing over 4.6 billion records [2]. So preventing the Sensitive type of data from unauthorized access is very important. This is very challenging to kept this sensitive data away from breaches. The data breaches are nothing but leakage of data intentionally or unintentionally by unauthorized access. During data transformation various types of data breaches can be happen. In previous researches the data breaches can be detected fastly using bloom filter with n-gram technique [1]. It can be used to detect and prevent the data breaches. Also we are going to use some sequence alignment technique for the detection of data breaches. The cuckoo filter will be useful to improve the space efficiency. In this paper we are going to use cuckoo filter with skip gram

2. PREVIOUS WORK

In previous researches data breach detection system like Symantec DLP uses N-gram set intersections with bloom filters[12]. The bloom filter is space saving but as bloom filters are used with N-gram it will creates accidental matches and false positive data. Also the alignments algorithms are used in various applications as in privacy preserving sequence matching have been studied [3]. Various types of tools are developed for secure the sensitive data on mobile devices [4-5]. The data leak from database issues are described in the research of anomaly detection [6]. In cloud computing privacy preserving risk is addressed in a data protection models for service provisioning in the cloud [7]. The privacy-preserving data leak detection was introduced and developed [8-9]. Where data leak detection operations are outsourced to a semi-honest third party.

2.1 Bloom Filter

Bloom filter is a data structure which will identify the membership to a static data set. The bloom filter allows the small rate of false positive data set. In spam filtering the cuckoo search is used for increase the performance of bloom filter. The bloom filter more space than the other data structure. In bloom filter the deletion of element cannot work. To improve performance of the bloom filter we are going to use cuckoo filter which is alternative for bloom.

2.2 N-GRAM

N-gram is used in various functions such as when growing a language model. It is also used for text mining. The benefits of n-gram model is simplicity and scalable to use. There are different types of n-gram model such as uni-gram, bi-gram, tri-gram, etc.

Consider sentence,

"There are seven days in week"

1. Uni-gram: There, are, seven, days, in, week.

2.Bi-gram:There-are,are-seven,seven-day,days-in, in-week.

3.Tri-gram:There-are-seven,are-seven-days,seven-days-in,

days-in-week.

It is also used in systematic appropriate string matching algorithm. By using this we can convert the different items in a set of n-gram. But the n-gram has some disadvantages like data sparsity problem. So, to removes this problem skip-gram technique is used.

3. PROPOSED SYSTEM

3.1 PROBLEM STATEMENT

The objective of proposed system is to develop a system to detect and prevent data breaches occurred in different applications like Military surveillance, Health Organization, Banking Systems, Education system etc. by using Cuckoo Filter with Skip-gram technique.

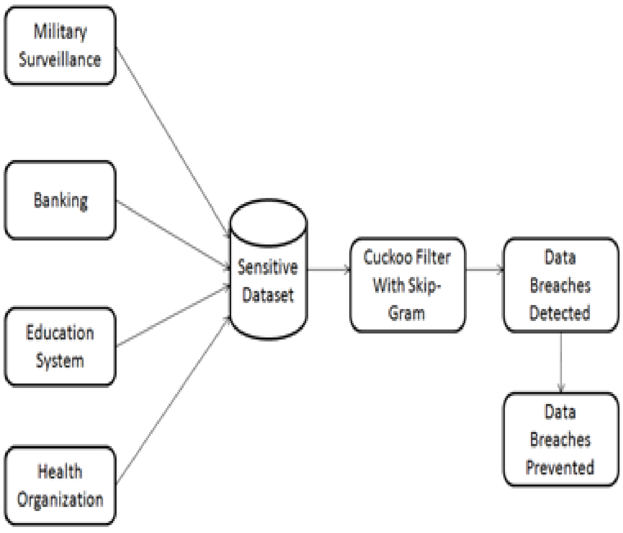

Fig-1: Block Diagram of proposed system.

3.2 WORKING SYSTEM

In our system we are going to perform the detection and prevention of data breaches occurred in various organizations. Fig.1 shows the block diagram of our system. Firstly, we are going to collect the data from those organizations. And the data which want to be kept private will be protected by our system. The systems consist of cuckoo filter and the skip-gram technique. The system will first find that whether the data breach is occurred or not. If there is a data breach is occurred then our system will going to detect that breach and by applying our algorithm, we are going to prevent that sensitive data. We are also using sequence alignment technique for preventing and detecting data breaches .And keep that sensitive data away from the unauthorized access.

A. CUCKOO FILTER

The cuckoo filter uses the cuckoo hashing to solve the problem of hash collisions. The cuckoo filter having same advantages like it supports adding and removing items dynamically. It gives higher lookup performance. The cuckoo filter gives high space holding capacity because it filter the item-placement decision earlier when new item is inserted. The lookup is performing fastly in a cuckoo filter than bloom filter. Cuckoo filter reduce the false positive probability. The research shows the cuckoo filter is practically better than a bloom filter [10].

B. SKIP-GRAM

Skip-grams are generalization of n-gram in which components (typically words) need not be consecutive in the text under consideration, but may leave gaps that are skipped over. They provide one way of overcoming the data sparsity problem found with conventional n-gram analysis [11].The skip-gram technique uses k-skip-n-gram. Fig3 shows the skip gram model. In our paper, we are using a skip-gram technique for sequence alignment.

Consider a sentence for k-skip-gram as a 2-skip-bi-grams and tri-grams compared with the bi-grams and tri-grams.

"There are seven days in week"

Bi-grams: - {There are, are seven, seven days, days in, in week}

2-skip-bi-gram:-{There are ,There seven, There days, There in, There week, are seven, are days, are in, are week, seven days, seven in, seven week, days in, days week, in week}

Tri-gram :-{ There are seven, are seven days, days in week}

2-skip-tri-gram:-{There are seven, There are days, There are in, There are week, There seven days, There seven in, There seven week, There days in, There days week, There in week, are seven days, are seven in, are seven week, are days in, are days week, are in week, seven days in, seven days week, seven in week, days in week}

The above example shows the working of skip-gram model.

4. APPLICATIONS

Following are some applications of our system.

4.1 Organizations:

Our system is useful to protect the data breaches occurred in organizations.

4.2 Banking System:

Bank system consist of sensitive type of data like credit card numbers, passwords, account numbers of clients these data may be breached by third party so to detect the data breach and prevent that data breach our system will be a helpful for banking system also.

4.3 Military Surveillance:

Military system also contains sensitive data so proposed system will be useful.

Everyone wants to kept their data private and safe from breaches our system will be helpful to detect and prevent that data breaches.

5. CONCLUSIONS

In our paper we introduced technique which will detect and prevent the data breaches from unauthorized access. We introduce the cuckoo filter and skip-gram technique with sequence alignment to detect and prevent the data breaches. Our approach is based on sequence alignment technique we are going to detect the data breaches occurred in organizations like military surveillance, education system etc. And if the data breach is occurred then we are going to prevent the sensitive type of data from breaches. For future work we plan to improve the performance of our approach.

REFERENCES

[1] Xiaokui Shu, Jing Zhang, Danfeng (Daphne) Yao WuChun Feng "Fast Detection of Transformed Data Leaks." IEEE Transaction On Information Forensics and Security, 3, March, 2016.

[2] Feb. 2015). *Data Breach QuickView: 2014 Data Breach Trends*[Online].Available:https://www.riskbasedsecurity.com/reports/2014-YEDataBreachQuickView.pdf, accessed Feb. 2015.

[3] M. Blanton, M. J. Atallah, K. B. Frikken, and Q. Malluhi, "Secure and efficient outsourcing of sequence comparisons," in *Proc. 17th Eur. Symp. Res. Comput. Secur.*, 2012, pp. 505–522.

[4] R. Hoyle, S. Patil, D. White, J. Dawson, P. Whalen, and A. Kapadia, "Attire: Conveying information exposure through avatar apparel," in *Proc. Conf. Comput. Supported Cooperat. Work Companion (CSCW)*, 2013, pp. 19–22.

[5] Z. Yang, M. Yang, Y. Zhang, G. Gu, P. Ning, and X. S. Wang,"AppIntent: Analyzing sensitive data transmission in Android for privacy leakage detection," in *Proc. 20th ACM Conf. Comput. Commun. Secur.*,2013, pp. 1043–1054.

[6] Bertino and G. Ghinita, "Towards mechanisms for detection and prevention of data exfiltration by insiders: Keynote talk paper," in *Proc. 6th ACM Symp. Inf., Comput. Commun. Secur. (ASIACCS)*, 2011,pp. 10–19

[7] Lin and A. Squicciarini, "Data protection models for service provisioning in the cloud," in *Proc. 15th ACM Symp. Access Control Models Technol.*, 2010, pp. 183–192.

[8] Shu, D. Yao, and E. Bertino, "Privacy-preserving detection of sensitive data exposure," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 5, pp. 1092–1103, May 2015.

[9] Liu, X. Shu, D. Yao, and A. R. Butt, "Privacy-preserving scanning of big content for sensitive data exposure with MapReduce," in *Proc. 5th ACM Conf. Data Appl. Secur. Privacy (CODASPY)*, 2015, pp. 195–206.

[10] Bin Fan, David G. Andersen, Michael Kaminskyy, Michael D. Mitzenmacherz," Cuckoo Filter: Practically Better Than Bloom".

[11] David Guthrie, Ben Allison, Wei Liu, Louise Guthrie, Yorick Wilks "A Closer Look at Skip-gram Modelling".

[12] Symantec. (2015). *Symantec Data Loss Prevention*. [Online].Available:http://www.symantec.com/data-loss-prevention, accessed Feb. 2015.