# Data Leakage Detector for Social Media Users Using an Improved Ant Bee Colony (ABC) Algorithm

Blessing C. Okoro[1], Friday E. Onuodu[2]

[1]*Department of Information Technology, National Open University of Nigeria, Nigeria*
[2]*Department of Computer Science, University of Port Harcourt, Nigeria*

***Abstract*— There is an alarming rate of data leakage in most social networks and not every social media user takes data security seriously. A major challenge faced by most Social Media users is the problem of data leakage. There are basically two major data leakage problems, they include Malicious Data Leakage (MDL) and Inadvertent Data Leakage (IDL) respectively. Despite the various improvements of data security by different encryption algorithms, there are still open problems and occurrences of data leakages on Social Media. In this work, we developed an improved Ant Bee Colony (ABC) algorithm for Data Leakage Detector for Social Media Users. We adopted Rapid Application Development Methodology (RAD) in this approach. We implemented with Hypertext Preprocessor (PHP) programming language using Ant Bee Colony Algorithm and MySQL Relational Database Management System as backend. We compared the existing system and proposed system of the Ant Bee Colony. The results obtained show that the performance accuracy is more efficient.**

***Keywords*— Data leakage Detector, Malicious Data Leakage, Inadvertent Data Leakage, Social Media, ABC Algorithm, RAD**

## I. Introduction

Improvement of data security by different encryption algorithms, there are still occurrences of data leakages on social media. Another interesting reason why data leakage has been an issue is the difficulty in tracing the data leaker. This is because data leakers are hard to trace and often results to several issues in any working environment. Secondly, most computer users click on pop-up adverts when browsing the web, and are then mandated to fill out forms in order to access certain site locations. Hence, confidential data is leaked to un-trusted third parties without proper investigation of the third parties [1].

According to [2], sensitive data must be shared to third-parties that are trusted. For instance, the records of patients in a hospital might be given to researchers for the purpose of preparing new treatments, or a company that shares the information of customers to its partners. Many past researchers considered applications where the original sensitive data cannot be perturbed. Perturbation is a very useful technique where the data is modified and made less sensitive. Andrienn [3] described the importance of data leak prevention. According to him; data leak prevention helps in the protection of confidential data such as finance data, customer data, employee data, etc. In addition, once most critical data and its location are identified on the network, it can be monitored to determine who is accessing and using it. Andrienn[3] also cited the vulnerability of most data leakage decision systems which is as a result of insider negligence. On the other hand, as data continuously changes, discovery and classification across large organizations adds complexities. Also, most organization’s data is unstructured such as text documents, emails, blogs, videos, spreadsheets, web pages, etc. This further implies that there is an urgent need for an Improved Data Leak Detector that will aid in the notification and prevention of potential data losses. Furthermore, the effectiveness and efficiency of an Improved Data Leak Detector can be deployed to the concept of Situation Awareness. This study addresses the deployment of Data Leak Detection and prevention especially its deployment and contributions to Situation Awareness. Confidential Data on social media can be leaked as a result of lack of awareness on malicious data leakage which occurs as a result of the application of malicious and stealthy software by hackers to steal sensitive or organizational data from a host. Secondly, another data leakage problem is the lack of awareness on inadvertent data leakage. This leakage problem occurs when there is an accidental leakage of sensitive data in the outbound traffic by a legitimate user. For instance, when the user forgets to use data encryption or carelessly forwards an internal email and attachment to outsiders. In addition, most hackers capitalize on the vulnerability of social media users especially in the area of data security.

## II. Related works

The uncontrolled transfer of confidential information to outsiders can be defined as data leakage. It poses a serious problem to numerous organizations as there is an increase in the occurrence of the incident. Despite the introduction of several methods of preventing data leakages, the problem still faces web and social media users. Secondly, data leakage detection systems perform the role of monitoring different interactive platforms on the web and further collect information about web documents according to the user’s preferences.

Dominich et al. [4] studied the modules of a data leakage detection system. According to them, the protected and confidential user information can be found in the document collection section. Thus, in order to protect these documents, the need for data encryption is highly indispensable.The usage of the vector space mathematical model by the system module indicates its importance. Tuza [5] further analyzed the vector space model in the system to be used as the basis of cryptographic and text mining module. In other words, the scoring module can match the vector space based mathematical representation of the web and confidential user documents.

The descriptive situation awareness model was presented by [6] in a generic dynamicdecision-making environment, studying the relevant factors and underlying mechanisms, which explains the interaction of individuals and environmental factors. Numerous approaches have been adopted for the implementation of the mentioned situation awareness models. The blackboard systems have been used to model all levels of situation awareness as defined by [6].

According to [7] most data leakers apply a spear-phishing method that targeted on key employees of victim organizations through social media in order to carry out reconnaissance and theft of confidential proprietary information. Secondly, there should be an efficient data theft detector for social media users. This detector should create awareness on irresponsible use of social media which often result to threats of data leakages.

Also, the role of the leak detector encompasses the monitoring of the network traffic for potential data leaks. The inspection can as well be carried out offline so as not to cause any real-time delay in the routing packets. Xiaokui et al. [8] proposed two important implementation models for the deployment of data leakage detector as a service on social media.

Panagiotis et al. [9] discussed about the agent guilt model for data leakage detection during the course of doing business with sensitive data online. According to them; the use of watermarking which involves the embedding of a unique code on each of the distributed copy of datasets, is used to traditionally handle data leakage detection. Thus, if that copy is later discovered in the hands of an authorized party, the leaker can be identified.

Chirag [10] developed the Detection and Prevention of Data Leakages on web servers. The work discussed data leakages in most web servers as an uncontrolled or unauthorized transmission of classified information to the outside. It poses a serious problem to companies as the cost of incidents continues to increase. Many software solutions were developed to provide data protection. However, data leakage detection systems cannot provide absolute protection. Thus, it is essential to discover data leakage as soon as possible.

Atif [11] researched on Information Leakage through online social networking; opening the doorway for Advanced Persistence and Threats. He illustrated that Data leakage is the big challenge in front of the industries and different institutes. Though there are number of systems designed for the data security by using different encryption algorithms, there is a big issue of the integrity of the users of those systems. It is very hard for any systemadministrator to trace out the data leaker among the system users.

Endsley [6] also analyzed algorithms for distributing objects to agents in a way that improves the possibilities of identifying a potential leaker.

## III. Materials and methods

### *A. Requirements and Data Source for the Proposed System Design.*

The development of the proposed system is a function of numerous requirements and data source which include:

- Personal Computer with specifications which include: 2GBRAM, 64 bits processing speed and Windows 7 Operating System.
- Journals on Data Leakage
- Journals on Social Media Security
- Journals on Relational Database Management Systems
- Journals on Xampp Servers
- Journals on Apache Servers
- Journals on Cyber Security
- Questionnaires on the occurrenceData Leakage on Social Media and adopted measures in curbing the menace.
- Journals on Rapid Application Design Methodology
- Journals on Artificial Bee Colony Algorithm
- Journals on Situation Awareness
- Related Works on Data Detection Systems
- Information Assistance from Research Supervisor

There are several existing algorithms on the Proposed System development. However, we chose to adopt Ant bee Colony Algorithm (ABC) for the Proposed System Design. This nature-inspired algorithm is motivated by a variety of biological and natural processes. Their popularity is based primarily on the ability of biological systems to efficiently adapt to frequently changeable environments. Evolutionary computation, neural networks, ant colony optimization, particle swarm optimization, artificial immune systems, and bacteria foraging algorithm are among the algorithms and concepts that were motivated by nature. Swarm behavior is one of the main features of different colonies of social insects (bees, wasps, ants, termites). This type of behavior is principally characterized by autonomy, distributed, functioning, and self-organizing. Swarm Intelligence is the area of Artificial Intelligence based on studying actions of individuals in various decentralized systems. When creating Swarm Intelligence models and techniques, researchers apply some principles of the natural swarm intelligence. According to [10], the ABC algorithm adopts an Artificial Intelligence approach in analyzing and validating Information Systems.

### *B. Comparative Analysis between the Proposed and Existing Algorithms.*

There are several existing algorithms on the Proposed System development. We decided to analyze and compare the algorithms. The analysis process was achieved through the deployment of online survey questionnaires to Big Data Analysts in order to get the performance feedback on the surveyed algorithms. During numerous applications of the algorithm, it was observed that the constructive version cannot successfully solve some combinatorial optimization problems. Therefore, several modifications of the algorithm were developed in order to solve hard optimization problems. The idea of improving alternatives could be developed in many different ways, and this approach certainly may be very useful for solving difficult combinatorial optimization problems. It has already been explored in the recent literature for solving some other hard optimization problems like network design and satisfactory problem in the logic with approximate conditional probability for the berth allocation algorithm.

*1) Ant Bee Colony Algorithm as Developed by [10]*

```
Step 1: Start
Step 2: Declare Variables
Step 3: Variables: SBEES, OBEES, QBEE, RP DEC,
        ATT RPT. Where SBEES is for Scout Bees,
        OBEES is for Onlooker Bees, and QBEE is for
        the Queen Bee, DEC is for Decision, RPT
        stands for Resource Pot Threat and ATT stands
        for attack.
Step 4: Initialize Variables
Step 5: `Initialize QBEE
        QBEE = SBEES + OBEES = (Problem
        Identification and Task Scheduling)
Step 7: Re-Initialize QBEE
Step 8: QBEE = (SBEES + OBEES) / RPT
Step 9: Re-Initialize QBEE
Step 10:QBEE = DEC
Step 11: DEC = (SBEES + OBEES) (ATT*RPT)
Step 12: Display DEC Result
Step 13: Stop
```

*2) Improved (Modified) Ant Bee Colony Algorithm (Proposed System)*

```
Step 1: Start
Step 2: Declare Variables
Step 3: Variables: SBEES, OBEES, QBEE, RP DEC,
        ATT RPT, FBU, FBUV, GMU, GMUV, TWU,
        TWUV, UN, PW. Where SBEES is for Scout
        Data Detectors, OBEES is for Monitoring Data
        Detectors, and QBEE is for the Central Data
        Detection System, DEC is for Decision, RPT
        stands for Confidential Data Threat, FBU
        stands for Facebook User, GMU stands for
        Gmail User, TWU stands for Twitter User,
        FBUV stands for Facebook User Validation,
        GMUV stands for Gmail User Validation,
        TWUV stands for Twitter User Validation, UN
        stands for Username, PW stands for Password
        and ATT stands for attack potential hacker.
Step 4: Initialize Variables
Step 5: Initialize FBUV
Step 6: FBUV = FBU * (UN + PW)
Step 7: Initialize GMUV
Step 8: GMUV = GMU * (UN + PW)
Step 9: Initialize TWUV
Step 10: TWUV = TWU * (UN + PW)
Step11: QBEE = SBEES + OBEES=
        (ProblemIdentification and Task Scheduling)
Step 12: Re-Initialize QBEE
Step 13: QBEE = (SBEES + OBEES) / RPT
Step 14: Re-Initialize QBEE
Step 15:  QBEE = DEC
Step 16:  DEC = (SBEES + OBEES) / (ATT*RPT)
Step 17: Display DEC Result
Step 18: Stop
Step 19: Quit System.
```

## IV. RESULT AND DISCUSSIONS

### *A. Algorithm / Program Performance Assessment Assessment Variables:*

E signifies Excellent, which is = 5.0

G signifies Good = 4.0

F signifies Fair = 3.0

P signifies Poor = 2.0

VP signifies Very Poor = 1.0

TABLE I

CHIRAG (2018) ALGORITHM ON THE DETECTION AND PREVENTION OF DATA LEAKAGES ON WEB SERVERS

| SN | PERFORMANCE AREA | VARIABLE | PERFORMANCE POINT |
|---|---|---|---|
| 1 | Datasets Used = Varchar/Numeric | E | 5.0 |
| 2 | System's Database Storage Capacity = 40GB | G | 4.0 |
| 3. | Processing Speed of the System = 5 Kilobytes per second | VP | 1.0 |
| 4. | Quick Response time of the system = 15 minutes | F | 3.0 |
| 5. | Open-Source Ability = None (Offline) | VP | 1.0 |
| 6. | GUI friendliness | E | 5.0 |
| TOTAL PERFORMANCE POINTS | | | 19.0 |

TABLE II

ATIF (2016) ALGORITHM ON INFORMATION LEAKAGE THROUGH ONLINE SOCIAL NETWORKING; OPENING THEDOORWAY FOR ADVANCED PERSISTENCE AND THREATS

| SN | PERFORMANCE AREA | VARIABLE | PERFORMANCE POINT |
|---|---|---|---|
| 1 | Datasets Used = OLE Objects | E | 5.0 |
| 2 | System's Database Storage Capacity = 18.5GB | G | 4.0 |
| 3. | Processing Speed of the System = 1.7 Kilobytes per second | VP | 1.0 |
| 4. | Quick Response time of the system = 5 minutes | G | 4.0 |
| 5. | Open-Source Ability = Enabled (Online) | G | 4.0 |
| 6. | GUI friendliness | F | 3.0 |
| TOTAL PERFORMANCE POINT | | | 21.0 |

TABLE III

BLESSING (2019), AN IMPROVED DATA LEAKAGE DETECTOR (IDLD) FOR SITUATION AWARENESS USING IMPROVED ANT BEE COLONY ALGORITHM (ABC)

| SN | PERFORMANCE AREA | VARIABLE | PERFORMANCE POINT |
|---|---|---|---|
| 1 | Datasets Used = OLE Objects | E | 5.0 |
| 2 | System's Database Storage Capacity = 100 GB (MySQL-Based) | E | 5.0 |
| 3. | Processing Speed of the System = 100 Kilobytes per second | E | 5.0 |
| 4. | Quick Response time of the system = 45 seconds | E | 5.0 |
| 5. | Open-Source Ability = Enabled (Online) | G | 4.0 |
| 6. | GUI friendliness | E | 5.0 |
| TOTAL PERFORMANCE POINT | | | 29.0 |

TABLE IV

SUMMARY OF ALGORITHM PERFORMANCE

| ALGORITHM | RESEARCH AREA AND DEPLOYMENT | GENERAL PERFORMANCE POINT |
|---|---|---|
| CHIRAG (2017) | The detection and prevention of Data Leakages on Web Servers | 19 |
| ATIF (2016) | Information Leakage through Online Social Networking; Opening the doorway for Advanced Persistence and Threats | 21 |
| BLESSING (2019) | An Improved Data Leakage Detector (IDLD) for Situation Awareness using Ant Bee Colony Algorithm (ABC) | 29 |

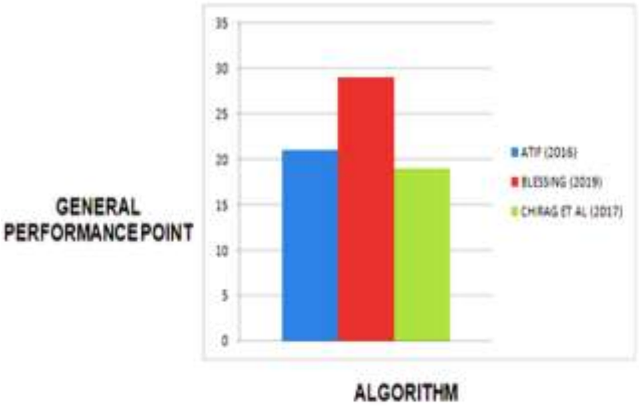

Fig 1. Algorithm Performance Chart

Figure 1 displays the general performance point of the three algorithms, which shows that the proposed system (Blessing 2019) has a better performance accuracy than others.

## V. CONCLUSIONS

The uncontrolled transfer of confidential information to outsiders can be defined as data leakage. It poses a serious problem to numerous organizations as there is an increase in the occurrence of the incident. Despite the introduction of several methods of preventing data leakages, the problem still faces web and social media users. Several modifications ofthe algorithm were developed in order to solve hard optimization problems. The idea of improving alternatives could be developed in many different ways, and the improved Ant Bee Colony approach certainly may be very useful for solving difficult combinatorial optimization problems such as network design.Secondly, data leakage detection systems perform the role of monitoring different interactive platforms on the web and further collect information about web documents according to the user's preferences. The collected web data sources are compared with the user's confidential documents.

## REFERENCES

[1] Sandip I. (2016), Data Allocation Strategies in Data Leakage and Detection, International Journal of Computer Applications (IJCA), 2(2), 1448 – 1452

[2] Prerna O. (2013), Review on Data Leakage Detection, International Journal of Engineering Research and Applications (IJERA), 1(3), 1088 – 1091, ISSN: 2248 – 9622

[3] Andrienn S, (2015). A Model for Data Leakage Detection, an International Article submitted to the Department of Computer Science, Stanford University, 353 Serra Street, Stanford, CA, 94305, USA

[4] Dominich P. (2013), Review on Data Leakage Detection, International Journal of Engineering Research and Applications (IJERA), 1(3)1088– 1091, ISSN: 2248 – 9622

[5] Tuza B. (2016), Data Allocation Strategies in Data Leakage and Detection, International Journal of Computer Applications (IJCA), 2(2), 1448 – 1452

[6] Endsley P. (2017), Design and Evaluation for Situation Awareness Enhancement, in proceedings of the Human Factors Society, 31st Annual Meeting, 1388 – 1392

[7] Nurul J. (2010), Pilot Situation Awareness: The challenge for the training community, in proceedings Of the Inter-service/Industry Training Systems Conference, 111 – 117

[8] Xiaokui U. (2014), Detecting Data Leakage using Data Allocation Strategies with fake objects, International Journal of Advance Research in Computer Engineering and Technology (IJARCET), 3(11), 3855 – 3862

[9] Panagiotis P. (2015), A model for Data Leakage Detection, an International Article submitted to the Department of Computer Science, Stanford University, 353 Serra Street, Stanford, CA, 94305, USA

[10] Chirag R, (2018). Data Leakage and Detection, an International Online Article on the Detection and Prevention of Data Leakages on web servers, International Journal of Computer Science and Information Technologies, 5(2), 2556 – 2558 (IJCSIT)

[11] Atif A. (2016), Information Leakage through online social networking; opening the doorway for Advanced Persistence, Threats, Proceedings of the 8th Australian Information Security Management Conference,http://ro.ecu.edu.au/ism/93