Journal of Physical Science and Innovation
ISSN: 2277-0119
Volume 9, No. 3, 2017

PRIVACY PRESERVING PUBLISHING SCHEME USING SIMULATED ATTACKER BACKGROUND KNOWLEDGE.

Agbator Oseremen Lawrence
Department of Computer Science,
Edo State Institute of Technology and Management, Usen, Edo
Email: mail4agbator@gmail.com

ABSTRACT

The enormous challenge of today's publishing privacy concerns in ubiquitous networks and computing include the ability to deploy large scale applications anywhere anytime. Cloud computing has revolutionized the way computing and software services are delivered to the clients on demand. It offers users the ability to connect to computing resources and access IT managed services with a previously unknown level of ease. Due to this greater level of flexibility, the cloud has become the breeding ground of a new generation of products and services. However, in this work we want direct attention away from the dataset which helps to guarantee the utility, and focus attention on the evasive attacker, simulate his knowledge and in other to create efficient security of datasets by developing a potential attacker background knowledge metric that will help in determining the extent to which anonymity algorithm should be applied to guarantee sufficient privacy for a particular domain without unnecessarily distorting the data utility with optimal anonymity.

Keywords: Attacker, Simulation, Anonimisation, Background-Knowledge, Metric.

INTRODUCTION

The enormous challenge of today's publishing privacy concerns in ubiquitous networks and computing include the ability to deploy large scale applications anywhere anytime. Today's applications are largely based on the seamless integration of lightweight mobile devices and cloud storage and computing

resources. Data Mining applications can now be easily deployed in mobile devices, data captured by smartphones can be stored and processed into a Cloud based expert system for intelligent analysis. The new challenges inherent in the ability to extract useful information from a vast amount of data which are intrinsically distributed is privacy of the entities identified by these vast amount of dataset. Research on Distributed Data Mining (DDM) has focused on the formulation of data mining algorithms for distributed computing environments, where each node processes its local data and contributes to compute a global solution. In many applications the solution is required to be available at every node. This is particularly important when considering applications in networked systems where each node is autonomous and active, like in peer-to-peer systems, mobile ad hoc networks, vehicular ad hoc networks, mobile social networks, wireless sensor networks. The coupling of data, software, standards and other technology from the traditional computing to mining of cloud and mobile data will indeed present new challenges. One of such challenge is privacy violation. We in this research work intend to focus on optimized application anonymity and integration algorithm that will guarantee the privacy of the entities whose attributes in the database management and knowledge based systems are been ported to a global cloud accessed by all kinds of mobile devices with their different kinds of operating systems, mobile applications and varied intentions.

PROBLEM STATEMENT

The explosion in the research on privacy preserving data publishing obviously has not been without its drawback, in one area it has been the inability of anonymization algorithm to efficiently guarantee the privacy of the published dataset and in another area it is the trade-off metric value which seeks to strike a balance between maximal data information and minimal privacy disclosure, this consideration sometimes advocate the reduction in the application of anonymity algorithm in other to prevent huge loss of information between the original information and the anonymised version. In

this research work we intend to solve this problem by looking at a different direction and that is the direction of the attacker background knowledge. Attacker background knowledge causes privacy breach in varying degree. This implies that data in different domain, based on attacker background knowledge is faced with different degree of vulnerability. It will be a part of this research to find out the way of simulating background knowledge of attackers, so that we can through the result of this research provide all-round protection for privacy.

OBJECTIVES OF THE STUDY

The main fundamental objectives of this study are;

1. To determine the varying degree of vulnerability that exists in dataset from different domain.
2. To reduce or eliminate the problem of trade off metric that exist between the level of anonymisation to be implemented and the distortion that result from loss of information between original dataset and anonymised dataset.
3. To eliminate or reduce the probabilistic nature of anonymisation schemes as a result of the assumption that a potential attacker have limited background knowledge.
4. To develop privacy publishing scheme that will provide all-round privacy by leveraging on attackers background knowledge.
5. To open a window of research tended towards the development of domain specific anonymity techniques.

LITERATURE REVIEW

Data mining is increasingly vital for decision makers to make a timely and accurate response from huge amounts of easily accessible information in the changing global environment. Accessing this published data for different analysis has presented its own challenges about privacy violations. In this literature review we are going to look at the research works that have

discussed methods and techniques of privacy preserving data publishing which are regarded as strong guarantee to avoid information disclosure and protect individuals' privacy. According to Yang Xu et al 2013, recent work focuses on proposing different anonymity algorithms for varying data publishing scenarios to satisfy privacy requirements, and keep data utility at the same time. K-anonymity has been proposed for privacy preserving data publishing, which can prevent linkage attacks by means of anonymity operation, such as generalization and suppression. Numerous anonymity algorithms have been utilized for achieving k-anonymity. In the paper published by Yang Xu, Tinghuai Ma, Meili Tang and Wei Tian in 2013. A Survey of Privacy Preserving Data Publishing using Generalization and Suppression, an overview of the development of privacy preserving data publishing, which is restricted to the scope of anonymity algorithms using generalization and suppression, was done.

The paper introduced privacy preserving models for attack. Thereafter an overview of several anonymity operations was done. They went further to cover anonymity algorithms and information metric which is essential ingredient of the algorithms. Many literatures reviewed here like Data mining in cloud Computing by Xia Geng and Zhi Yang in ISCA, 2013 conference introduces the basic concept of cloud computing and data mining, with emphasis on how data mining is used in the cloud and summarises the research on parallel programming model and mass data mining services based on cloud computing.Efficiently Answering Top-k Typicality Queries on Large Databases by Ming Hua, Jian Pei, Ada W. C. Fu, Xuemin Lin and Ho-Fung Leung 2007 apply the idea of typicality analysis from psychology and cognition science to database query answering, and study the novel problem of answering top-k typicality queries. The work model typicality in large data sets systematically.

In the work done by Jie Wang Weijun Zhong, Shuting Xu and Jun Zhang in 2005 titled Selective Data Distortion via Structural Partition and SSVD for Privacy Preservation, they posited that accurate information extracted from datasets is required for making reasonable decisions using data mining algorithms, and that privacy preservation has become one of the top priorities in the design of various data mining applications. In the paper, a novel data distortion strategy based on structural partition and sparsified Singular Value Decomposition (SSVD) technique was proposed. Three schemes, object-based partition, feature-based partition and hybrid partition, were defined to permit a tradeoff between privacy protection on centralized datasets and accuracy of data mining techniques. They also used some metrics to measure privacy preservation to examine the performance of their proposed new strategies. Data utility of the three proposed schemes was examined by a binary classification based on the support vector machine. Furthermore, the effect of different ranks of SVD and the threshold value of SSVD on data distortion and utility were also tested by the researches.

Their findings showed in comparison with standard data distortion techniques, the proposed schemes were very efficient in achieving a good tradeoff between data privacy and data utility, and it affords a feasible solution, with a significant reduction on the computational cost from SVD, to protect sensitive information and promise high accuracy in decision making. The work done by Xuyun Zhang, Wanchun Dou and Jian Pei in 2013, is one of the research areas that has done something also on privacy issues in cloud data mining, in their work they did investigate the the local-recoding problem for big data anonymization against proximity privacy breaches and attempt to identify a scalable solution to this problem associated with a practical and widely-adopted technique for data privacy preservation which is to anonymize data via generalization to satisfy a given privacy model, This is because privacy preserving approaches tailored to small-scale data sets often fall short when encountering big data, due to their insufficiency or poor scalability.

Specifically, they presented a proximity privacy model allowing semantic proximity of sensitive values and multiple sensitive attributes, and model the problem of local recoding as a proximity-aware clustering problem. A scalable two-phase clustering approach consisting of t-ancestors clustering (similar to k-means) algorithm and a proximity-aware agglomerative clustering algorithm were later proposed by the researchers to address the problem associated with data anonimization in traditional computing. The rigorous definition of privacy protection by Dalenius as referenced by Yang Xu et al is that addressing to the published dataset should not increase any possibility of adversary to gain extra information about individuals, even with the presence of background knowledge. Many literatures in this subject area have always submitted that it is not possible to quantize the scope of background knowledge of a potential attacker and therefore, a transparent hypothesis taken by many PPDP literatures according to Yang Xu et al is that adversary has limited background knowledge. However, in this work we want to develop a potential attacker background knowledge metric that will help in determining the extent to which anonymity algorithm should be applied to guarantee sufficient privacy for a particular domain without unnecessarily distorting the data utility with optimal anonymity.

Methodology. The approach we intend to adopt will flow from our ability to firmly establish the presence of significant potential background knowledge as against the present assumption held by privacy preserving data publishing literatures, the transparent hypothesis that an adversary has limited background knowledge. Generalization and suppression are the most common anonymity operations used to implement k-anonymity and its extension. Generalization in a nutshell means replacing specific value of quasi-identifiers with more general value. Suppression, which is the highest form of generalization is the operation which uses special symbolic character to replace its authentic value (e.g. *, &, #), and makes the value meaningless. Unlike generalization and suppression, anatomization and permutation does

not make any modification of original dataset, while decrease the correlation of quasi-identifiers and sensitive attribute. Generally, quasi-identifiers and sensitive attribute are published separately. Quite a few researches make use of these two anonymity operations. When just referring to the purpose of information statistic, perturbation operation has merits of simplicity and efficiency. The main idea of perturbation is to substitute original value for synthetic data, and, ensures the statistical characteristic of original dataset. After perturbation operation, the dataset is completely not the presentation of original dataset which is its remarkable trait. Adding noise, swapping data and generating synthetic data are the three common means of perturbation. (Yang Xu et al, 2013).

All these different anonymisation methods and operations are aimed at preventing a potential attacker from using his background knowledge to perform two major types of attacks, which is linkage attack and probabilistic attack. This background knowledge of a potential attacker is derivable from his access to previously exposed dataset or rather publicly published dataset. The popularity and wide adoption and use of cloud infrastructure in publishing, naturally offers one medium for global interaction. Therefore leveraging on the creation of logs by systems which are not directly under the supervision of the user, we plan to simulate dataset level of exposure which statistically correlate with potential attacker background knowledge of dataset in a specific domain.

CONCLUSION

Cloud computing has revolutionized the way computing and software services are delivered to the clients on demand. It offers users the ability to connect to computing resources and access IT managed services with a previously unknown level of ease. Due to this greater level of flexibility, the cloud has become the breeding ground of a new generation of products and services. However, the flexibility of cloud-based services comes with the risk of the

security and privacy of users' data. Thus, security concerns among users of the cloud have become a major barrier to the widespread growth of cloud computing especially in the area of privacy preserving publishing. One of the security concerns of cloud is data mining based privacy attacks that involve analyzing data over a long period to extract valuable information. In particular, in current cloud architecture a client entrusts a single cloud provider with his data. It gives the provider and outside attackers having unauthorized access to cloud, an opportunity of analyzing client data over a long period to extract sensitive information that causes privacy violation of clients. This is a big concern for many clients of cloud computing, it is our hope that this research when successfully completed will eliminate or bring to a significant reduction, the barrier to the growth of cloud computing as a result of the security concerns among users of the cloud and by extension cloud data mining.

REFERENCES

Flavia Moser, Recep Colak, Arash Rafiey, Martin Ester (2009) Mining Cohesive Patterns from Graphs with Feature Vectors Simon Fraser University, Canada.

Jie Wang Weijun Zhong, Shuting Xu and Jun Zhang (2005) Selective Data Distortion via Structural Partition and SSVD for Privacy Preservation,

Ming Hua, Jian Pei, Ada W. C. Fu, Xuemin Lin and Ho-Fung Leung (2007) Efficiently Answering Top-k Typicality Querieson Large Databases

Xia Geng and Zhi Yang (2013) Data mining in cloud Computing in ISCA, 2013 conference.

Xuyun Zhang, Wanchun Dou, Jian Pei(2013) Proximity-Aware Local-Recoding Anonymization with MapReduce for Scalable Big Data Privacy Preservation in Cloud IEEE TRANSACTIONS ON COMPUTERS, TC-2013-12-0869 1

Yang Xu, Tinghuai Ma, Meili Tang and Wei Tia (2013), A Survey of Privacy Preserving Data Publishing using Generalization and Suppression. Appl. Math. Inf. Sci. **8**, No.3, 1103-1116 (2014) / www.naturalspublishing.com/Journals.asp

Reference to this paper should be made as follows: Agbator Oseremen Lawrence (2017), Privacy Preserving Publishing Scheme Using Simulated Attacker Background Knowledge. J. of Physical Science and Innovation, Vol. 9, No. 3, Pp. 64-72