Data Utility Verification Using Private Publishing Schemes

Dr.Vandana S. Bhat[1], PawanKumar Mashimode[2]
[1, 2] Dept of ISE
[1, 2] SDM college of Engineering &Technology, Dharwad
Karnataka, India

Abstract- *Service providers have the ability to collect large amounts of user data. Sometimes, a set of providers may try to aggregate their data and then anonymizing it before publishing for specific data mining tasks. In this process, how to protect user's privacy is extremely critical. This is the so-called privacy-preserving collaborative data publishing problem. In such scenarios, the data users may have a strong demand to measure the utility of the published data, since most anonymization techniques have side effects on data utility. This task is non-trivial, because the utility measuring usually requires the aggregated raw data, which is not revealed to the data users due to privacy concerns. Furthermore, the data publishers may even cheat in the raw data, since no one, including the individual providers, knows the full data set. We consider the collaborative data publishing problem for anonymizing horizontally partitioned data at multiple data providers. We consider a new type of "insider attack" by colluding data providers who may use their own data records (a subset of the overall data) to infer the data records contributed by other data providers. We propose a privacy-preserving utility verification mechanism based upon cryptographic technique for DiffPart—a differentially private scheme designed for set-valued data. This proposed system can measure the data utility based upon the encrypted frequencies of the aggregated raw data instead of the plain values, which thus prevents privacy breach. Moreover, it is enabled to privately check the correctness of the encrypted frequencies provided by the publisher, which helps detect dishonest publishers. We also extend this mechanism to DiffGen—another differentially private publishing scheme designed for relational data.*

Keywords- Collaborative, data publishing, utility verification, differential privacy.

I. INTRODUCTION

Now days lots of data is stored in cloud hence to avoid storage memory, in cloud it is easy and highly fast to store, retrieve and process the large amount of data, so it is important to maintain the privacy of the data and integrity of the data. In this work we have take the two schemes to secure the privacy of the data and secure the integrity of the data such as Diffpart and DiffGen and RSA algorithm is also used for encryption and decryption of the data ,so that data integrity should be maintain in the cloud .Due to fast headway on storing, processing, also systems administration abilities for registering devices, there need been a colossal Growth in the accumulation about advanced majority of the data over people. And the rise from claiming new registering paradigms, for example, cloud computing, expands the plausibility about vast scale conveyed information gathering starting with numerous sources. Same time the gathered information offer colossal chances for mining suitable information, there may be likewise An danger to security a result information done crude type frequently hold delicate data over people. Administration suppliers have the capacity to gather a lot from claiming client information. Sometimes, An situated for suppliers might attempt on aggravator their information for particular information mining assignments. For example, the doctor's facilities nation-wide might outsource their restorative records should an exploration one assembly for mining the spreading examples from claiming flu epidemics. In this process, how to ensure users' security may be greatly discriminating. This may be those alleged privacy-preserving community oriented information distributed issue.

A utility-preserving system to differentially private information discharges is introduced. In for k-anonymity, it has the ability to prepare universally useful secured datasets. Information may be transformed through individual positioning micro amassed to decrease its affectability. Subtle elements ahead how will apply those system to numerical What's more unmitigated information need aid given. The information must make anonymized When discharge should protect those protection of the subjects should whom those records relate. Differential protection will be An protection model to anonymization that offers that's only the tip of the iceberg hearty security certifications over past models, for example, such that k-anonymity Also delay. anyhow, frequently dismissed those utility of transmission individual outputs will be very limited, Possibly due to measure about commotion that needs with make included with acquire them alternately in light utility is just safeguarded for An confined kind or An Constrained amount about query.on the information discharges aggravate no presumptions on the employments of the ensured information.

Differential protection will be an much that's only the tip of the iceberg thorough protection model. It obliges that those discharged information is uncaring of the expansion alternately evacuation of a solitary record. To actualize all the this model, those comparing anonymization instruments typically must include commotion of the distributed data, alternately probabilistically sum up the crude information. Obviously, at these information anonymization components need not kidding side impacts on the information utility. Likewise An result, those clients of the distributed information normally have An solid interest on confirm the genuine utility of the anonymized information. On change a crude information table on fulfill a specified protection requirement, a standout amongst those The majority prevalent systems will be generalization. Generalization replaces a particular esteem with An that's only the tip of the iceberg general esteem to make those data lesquerella exact same time preserving those “truthfulness” from claiming data.

II. LITERATURE SURVEY

In this chapter various papers have been surveyed to study the existing system and approaches. In this paper “Secure distributed framework for achieving _-differential privacy”, [D. Alhadidi, N. Mohammed, B. C. M. Fung, and M. Debbabi]. Address the problem of private data publishing where data is horizontally divided among two parties over the same set of attributes. In particular, they present the first generalization-based algorithm for differentially private data release for horizontally-partitioned data between two parties in the semi honest adversary model. The generalization algorithm correctly releases differentially-private data and protects the privacy of each party according to the definition of secure multi-party computation. To achieve this, they first present a two-party protocol for the exponential mechanism. This protocol can be used as a sub protocol by any other algorithm that requires exponential mechanism in a distributed setting.

Disadvantages:

- Need different heuristics for different data mining tasks.
- Need to increase robustness.

In this paper “Probabilistic relational reasoning for differential privacy ”,[G. Barthe, B. Köpf, F. Olmedo, and S. Z. Béguelin] states differential privacy is a notion of confidentiality that protects the privacy of individuals while allowing useful computations on their private data. Deriving differential privacy guarantees for real programs is a difficult and error-prone task that calls for principled approaches and tool support. Approaches based on linear types and static analysis has recently emerged; however, an increasing number of programs achieve privacy using techniques that cannot be analyzed by these approaches and Presents CertiPriv, it is a machine-checked framework that supports fine-grained reasoning about an expressive class of privacy policies in the Coq proof assistant. In contrast to previous language-based approaches to differential privacy, CertiPriv allows to reason directly about probabilistic computations and to build proofs from first principles. As a result, CertiPriv achieves flexibility, expressiveness, and reliability, and appears as a plausible starting point for capturing and analyzing formally new developments in the field of differential privacy.

Disadvantages:

- Need to use the game playing technique for verifying in CertiPriv
- Not scalable

In this “Differentially private trajectory data publication,” [R. Chen, B. C. M. Fung, and B. C. Desai], propose a non-interactive data-dependent sanitization algorithm to generate a differentially private release for trajectory data. The efficiency is achieved by constructing a noisy prefix tree, which adaptively guides the algorithm to circumvent certain output sub-domains based on the underlying database. And design a statistical process for efficiently constructing a noisy prefix tree under Laplace mechanism. This is vital to the scalability of processing datasets with large location universe sizes. We make use of two sets of inherent constraints of a prefix tree to conduct constrained inferences, which helps generate a more accurate release.

Disadvantages:

- Need to work on utility of sanitized data on other data mining tasks, for example, classification and clustering.

In this “Differentially private transit data publication: [R. Chen, B. C. M. Fung, B. C. Desai, and N. M. Sossou], A case study on the Montreal transportation system”, present solution to transmit data publication under the rigorous differential privacy model for the Societies transport the Montréal (STM). And propose an efficient data-dependent yet differentially private transit data sanitization approach based on a hybrid-granularity prefix tree structure. Moreover, as a post-processing step, make use of the inherent consistency constraints of a prefix tree to conduct constrained inferences, which lead to better utility. Proposed solution not only applies to general sequential data, but also can be seamlessly extended to trajectory data.

Disadvantage:

- Not robust
- Not scalable

In this “Publishing set-valued data via differential privacy”, [R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai, and L. Xiong], study the problem of publishing set-valued data for data mining tasks under the rigorous differential privacy model. And propose a probabilistic top-down partitioning algorithm for publishing set-valued data in the framework of differential privacy. Compared to the existing works on set-valued data publishing, our approach provides stronger privacy protection with guaranteed utility. Also contributes to the research of differential privacy by demonstrating that an efficient non-interactive solution could be achieved by carefully making use of the underlying dataset.

Disadvantages:

- Need to increase security and efficiency

In this paper “FAST: Differentially private real-time aggregate monitor with filtering and adaptive sampling”, [L. Fan, L. Xiong, and V. Sunderam], a tool for monitoring real-time aggregates under differential privacy with filtering and adaptive sampling. The key innovation is that FAST utilizes feedback loops based on observed (perturbed) values to dynamically adjust the filtering model as well as the sampling rate. Studies across multiple data sets confirm the effectiveness and the superior performance of FAST algorithms with respect to the state-of-the art methods. The real-time feature and accurate release provided by FAST will facilitate data holders to continuously share private aggregate, thus enabling important data monitoring applications, such as disease surveillance and traffic monitoring.

III. METHODOLOGY

We have 3 modules
1. Creator
2. Publisher
3. Reader

Module Description:

Module 1: Creator

In this module maker may be enrolled with the publisher, after that maker will login to publisher so as will transfer their information. Those makers hold the crude information & need to safely transfer their crude information should a focal publisher who may be guaranteed on never unveil this information on other gatherings including those suppliers. When uploading any information to publisher, maker will scramble Furthermore transfer to security end goal. Those information managers are answerable for scanning. Also uploading those records on Publisher

Module 2: Publisher

Publisher. The Publisher may be capable looking into sake of the record substance supplier for both allocating those suitable amount of assets in the cloud, furthermore reserving the period again which the required assets would allocate. It will dispense space, get information from holder Also arrange those data, What's more store clinched alongside numerous servers. Publisher server will saves every last one of information holder majority of the data and saves every last one of onlooker majority of the data Furthermore it additionally permits entry of the majority of the data through imp organize.

Module 3: Reader

In this module onlooker might download the record substance. In the recent past downloading client need with register In later client could login of the publisher What's more download the files. Spectator could additionally see those uploaded Files Furthermore they might entry those document. Commissioned client each person download those files. Following accepting that document they might confirm the utility of the information.

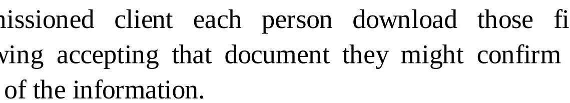
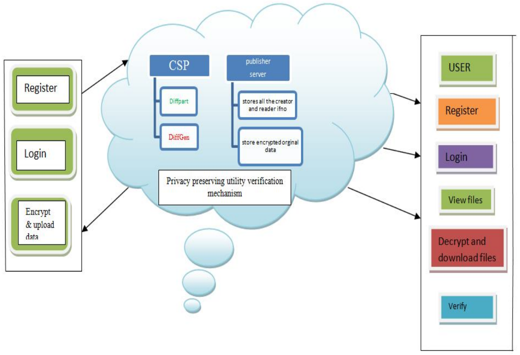
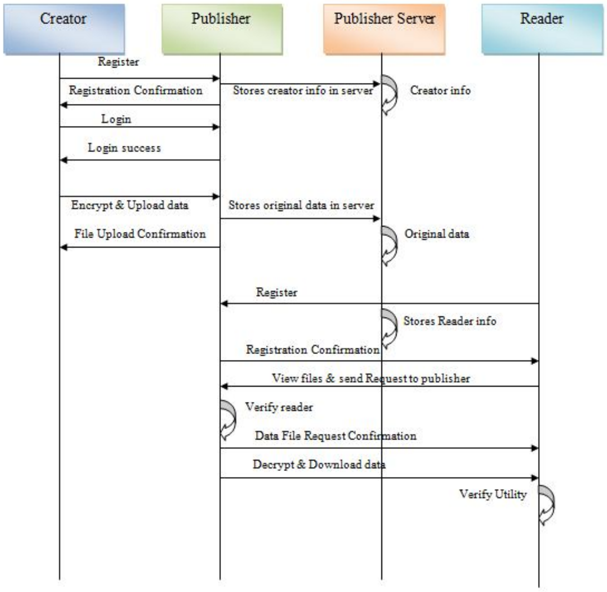


Fig: Overview of DiffPart or DiffGen.

Architecture


Register
Login
Encrypt & upload data
CSP
Diffpart
DiffGen
publisher server
stores all the creator and reader ifno
store encrypted orginal data
Privacy preserving utility verification mechanism
USER
Register
Login
View files
Decrypt and download files
Verify

Sequence Diagram



IV. RESULTS

If creator wants store data in cloud then he has to create an account in cloud by using user name and password. For creating an account you have register to cloud first. After creating an account in the cloud, the user should login by using user name and password. Hence the creator will have an account in cloud, here creator can browse Files from the location where you kept the files and Upload them, If reader wants to read the Uploaded files click on Read button and enter the file credential, If creator wants to update the Already uploaded files, update files in the text area and click on Update button, If creator wants to delete the files, click on Delete button. If any Reader wants to read any files, before accessing any files they should register first, After registration is successful, login by using username and password, Click on view publisher button to view all publisher files. If reader wants to access the files, they should send a request to creator, Then creator will view all requests from reader by click on view Reader Requests, Later if that requested file is available, creator will send a response to reader, Reader can view all the responses from creator by click on View Received Files button, Then Request for Publisher key by click on Request publisher key button---- Note down the key, Request creator key ----Note down the key, By using both publisher and creator key Request for file content by click on content button, If you have entered valid keys, then you will receive an file content, Finally you can save the files by click on SAVE button.

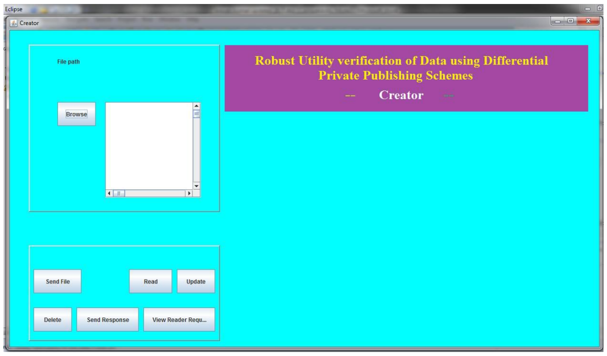


Fig: figure of creator account.

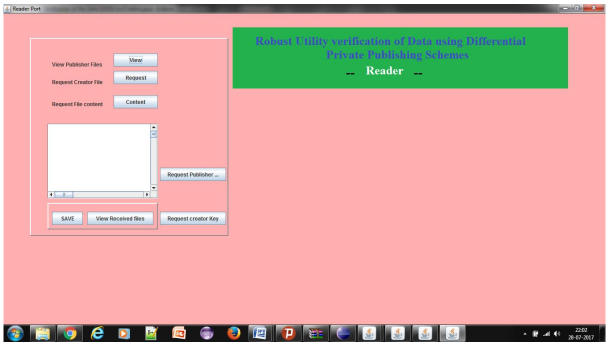


Figure of reader account in the cloud

V. CONCLUSIONS

Now days lots of data is stored in cloud hence to avoid storage memory, in cloud it is easyand highly fast to store, retrieve and process the large amount of data, so it is important to maintain the privacy of the data and integrity of the data. In this work we have take the two schemes to secure the privacy of the data and secure the integrity of the data such as Diffpart and DiffGen and RSA algorithm is also used for encryption and decryption of the data ,so that data integrity should be maintain in the cloud . We consider the issue of verifying the data utility and data secur.ity of the data stored in the cloud released by non interactive differentially private strategies. Comparable systems areproposed to accomplish the goals for set value and relational data respectively.

REFERENCES

[1] Jingyu Hua, An Tang, Yixin Fang, Zhenyu Shen, and Sheng Zhong,” Privacy-Preserving Utility Verification of the Data Published by Non-Interactive Differentially Private Mechanisms”, ieee transactions on information forensics and security, vol. 11, no. 10, october 2016.

[2] D. Alhadidi, N. Mohammed, B. C. M. Fung, and M. Debbabi, “Secure distributed framework for achieving _-differential privacy,” in Privacy Enhancing Technologies (Lecture Notes in Computer Science), vol. 7384. Berlin, Germany: Springer, 2012, pp. 120–139.

[3] G. Barthe, B. Köpf, F. Olmedo, and S. Z. Béguelin, “Probabilistic relational reasoning for differential privacy,” ACM SIGPLAN Notices, vol. 47, no. 1, pp. 97–110, Jan. 2012.

[4] D. Boneh, E.-J. Goh, and K. Nissim, “Evaluating 2-DNF formulas on ciphertexts,” in Theory of Cryptography. Berlin, Germany: Springer, 2005, pp. 325–341.

[5] R. Chen, B. C. M. Fung, and B. C. Desai. (2011). “Differentially private trajectory data publication.” [Online]. Available: http://arxiv.org/abs/1112.2020.

[6] R. Chen, B. C. M. Fung, B. C. Desai, and N. M. Sossou, “Differentially private transit data publication: A case study on the montreal transportation system,” in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), 2012, pp. 213–221.

[7] R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai, and L. Xiong, “Publishing set-valued data via differential privacy,” Proc. VLDB Endowment, vol. 4, no. 11, pp. 1087–1098, 2011.

[8] C. Dwork, “Differential privacy,” in Automata, Languages and Programming. Berlin, Germany: Springer, 2006, pp. 1–12.

[9] C. Dwork, “Differential privacy: A survey of results,” in Theory and Applications of Models of Computation. Berlin, Germany: Springer, 2008, pp. 1–19.

[10] C. Dwork, “Differential privacy in new settings,” in Proc. 21st Annu. ACM-SIAM Symp. Discrete Algorithms (SODA), 2010, pp. 174–183.