Robust Utility Verification of Data using Differential Private Publishing Schemes

Pawan Kumar Mashimode
M.Tech(IT),Dept of ISE,
SDM College of Engineering
&Technology, Dharwad,
Karnataka, India

Dr. Vandana S. Bhat
Dept of ISE,
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.**

General Terms:- Collaborative, data publishing, utility verification, differential privacy.

1. INTRODUCTION

Due to the rapid advancement in storing, processing, and networking capabilities of computing devices, there has been a tremendous growth in the collection of digital information about individuals. And the emergence of new computing paradigms, such as cloud computing, increases the possibility of large-scale distributed data collection from multiple sources. While the collected data offer tremendous opportunities for mining useful information, there is also a threat to privacy because data in raw form often contain sensitive information about individuals. Service providers have the ability to collect large amounts of user data. Sometimes, a set of providers may try to aggregate their data for specific data mining tasks. For example, the hospitals nation-wide may outsource their medical records to a research group for mining the spreading patterns of influenza epidemics. In this process, how to protect users' privacy is extremely critical. This is the so-called privacy-preserving collaborative data publishing problem.

An utility-preserving method for differentially private data releases is presented. Like with k-anonymity, it is able to produce general-purpose protected datasets. Data is processed via individual ranking micro aggregation to reduce its sensitivity. Details on how to apply the method to numerical and categorical data are provided. The data must be anonymized before release to protect the privacy of the subjects to whom the records relate. Differential privacy is a privacy model for anonymization that offers more robust privacy guarantees than previous models, such as k-anonymity and its extensions. However, it is often disregarded that the utility of differentially private outputs is quite limited, either because of the amount of noise that needs to be added to obtain them or because utility is only preserved for a restricted type and/or a limited number of queries. On the contrary, k-anonymity-like data releases make no assumptions on the uses of the protected data.
Differential privacy is a much more rigorous privacy model. It requires that the released data is insensitive to the addition or removal of a single record. To implement this model, the corresponding anonymization mechanisms usually have to add noise to the published data, or probabilistically generalize the raw data. Obviously, all these data anonymization mechanisms have serious side effects on the data utility. As a result, the users of the published data usually have a strong demand to verify the real utility of the anonymized data. To transform a raw data table to satisfy a specified privacy requirement, one of the most popular techniques is generalization. Generalization replaces a specific value with a more general value to make the information less precise while preserving the "truthfulness" of information.

2. LITERATURE SURVEY

"D. Alhadidi, N. Mohammed, B. C. M. Fung, and M. Debbabi"
The paper "Secure distributed framework for achieving _-differential privacy", address the problem of private data publishing where data is horizontally divided among two parties over the same set of attributes. In particular, we

present the first generalization-based algorithm for differentially private data release for horizontally-partitioned data between two parties in the semihonest 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, we first present a two-party protocol for the exponential mechanism. This protocol can be used as a subprotocol 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.

"G. Barthe, B. Köpf, F. Olmedo, and S. Z. Béguelin"
The paper "Probabilistic relational reasoning for differential privacy", 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 finegrained 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

"R. Chen, B. C. M. Fung, and B. C. Desai."
The paper "Differentially private trajectory data publication.", 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.

"R. Chen, B. C. M. Fung, B. C. Desai, and N. M. Sossou"
The paper "Differentially private transit data publication: A case study on the montreal transportation system", present solution to transit data publication under the rigorous differential privacy model for the Société de transport de 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.

Disadvantages:

- Not robust
- Not scalable

"R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai, and L. Xiong"
The paper "Publishing set-valued data via differential privacy", 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

"L. Fan, L. Xiong, and V. Sunderam"
The paper "FAST: Differentially private real-time aggregate monitor with filtering and adaptive sampling", proposed FAST, 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.

3. METHODOLOGY

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

Module Description:

Module 1: Creator
In this module Creator is registered with the publisher, after that creator will login to publisher in order to upload their data. The creator holds the raw data & wants to securely upload their raw data to a central publisher who is guaranteed to never disclose these data to other parties including the providers. Before uploading any data to publisher, creator will encrypt and upload for security purpose. The Data Owner is responsible for browsing and uploading the file to Publisher.

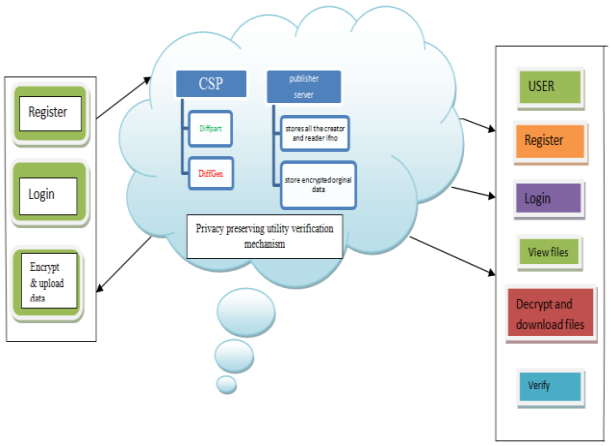
Module 2: Publisher
The Publisher is responsible on behalf of the file content provider for both allocating the appropriate amount of resources in the cloud, and reserving the time over which the required resources are allocated. It will allocate space, receive data from owner and classify the data, and store in multiple servers.

Publisher server will stores all the data owner information and stores all the reader information and it also allows access to the information through IP network.

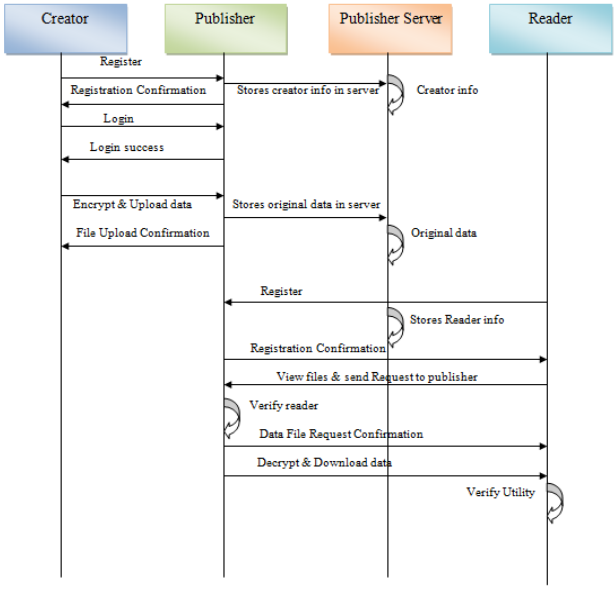
Module 3: Reader
In this module Reader can download the file content. Before downloading user has to register first, later user can login to the publisher and download the files. Reader can also view the uploaded Files and they can access the file. Authorized user every one download the files. After receiving the file they can verify the utility of the data.
In this work we first propose a privacy-preserving utility verification mechanism for DiffPart, consist of two main steps: recursive partition (i.e., specializations) and perturbation. In each round of partition, they split records into more specific equivalent groups and then count each of the groups in a noisy way. Those groups that are empty or extremely small are discarded for improving time efficiency. Specifically,the partition of *DiffPart* in each round is based on a contextfree
Taxonomy tree that ignores the underground dataset,and thus is deterministic. While *DiffGen* uses an exponential mechanism to select a candidate for specialization in each round based on the underground dataset, and thus the partition
Is probabilistic.

Fig: Overview of DiffPart or DiffGen.

Architecture



Sequence Diagram



4. ANALYSIS AND DISCUSSION

Collaborative data publishing can be considered as a multi-party computation problem, in which multiple providers wish to compute an anonymized view of their data without disclosing any private and sensitive information. We assume the data providers are semi-honest, commonly used in distributed computation setting. A trusted third party (TTP) or Secure Multi-Party Computation (SMC) protocols can be used to guarantee there is no disclosure of intermediate information during the anonymization. However, neither TTP nor SMC protects against inferring information using the anonymized data.

A lot of privacy models and corresponding anonymization mechanisms have been proposed, such as k-anonymity and differential privacy. K-anonymity and its variants protect privacy by generalizing the records such that they cannot be distinguished from some other records. Differential privacy is a much more rigorous privacy model. It requires that the released data is insensitive to the addition or removal of a single record. To implement this model, the corresponding anonymization mechanisms

usually have to add noise to the published data, or probabilistically generalize the raw data. Obviously, all these data anonymization mechanisms have serious side effects on the data utility. As a result, the users of the published data usually have a strong demand to verify the real utility of the anonymized data.

Due to the rapid advancement in storing, processing, and networking capabilities of computing devices, there has been a tremendous growth in the collection of digital information about individuals. And the emergence of new computing paradigms, such as cloud computing, increases the possibility of large-scale distributed data collection from multiple sources. While the collected data offer tremendous opportunities for mining useful information, there is also a threat to privacy because data in raw form often contain sensitive information about individuals. Service providers have the ability to collect large amounts of user data. Sometimes, a set of providers may try to aggregate their data for specific data mining tasks. We aim to solve this challenge and propose a mechanism that can verify whether the utility of the published data is equal to the utility claimed by the publisher without compromising the data privacy, namely disclosing the raw data, even when the publisher is dishonest. Since the differential privacy model is becoming de facto standard for privacy preserving as it can provide rigorous privacy protection, our work in this paper focuses on differentially private data publishing mechanisms.
In this work privacy-preserving utility verification mechanism for DiffPart, a differentially private anonymization algorithm designed for set-valued data. DiffPart perturbs the frequencies of the records based on a context-free taxonomy tree and no items in the original data are generalized. Our proposal solves the challenge to verify the utility of the published data based on the encrypted frequencies of the original data records instead of their plain values. As a result, it can protect the original data from the verifying parties (i.e., the data users) because they cannot learn whether or how many times a specific record appears in the raw dataset without knowing its real frequency. In addition, since the encrypted frequencies are provided by the publisher, we also present a scheme for the verifying parties to incrementally verify its correctness.

We then extend the above mechanism to DiffGen, a differentially private anonymization algorithm designed for relational data. Different from DiffPart, DiffGen may generalize the attribute values before perturbing the frequency of each record. Information losses are caused by both the generalization and the perturbation. These two kinds of information losses are measured separately by distinct utility metrics. We take both into consideration. Our analysis shows that the utility verification for generalization operations can be carried out with only the published data. As a result, this verification does not need any protection. The utility metric for the perturbation is similar with that for DiffPart. We thus adapt the proposed privacy-preserving mechanism to this verification

5. CONCLUSIONS

In this work, we consider the problem of verifying the Utility of data released by non-interactive differentially private Methods. Similar mechanisms are proposed to achieve the goal for set-valued and relational data respectively. The proposed solutions require the publisher to provide auxiliary datasets in cipher text along with the publishing data. The providers then sequentially verify the auxiliary datasets to see whether their data is correctly involved.

6. 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.
[11] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in Proc. Theory Cryptogr., 2006, pp. 265–284.
[12] L. Fan, L. Xiong, and V. Sunderam, "FAST: Differentially private real-time aggregate monitor with filtering and adaptive sampling," in Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD), 2013, pp. 1065–1068.
[13] D. M. Freeman, "Converting pairing-based cryptosystems from composite-order groups to prime-order groups," in Proc. 29th Annu. Int. Conf. Theory Appl. Cryptogr. Techn. (EUROCRYPT), 2010, pp. 44–61.
[14] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Comput. Surv., vol. 42, no. 4, 2010, Art. no. 14.
[15] Y. Hong, J. Vaidya, H. Lu, P. Karras, and S. Goel, "Collaborative search log sanitization: Toward differential privacy and boosted utility," IEEE Trans. Dependable Secure Comput., vol. 12, no. 5, pp. 504–518, Sep./Oct. 2015.
[16] W. Jiang and C. Clifton, "A secure distributed framework for achieving k-anonymity," Int. J. Very Large Data Bases, vol. 15, no. 4, pp. 316–333, Nov. 2006.