OAIJSE

OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

PRIVACY PRESERVATION AND PUBLISHING USING DEPTH TRACING ALGORITHM

Ms. Sonal Kadam[1], Prof. Harish Barapatre[2]
Department of Computer Engineering Yadavrao Tasgaonkar Institute of Engineering and Technology Maharashtra,India[1]
Email:sonalkadam0312@gmail.com
Department Of Computer Engineering Yadavrao Tasgaonkar Institute Of Engineering and Technology Maharashtra, India[2]
Email:harish.barapatre@tasgaonkartech.com

Abstract- The collection of digital information by governments, corporations and individuals has created tremendous opportunities for knowledge and information-based decision making. Driven by mutual benefits, or by regulations that require certain data to be published, there is a demand for the exchange and publication of data among various parties. Data in its original form, however, typically contains sensitive information about individuals, and publishing such data will violate individual privacy. The current practice in data publishing relies mainly on policies and guidelines as to what type of data can be published and on agreements on the use of published data. This approach alone may lead to excessive data distortion or insufficient protection. Privacy-preserving data publishing (PPDP) provides methods and tools for publishing useful information while preserving data privacy. Recently, PPDP has received considerable attention in research communities, and many approaches have been proposed for different data publishing scenarios. In this survey, we will systematically summarize and evaluate different approaches to PPDP, study the challenges in practical data publishing, clarify the differences and requirements that distinguish PPDP from other related problems, and propose future research directions.

Keywords—*depth-tracing, privacy preservation, publishing, information- based decision making, data market.*

I INTRODUCTION

To integrate truthfulness and privacy preservation in practical data market, there are four major challenges. The first and the thorniest design challenge is that verifying the truthfulness of data collection and preserving the privacy seem to be contradictory objectives. Ensuring the truthfulness of data collection allows the data consumers to verify the validities of data contributor's identities and the content of raw data, whereas privacy preservation tends to prevent them from learning these confidential contents. Specifically, the property of non-repudiation in classical digital signature schemes implies that the signature is unforgeable, and any third party is able to verify the authenticity of a data submitter using her public key and the corresponding digital certificate, i.e., the truthfulness of data collection in our model. However, the verification in digital signature schemes requires the knowledge of raw data, and can easily leak a data contributor's real identity [9]. Regarding a message authentication code (MAC), the data contributors and the data consumers need to agree on a shared secret key, which is unpractical in data markets.

Yet, another challenge comes from data processing, which makes verifying the truthfulness of data collection even harder. Now a days, more and more data markets provide data services rather than directly offering raw data. The following three reasons account for such trend:1) For the data contributors, they have several privacy concerns [8]. Nevertheless, the service-based trading mode, which has hidden the sensitive raw data, alleviates their concerns. 2) For the service provider, semantically rich and insightful data services can bring in more profits. [10] 3) For the data consumers, data copyright infringement [11] and datasets resale [12] are serious. However, such a data trading mode differs from most of conventional data sharing scenarios, e.g., data publishing [13]. Besides, the result of data processing may no longer be semantically consistent with the raw data [14], which makes the data consumer hard to believe the truthfulness of data collection. In addition, the digital signatures on raw data become invalid for the data processing result, which discourages the data consumer from doing verification as mentioned above.

The third challenge lies in how to guarantee the truthfulness of data processing, under the information asymmetry between the data consumer and the service provider due to data confidentiality. In particular, to ensure data confidentiality against the data consumer, the service provider can employ a conventional symmetric/asymmetric cryptosystem, and can let the data contributors encrypt their raw data. Unfortunately, a hidden problem arisen is that the data consumer fails to verify the correctness and completeness of a returned data service. Even worse, some greedy service providers may exploits this vulnerability to reduce operation cost during the execution of data processing, e.g., they might return an incomplete data service without processing the whole raw data set, or even return an outright fake result without processing the data from designated data sources.

Last but not least, the fourth design challenge is the efficient requirement of data markets, especially for data acquisition, i.e., the service provider should be able to collect data from a large number of data contributors with low latency. Due to the timeliness of some kinds of person specific data, the service provider has to periodically collect fresh raw data to meet the diverse demands of high quality data services. For example, 25 billion data collection activities take place on Gnip, every day [2]. Meanwhile, the service provider needs to verify data authentication and data integrity. One basic approach is to let each data contributor sign her raw data. However, classical digital signature schemes, which verify the received signatures one after another, may fail to satisfy the stringent time requirement of data markets. Furthermore, the maintenance of digital certificate under the traditional Public Key Infrastructure (PKI) also incurs significant communication overhead. Under such circumstances, verifying a large number of signatures sequentially certainly becomes the processing bottleneck at the service provider.

II EXISTING SYSTEM

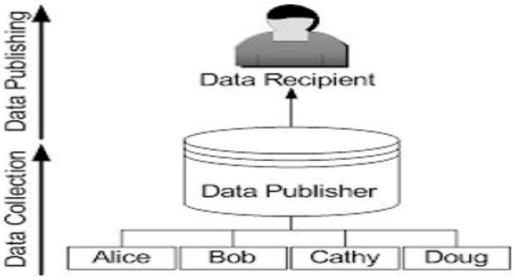

Fig.1:A two-layer system model data markets

We consider a two-layer system model for data markets as shown in figure 1. The model has a data acquisition layer and a data trading layer. There are four major kinds of entities, including data contributors, a service provider, data consumers, and a registration centre.

There are two models of data publishers [Gehrke 2006]. In the untrusted model, the data publisher is not trusted and many attempt to identify sensitive information from record owners. Various cryptographic sloutions [Yang et al.2005]; anonymous communications [Chaum 1981; Jakobsson et al. 2002]; and statstical methods [Warner 1965] were proposed to collect records anonymously from their owners without revealing the owner’s identity. In the trusted model, the data publisher is trustworthy and record owners are willing to provide their personal information to the data publisher; however, the trust is not trsnsitive to the data receipent. We will assume the trusted model of data publishers and consider privacy issues in the data publishing phase.

III AIM OF PROPOSED SYSTEM

In recent year, due to the widespread use of the Internet of Things (IoTs) and Big data technologies, the amount of personal information collected by some large internet companies and social networking services has reached an unprecedented level [1]. These personal data are very valuable for the public and private sectors to improve their products or services. However, raw data may contain sensitive information about individuals, access to personal data is strictly limited. In particular, some companies and organizations have collected a large amount of personal information that is very useful to third parties.

In fact, some start-ups currently developing applications to support this trend. For example, data coup [5] have created the world’s first personal data market, which contains thousands of personal information about its users (i.e., location, profession, health, etc.). However, they did not provide an effective pricing mechanism. The subscription fee for Data coup is fixed. In addition, the data provider’s privacy attitude is different, so the utility [6,7]of publishing data is different, which involves compensation for data providers during the transaction and is closely related to the operating costs of the data market. This is the reason we are develop a system that satisfy: 1) It is the first secure mechanism for data markets achieving both truthfulness and privacy preservation. 2) System is structured internally in a way of Encrypt-then-sign using partially homomorphic encryption and identity-baed signature. It enforces the service provider to truthfully collect and to process real data. Besides, TPDM incorporates a two-layer batch verification scheme with an efficient outcome verification scheme, which can drastically reduce computation overhead. 3) We instructively instantiate this system with two kinds of practical data services, namely profile matching and data distribution.

IV SYSTEM ARCHITECURE

The data publisher is not required to have the knowledge to perform data mining on behalf of the data recipient. Any data mining activities have to be performed by the data recipient after receiving the data from the data publisher. Sometimes, the data publisher does not even know who the recipients are at the time of publication, or has no interest in data mining.

In this system, one assumption is that the data recipient could also be an attacker. For example, the data recipient , says a drug research company, is a trustworthy entity; however, it is difficult to guarantee that all staff in the company is trustworthy as well. It emphasizes publishing data records about individuals (i.e., micro data). Clearly, this requirement is more stringent than publishing data mining results, such as classifiers, association rules, or statistics about groups of individuals.

In some data publishing scenarios, it is important that each published record corresponds to an existing individual in real life. Consider the example of patient records. The pharmaceutical researcher (the data recipient) may need to examine the actual patient records to discover some previously unknown side effects of the tested drug. If a published record does not correspond to an existing patient in real life, it is difficult to deploy data mining results in the real world. Randomized and synthetic data do not meet this requirement. Although an encrypted record corresponds to areal life patient, the encryption hides the semantics required for acting on the patient represented.

In this system firstly the efficient secure scheme for data markets simultaneously guarantees data truthfulness and privacy preservation. In this system, user purchase product than he/she can send review to the system than system first check whether the contributors are authorized person or not. Under a specific data service, this system provides privacy preservation and variability. To counter partial data collection attack, each data consumer should be enabled to verify whether raw data are really provided by registered data contributors, i.e., truthfulness of data collection in the data trading layer. On the other hand, the data consumer should have the capability to verify the correctness and completeness of a returned data service in order to combat no/partial data processing attack.

We here use the term truthfulness of data processing in the data trading layer to represent the integrated requirement of correctness and completeness of data processing results. Using terminology from the sign-encryption scheme [2], TPDM is structed internally in a way of Encrypt-then-sign, using partially homomorphic encryption and identity-based signature. It enforces the service provider to truthfully collect and process real data. The essence of TPDM is to first synchronize data processing and signature verification into the same ciphertext space, and then to tightly integrate data processing with outcome verification via the homomorphic properties.

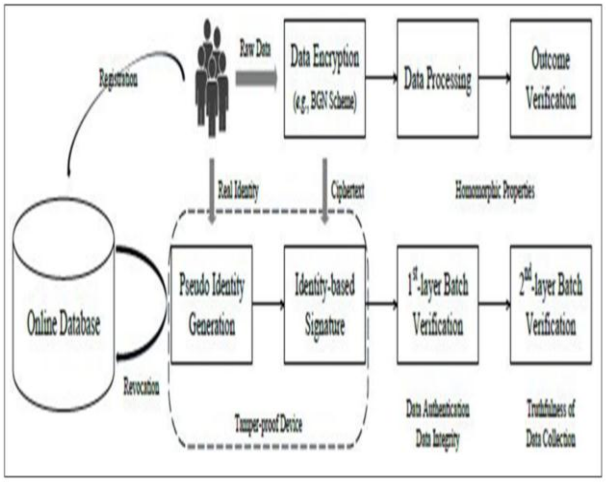
Registration
Raw Data
Data Encryption (e.g., BGN Scheme)
Data Processing
Outcome Verification
Real Identity
Ciphertext
Homomorphic Properties
Online Database
Pseudo Identity Generation
Identity-based Signature
1st-layer Batch Verification
2nd-layer Batch Verification
Revocation
Tamper-proof Device
Data Authentication Data Integrity
Truthfulness of Data Collection

Fig.2: Prosed System Flow

Relevant objective of the proposed system are: 1) To propose a efficient secure scheme for data markets, which simultaneously guarantees data truthfulness and privacy preservation. 2) To achieve the ultimate goal of the service provider in a data market is to maximize their profit. 3) To first secure mechanism for data markets achieving both data truthfulness and privacy preservation. 4) To ensure truthfulness and protecting the privacies of data contributors. 5) To hide content of raw data from data consumers to guarantee data confidentiality, even if the real identities of the data contributors are hidden.

V METHODOLOGY

Phase I: Initialization

We assume that the registration center sets up the system parameters at the beginning of data trading.

Phase II: Signing Key Generation

To achieve anonymous authentication in data markets, the tamper-proof device is utilized to generate a pair of pseudo identity and secrete key for each registered data- contributors.

Phase III: Data submission

For secure submission of raw data, we need to consider several requirements, including confidentiality, authentication, and integrity. To provide data confidentiality, we employ partially homomorphic encryption. Besides, To guarantee data authentication and data integrity the encrypted raw data should be signed before submission, and be verified after reception.

Phase IV: Data Processing and Verification

In this phase, we consider two-layer batch verifications, i.e., verifications conducted by both the service provider and the data consumer. Between the two-layer batch verifications, we introduce data processing and signatures aggregation done by the service provider.

At last, we present outcome verification conducted by the data consumer.

Phase V: Tracing and Revocation

The two-layer batch verifications only hold when all the signatures are valid, and fail even when there is a single invalid signature. In practice, a signature batch may contain invalid one(s) caused by accidental data corruption or possibly malicious activities launched by an external attacker. Traditional batch verified would reject the entire batch, even if there is a single invalid signature, and thus waste the other valid data items. Therefore, tracing and/or recollecting invalid data items and their corresponding signatures are important in practice. If the second-layer batch verification fails, the data consumer can require the service provider to find out the invalid signature(s). Similarly, if the first-layer batch verification fails, the service provider has to find out the invalid one(s) by herself.

Algorithm

ALGORITHM 1:- MODIFIED DEPTH-TRACING

```
Initialization: S = {σ1,··· ,σn}, head = 1, tail = n, limit = l
whitelist = ∅, blacklist = ∅, resubmitlist = ∅
Function l-DEPTH-TRACING (S, head, tail, limit)
    If |whitelist| + |blacklist| = n or limit = 0 then
        return
    else if CHECK-VALID (S, head, tail) = true then
        ADD-TO-WHITELIST (head, tail)
    else if head = tail then Single signature verification
        ADD-TO-BLACKLIST (head, tail)
    else Batch signatures verification from σ head to σ tail
        mid = [head + tail]/2
        l-DEPTH-TRACING (S, head, mid, limit−1)
        l-DEPTH-TRACING (S, mid + 1, tail, limit−1)
```

VI APPLICATION

The proposed system is used in real time basis for eg. in social media, Yahoo, Amazon, Flipkart etc.

An advanced aggregate statistic, where the service provider wants to capture the underlying distribution over the collected dataset, and to offer such a distribution as a data service to the data consumer. For example, an analyst, as the data consumer, may want to learn the distribution of residential energy consumptions

Classic data service in social networking, i.e., fine-grained profile matching. In particular, a data consumer's friending strategy can be derived from a large scale of data contributions.

REFERENCES

[1] Qingbei Guo , Lin Wang , Shouning Qu,"Parallel rational world based privacy preservation mechanism for group privacy" IEEE 2016

[2] Chaoyue Niu ; Zhenzhe Zheng ; Fan Wu ; Xiaofeng Gao ; Guihai Chen, "Trading Data in Good Faith: Integrating Truthfulness and Privacy Preservation in DataMarkets" IEEE 2017

[3] Wenny Rahayu, " Privacy Preservation in Big Databases" IEEE 2017

[4] M. Barbaro, T. Zeller, and S. Hansell," A [2006] face is exposed for AOL searcher no. 4417749," New York Times, Aug. 2006.

[5] "TRUSTe/NCSA Consumer Privacy Infographic – US Edition," [2016] https://www: truste: com /resources/privacy-research / ncsa-consumerprivacy-index-us/.

[6] K.Ren,W.Lou,K.Kim, and R.Deng," [2006] A novel privacy preserving authentication and access control scheme for pervasive computing environments," IEEE Transactions on Vehicular Technology, vol.55,no.4,pp.1373–1384, 2006.

[7] Prajakta Tambe , Deepali Vora, "Privacy preservation on social network using data sanitization" IEEE 2016

[8] M. Balazinska, B. Howe, and D. Suciu, [2011] "Data markets in the cloud: An opportunity for the database community," PVLDB, vol. 4, no. 12, pp. 1482–1485, 2011.

[9] [1994] Digital Signature Standard, Federal Information Processing Standards Publication 186, May 1994.

[10] Odlyzko A., [2000] Discrete logarithms: The past and the Future; Designs, Codes and Cryptography, (2000), 129-145.

[11] McCurleyK., [1990] the discrete logarithm problem, Proceedings of Symposiain Applies Mathematica, Vol.42, 1990, 49-74.

[12] Lidl, Niederreiter [1997], Finite Fields (2nd ed.), Cambridge University, Press.

[13] Neal Koblitz, Algebraic Aspects of Cryptography, Springer.

[14] Lilly P.L, Saju M.I., [2014] A method of designing a public-key cryptosystem based on discrete logarithm problem, IRJPA-4(11), 2014, 628-630.

[15] Diffie W., Helman M.E., [1976] New Directions in Cryptography, IEEE Transactions on information theory, Vol. IT-22, Nov.1976, 644- 654.