Achieving Data Truthfulness and Privacy Preservation in Data Markets

First Author: V Sreenadh, Assistant professor in Department of CSE, Ramireddy Subbarami Reddy (Rsr) Engineering College, Kadanuthala, Andhra Pradesh.

Second Author: Sk Badulla, Assistant Professor in Department of CSE, Ramireddy Subbarami Reddy (Rsr) Engineering College, Kadanuthala, Andhra Pradesh

Abstract

Many online information platforms have developed as an important business model to meet society's demands for person-specific data, where a service provider gathers raw data from data contributors and then provides value-added data services to data consumers. However, in the layer of data trading, consumers of data face a critical issue: how to confirm that the service provider has accurately gathered and processed their data. Additionally, the data contributors typically don't want the data consumers to know their true identities or sensitive personal information. TPDM, which effectively blends Truthfulness and Privacy protection in Data Markets, is the solution we provide in this study. Internally, TPDM is organized as an Encrypt-then-Sign system that makes use of identity-based signatures and partially homomorphic encryption. It enables batch verification, data processing, and outcome verification simultaneously. while preserving data privacy and identity protection.

1. Introduction

Data mining is that the method of analyzing knowledge from totally different views and summarizing it into helpful info. Data processing software package is one in all variety of analytical tools for analyzing knowledge. It permits users to investigate knowledge from many various dimensions or angles, categorise it, and summarize the relationships known. Technically, data mining is that the process of finding correlations or patterns among dozens of fields in massive relative databases. Data processing involves six common categories of tasks: Anomaly detection the identification of surprising knowledge records, which may be attention grabbing or knowledge errors that need more investigation. Dependency modelling searches for relationships between variables. For instance a grocery would possibly gather knowledge on client buying habits. Exploitation association rule learning, the grocery will confirm that merchandise are often bought along and use this info for promoting functions. This is often generally stated as market basket analysis. Bunch is that the task of discovering teams and structures within the knowledge that are in a way or another "similar", while not exploitation better known structures within the knowledge. Therefore, so as to reduce the expenditure for knowledge acquisition, associate timeserving method for the service supplier is to mingle some imitative or artificial data into the information sets. Yet, to scale back operation price, a strategic service supplier could give data services supported a set of the entire information set, or perhaps come a faux result while not process the Classification is that the task of generalizing better known structure to use to new knowledge. For instance, associate e-mail program would possibly try to classify an e-mail as "legitimate" or as "spam". Regression makes an attempt to seek out a operate that models the information with the smallest amount error. Summarisation providing a lot of compact illustration of the information set, together with image and report generation. In the era of huge knowledge, society has developed associate insatiable appetency for sharing personal knowledge. Realizing the potential of non public data's value in higher cognitive process and user expertise sweetening, many open info platforms have emerged to change person specific knowledge to be changed on the web. However, there exists a crucial security downside in these market-based platforms, i.e., it's troublesome to ensure the honesties in terms of information assortment and processing, particularly once privacies of the information contributors are required to be preserved. Guaranteeing honesties and protective the privacies of information contributors are each necessary to the long run healthy development of data markets. On one hand, the last word goal of the service supplier during acknowledge market is to maximise information from selected data sources. The service supplier ought to be ready to collect data from an oversized range of information contributors with low latency. Because of the timeliness of some varieties of person-specific knowledge, the service supplier should sporadically collect recent information to satisfy the varied demands of high-quality data services. For instance, twenty five billion knowledge assortment activities happen. Meanwhile, the service supplier has to verify knowledge authentication and data integrity. One basic approach is to let every data contributor sign her information. However, classical digital signature schemes, that verify the received signatures one once another, could fail to satisfy the rigorous time demand of information market.

2. Literature Survey

In recent years, data market design has gained increasing interest, especially from the database community. The seminal paper [10] by Balazinska et al. discusses the implications of the emerging digital data markets, and lists the research opportunities in this direction. Li et al. proposed a theory of pricing private data based on differential privacy. Upadhyaya et al. developed a middleware system, called DataLawyer, to formally specify data use policies, and to automatically enforce these pre-defined terms during data usage. Jung et al. [12] focused on the datasets resale issue at the dishonest data consumers.

To get a tradeoff between functionality and performance, partially homomorphic encryption (PHE) schemes were exploited to enable practical computation on encrypted data. Unlike those prohibitively slow fully homomorphic encryption (FHE) schemes that support arbitrary operations, PHE schemes focus on specific function(s), and achieve better performance in practice. A celebrated exam- ple is the Paillier cryptosystem , which preserves the group homomorphism of addition and allows multiplication by a constant. Thus, it can be utilized in data aggregation [19] and interactive personalized recommendation [23]. Yet, another one is ElGamal encryption [22], which supports homomorphic multiplication, and it is widely employed in voting. Moreover, the BGN scheme [18] facilitates one extra multiplication followed by multiple addition- s, which in turn allows the oblivious evaluation of quadratic multivariate polynomials, e.g., shortest distance query [27] and optimal meeting location decision. Lastly, several stripped-down homomorphic encryption schemes were employed to facilitate practical machine learning algorithms on encrypted data, such as linear means classifier, naïve Bayes, neural networks [25], and so on.

Ensuring truthfulness and protecting the privacies of da-ta contributors are both important to the long term healthy development of data markets. On one hand, the ultimate goal of the service provider in a data market is to maximize her profit. Therefore, in order to minimize the expenditure for data acquisition, an opportunistic way for the service provider is to mingle some bogus or synthetic data into the raw data sets. Yet, to reduce operation cost, a strategic service provider may provide data services based on a subset of the whole raw data set, or even return a fake result without processing the data from designated data sources. However, if such speculative and illegal behaviors cannot be identified and prohibited, it will cause heavy losses to the data consumers, and thus destabilize the data market. On the other hand, while unleashing the power of personal data, it is the bottom line of every business to respect the privacies of data contributors. The debacle, which follows AOL's public release of "anonymized" search records of its customers, highlights the potential risk to individuals in sharing personal data with private companies [7]. Besides, according to the survey report of 2016 TRUSTe/NCSA Consumer Privacy Infographic - US Edition [8], 89% say they avoid companies that do not protect their privacies. Therefore, the content of raw data should not be disclosed to data consumers to guarantee data confidentiality, even if the real identities of the data contributors are hidden.

To integrate truthfulness and privacy preservation in a 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 truthful-ness of data collection allows the data consumers to verify the validities of data contributors' 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 sub-mitter using her public key and the corresponding digital certificate, i.e., the truthfulness of data collection in our mod-el. 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. Nowadays, more and more data markets pro-vide data services rather than directly offering raw data. The following three reasons account for such a 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 process-ing result, which discourages the data consumer from doing verification as mentioned above. Moreover, although data provenance [15] helps to determine the derivation history of a data processing result, it cannot guarantee the truthfulness of data collection.

The third challenge lies in how to guarantee the truthful-ness of data processing, under the information asymmetry between the data consumer and the service provider due to data confidentiality. In particular, to ensure data confi-dentiality against the data consumer, the service provider can employ a conventional symmetric/asymmetric crypto-system, 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 exploit 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 efficiency 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 col-lect 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 certificates under the traditional Public Key Infrastructure (PKI) also incurs significant communication over-head. Under such circumstances, verifying a large number of signatures sequentially certainly becomes the processing bottleneck at the service provider.

3. Proposed Model

By jointly considering above four challenges, we propose TPDM, which achieves both Truthfulness and Privacy preservation in Data Markets. TPDM first exploits partially homomorphic encryption to construct a ciphertext space, which enables the service provider to launch data services and the data consumers to verify the correctness and completeness of data processing results, while maintaining data confidentiality. In contrast to classical digital signature schemes, which are operated over plaintexts, our new identity-based signature scheme is conducted in the ciphertext space. Furthermore, each data contributor's signature is derived from her real identity, and is unforgeable against the service provider or other external attackers. This appealing property can convince data consumers that the service provider has truthfully.

collected data. To reduce the latency caused by verifying a bulk of signatures, we propose a two-layer batch verification scheme, which is built on the bilinearity of admissible pairing. At last, TPDM realizes identity preservation and revocability by carefully adopting ElGamal encryption and introducing a semi-honest registration center.

To the best of our knowledge, TPDM is the first secure mechanism for data markets achieving both data truthfulness and privacy preservation.

TPDM is structured 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 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.

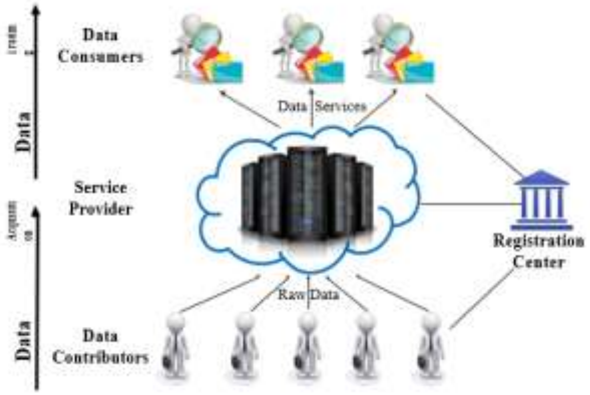

Fig. 1. A two-layer system model for data markets.

In the data acquisition layer, the service provider pro-cures massive raw data from the data contributors, such as social network users, mobile smart devices, smart meters, and so on. In order to incentivize more data contributors to actively submit high-quality data, the service provider needs to reward those valid ones to compensate their data collection costs. For the sake of security, each registered data contributor is equipped with a tamper-proof device. The tamper-proof device can be implemented in the form of either specific hardware [16] or software [17]. It prevents any adversary from extracting the information stored in the device, including cryptographic keys, codes, and data.

We consider that the service provider is cloud based, and has abundant computing resources, network bandwidths, and storage space. Besides, she tends to offer semantically rich and value-added data services to data consumers rather than directly revealing sensitive raw data, e.g., social network analyses, data distributions, personalized recommendations, and aggregate statistics.

The registration center maintains an online database of registrations, and assigns each registered data contributor an identity and a password to activate the tamper-proof device. Besides, she maintains an official website, called certificated bulletin board [18], on which the legitimate system participants can publish essential information, e.g., whitelists, blacklists, resubmit-lists, and reward-lists of data contributors. Yet, another duty of the registration center is to set up the parameters for a signature scheme and a cryptosystem. To avoid being a single point of failure or bottleneck, redundant registration centers, which have identical functionalities and databases, can be installed.

Challenger

In this section, we focus on attacks in practical data markets, and define corresponding security requirements.

First, we consider that a malicious data contributor or an external attacker may impersonate other legitimate data contributors to submit possibly bogus raw data. Besides, some malicious attackers may deliberately modify raw data during submission. Hence, the service provider needs to confirm that raw data are indeed sent unaltered by registered data contributors, i.e., to guarantee data authentication and data integrity in the data acquisition layer.

Second, the service provider in the data market might be greedy, and attempts to maximize her profit by launching the following two types of attacks:

Partial data collection: To cut down the expenditure on data acquisition, the service provider may insert bogus data into the raw data set.

No/Partial data processing: To reduce the operation cost, the service provider may try to return a fake result without processing the data from designated sources, or to provide data services based on a subset of the whole raw data set.

On one hand, 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.

Third, we assume that some honest-but-curious data contributors, the service provider, the data consumers, and external attackers, e.g., eavesdroppers, may glean sensitive information from raw data, and recognize real identities of data contributors for illegal purposes, e.g., an attacker can infer a data contributor's home location from her GPS records. Hence, raw data of a data contributor should be kept secret from these system participants, i.e., data confidentiality. Besides, an outside observer cannot reveal a data contributor's real identity by analyzing data sets sent by her, i.e., identity preservation.

Fourth, a minority of data contributors may try to behave illegally, e.g., launching attacks as mentioned above, if there is no punishment. To prevent this threat, the registration center should have the ability to retrieve a data contributor's real identity, and revoke it from further usage, when her signature is in dispute, i.e., traceability and revocability.

4. Conclusion

we have proposed the first efficient secure scheme TPDM for data markets, which simultaneously guarantees data truthfulness and privacy preservation. In TPDM, the data contributors have to truthfully submit their own data, but cannot impersonate others. Besides, the service provider is enforced to truthfully collect and process data. Furthermore, both the personally identifiable information and the sensitive raw data of data contributors are well protected. In addition, we have instantiated TPDM with two different data services, and extensively evaluated their performances on two real-world datasets. Evaluation results have demonstrated the scalability of TPDM in the context of large user base, especially from computation and communication overheads

5. References

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

[2] "2016 TRUSTe/NCSA Consumer Privacy Infographic – US Edition,"

https://www.truste.com/resources/privacy-research/ ncsa-consumer-privacy-index-us/.

[3] K. Ren, W. Lou, K. Kim, and R. Deng, "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.

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

[5] P. Upadhyaya, M. Balazinska, and D. Suciu, "Automatic enforce- ment of data use policies with datalawyer," in SIGMOD, 2015.

[6] T. Jung, X.-Y. Li, W. Huang, J. Qian, L. Chen, J. Han, J. Hou, and C. Su, "AccountTrade: accountable protocols for big data trading against dishonest consumers," in INFOCOM, 2017.

[7] G. Ghinita, P. Kalnis, and Y. Tao, "Anonymous publication of sensitive transactional data," IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 2, pp. 161–174, 2011.

[8] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Comput- ing Surveys, vol. 42, no. 4, pp. 1–53, Jun. 2010.

[9] R. Ikeda, A. D. Sarma, and J. Widom, "Logical provenance in data-oriented workflows?" in ICDE, 2013.

[10] M. Raya and J. Hubaux, "Securing vehicular ad hoc networks," Journal of Computer Security, vol. 15, no. 1, pp. 39–68, 2007.

[11] T. W. Chim, S. Yiu, L. C. K. Hui, and V. O. K. Li, "SPECS: secure and privacy enhancing communications schemes for VANETs," Ad Hoc Networks, vol. 9, no. 2, pp. 189 – 203, 2011.

[12] D. Boneh, E. Goh, and K. Nissim, "Evaluating 2-dnf formulas on ciphertexts," in TCC, 2005.

[13] R. A. Popa, A. J. Blumberg, H. Balakrishnan, and F. H. Li, "Privacy and accountability for location-based aggregate statistics," in CCS, 2011.

[14] J. H. An, Y. Dodis, and T. Rabin, "On the security of joint signature and encryption," in EUROCRYPT, 2002.

[15] D. Boneh and M. Franklin, "Identity-based encryption from the weil pairing," in CRYPTO, 2001.

[16] T. ElGamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," IEEE Transactions on Information Theory, vol. 31, no. 4, pp. 469–472, 1985.

[17] R. Zhang, Y. Zhang, J. Sun, and G. Yan, "Fine-grained private matching for proximity-based mobile social networking," in IN- FOCOM, 2012.

[18] D. Eastlake and P. Jones, "US Secure Hash Algorithm 1 (SHA1)," IETF RFC 3174, 2001.

[19] R. Gilad-Bachrach, N. Dowlin, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing, "CryptoNets: Applying neural networks to encrypted data with high throughput and accuracy," in ICML, 2016.

[20] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in EUROCRYPT, 1999.

[21] X. Meng, S. Kamara, K. Nissim, and G. Kollios, "GRECS: graph encryption for approximate shortest distance queries," in CCS, 2015.

[22] Z. Erkin, T. Veugen, T. Toft, and R. L. Lagendijk, "Generating pri- vate recommendations efficiently using homomorphic encryption and data packing," IEEE Transactions on Information

Forensics and Security, vol. 7, no. 3, pp. 1053–1066, 2012.

[23] C. Niu, Z. Zheng, F. Wu, X. Gao, and G. Chen, "Achieving data truthfulness and privacy preservation in data markets," Tech. Rep., 2018. [Online]. Available: https://www.dropbox.com/s/7m3jwcio18q4sy0/Technical Report for TPDM.pdf?dl=0

[24] Z. Zheng, Y. Peng, F. Wu, S. Tang, and G. Chen, "Trading data in the crowd: Profit-driven data acquisition for mobile crowdsens- ing," IEEE Journal on Selected Areas in Communications, vol. 35, no. 2, pp. 486–501, 2017.