International Journal of Research
Available at https://journals.pen2print.org/index.php/ijr/
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 03
March 2019

Achieving Data Truthfulness and Privacy Preservation in Data Markets

Ms. Ruqqiya Zufeen, Ms. Maliha Naureen, Ms. Sakina Fatima, Mr. Feroz Amer
[1]Dept. of IT, Deccan College of Engineering and Technology, Hyderabad.
[2]Dept. of IT, Deccan College of Engineering and Technology, Hyderabad.
[3]Dept. of IT, Deccan College of Engineering and Technology, Hyderabad.
[4]Assistant Professor, Dept. of IT, Deccan College of Engineering and Technology, Hyderabad.

Abstract: *As a significant business paradigm, many online information platforms have emerged to satisfy society's needs for person-specific data, where a service provider collects raw data from data contributors, and then offers value-added data services to data consumers. However, in the data trading layer, the data consumers face a pressing problem, i.e., how to verify whether the service provider has truthfully collected and processed data? Furthermore, the data contributors are usually unwilling to reveal their sensitive personal data and real identities to the data consumers. In this paper, we propose TPDM, which efficiently integrates Truthfulness and Privacy preservation in Data Markets. TPDM is structured internally in an Encrypt-then-Sign fashion, using partially homomorphic encryption and identity-based signature. It simultaneously facilitates batch verification, data processing, and outcome verification, while maintaining identity preservation and data confidentiality. We also instantiate TPDM with a profile matching service and a data distribution service, and extensively evaluate their performances on Yahoo! Music ratings dataset and 2009 RECS dataset, respectively. Our analysis and evaluation results reveal that TPDM achieves several desirable properties, while incurring low computation and communication overheads when supporting large-scale data markets.*

Keywords: Data markets, data truthfulness, privacy preservation.

1. Introduction

In the era of big data, society has developed an insatiable appetite for sharing personal data. Realizing the potential of personal data's economic value in decision making and user experience enhancement, several open information platforms have emerged to enable person-specific data to be exchanged on the Internet [1], [2], [3], [4], [5]. For example, Gnip, which is Twitter's enterprise API platform, collects social media data from Twitter users, mines deep insights into customized audiences, and provides data analysis solutions to more than 95% of the Fortune 500 [2]. However, there exists a critical security problem in these market-based platforms, i.e., it is difficult to guarantee the truthfulness in terms of data collection and data processing, especially when privacies of the data contributors are needed to be

preserved. Let's examine the role of a pollster in the presidential election as follows. As a reliable source of intelligence, the Gallup Poll [6] uses impeccable data to assist presidential candidates in identifying and monitoring economic and behavioral indicators. In this scenario, simultaneously ensuring truthfulness and preserving privacy require the Gallup Poll to convince the presidential candidates that those indicators are derived from live interviews without leaking any interviewer's real identity (e.g., social security number) or the content of her interview. If raw data sets for drawing these indicators are mixed with even a small number of bogus or synthetic samples, it will exert bad influence on the final election result. Ensuring truthfulness and protecting the privacies of data 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 truthfulness 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 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. Nowadays, more and more data markets provide 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]

2. Literature Survey:

Power-law distribution of the World Wide Web

In this model, the number of new links a site receives at each time step is a random fraction of the number of links the site already has. New sites, each with a different growth rate, appear at an exponential rate. This model yields scatter plots similar to Fig. 1B, and can produce any power-law exponent 1.

Finding and Evaluating Community Structure in Networks

We propose a measure for the strength of the community structure found by our algorithms, which gives us an objective metric for choosing the number of communities into which a network should be divided. We demonstrate that our algorithms are highly effective at discovering community structure in both computer-generated and real-world network data, and show how they can be used to shed light on the sometimes dauntingly complex structure of networked systems.

Improved network community structure improves function prediction

First, we apply a novel method that generates improved modularity solutions than the current state of the art. Second, we develop a better method to use this community information to predict proteins' functions. We discuss when and why this community information is important. Our results should be useful for two distinct scientific communities: first, those use various cost functions to detect community structure, where our new optimization approach will improve solutions, and second, those working to extract novel functional information about individual nodes from large interaction datasets.

Fast parallel algorithm for unfolding of communities in large graphs

The proposed distributed memory parallel algorithm targets the costly first iteration of the initial method by parallelizing it. Experimental results on a MPI setup with 128 parallel processes shows that up to ≈5× performance improvement is achieved as compared to the sequential version while not compromising the correctness of the final result.

3. System Analysis:

3.1 Objective:

In this project, by jointly considering issues in data sharing, we propose TPDM, which achieves both Truthfulness and Privacy preservation in Data Markets.

3.2 Problem Statement:

Ensuring truthfulness and protecting the privacies of data 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 maximizeher 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

3.3 EXISTING SYSTEM

- 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 truth fullness of data collection allows the data consumers to verify the validities of data contributors' identities and the content of raw data.
- Yet, another challenge comes from data processing, which makes verifying the truthfulness of data collection even harder. 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.
- Fourth design challenge is the efficiency requirement of data markets, especially for data acquisition, service provider has to periodically collect fresh raw data to meet the diverse demands of high quality data services.

3.4 DISADVANTAGES OF EXISTING SYSTEM

- Verification in digital signature schemes requires the knowledge of raw data, and can easily leak a data contributor's real identity. 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.

3.5 PROPOSEDSYSTEM

- 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. TPDM realizes identity preservation and revocability by carefully adopting ElGamal encryption.

3.6 ADVANTAGES OF PROPOSED SYSTEM

- 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.

MODULES:

Admin:

Admin can view users who are registered and admin can authorize users. Admin can see all friend requests information. Along with these details admin can view information of different communities available on network and users who are part of that community and check which community is more popular and find most parallel community which is becoming more popular compare to competitive community this is done using mountain model and landslide strategy.

Mountain model:

The Mountain model is integral in this research, and is based on modularity, approximate optimization, and graph theory. It sorts the chain groups by the weights of edges. Owing to the feature of

International Journal of Research
Available at https://journals.pen2print.org/index.php/ijr/
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 03
March 2019

community structures, some chain groups in a community may fall down while surrounding community may rise like mountains. Resolutely, a suitable number of chain groups at the top of mountains are chosen to form new communities.

Landslide strategy:

Number of nodes and edges in the networks remain unchanged, after the community merging operation, the number of edges in the new community equals the sum of the edges in and between the two merged communities. Moreover, the number of edges between the new community and the other communities equals the sum of edges between the merged communities and other communities.

User:

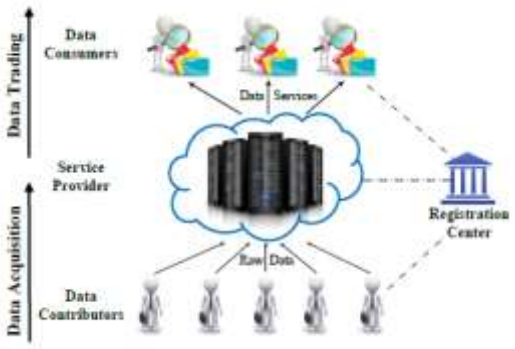
OSN System Construction Module

- In the first module, we develop the Online Social Networking (OSN) system module. We build up the system with the feature of Online Social Networking. Where, this module is used for new user registrations and after registrations the users can login with their authentication.
- Where after the existing users can send messages to privately and publicly, options are built. Users can also share post with others. The user can able to search the other user profiles and public posts. In this module users can also accept and send friend requests.
- With all the basic feature of Online Social Networking System modules is build up in the initial module, to prove and evaluate our system features.

Community creation:

In these modules users can create community and users who are registered with application can post reply to posts posted by respective community.

4. System Design:



4.1 System Architecture

International Journal of Research
Available at https://journals.pen2print.org/index.php/ijr/
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 03
March 2019

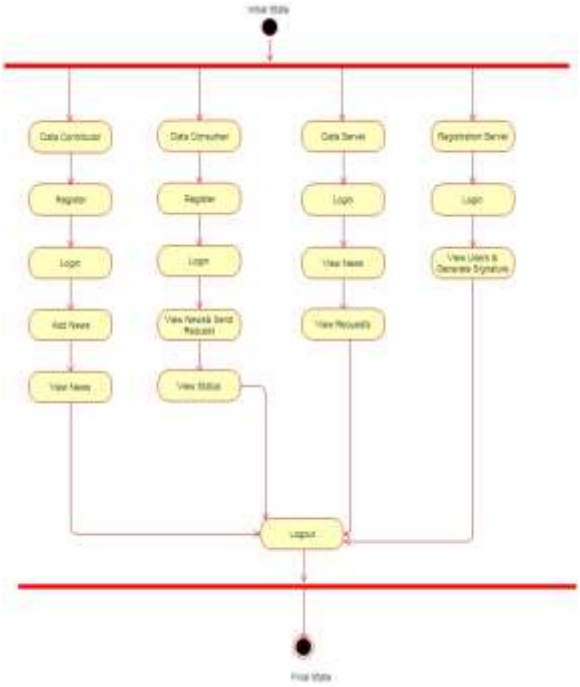


Image 4.2 Activity Diagram

5. Output Results:

Fig 5.1: Home page

Fig 5.2: Data Contributor Login

Fig 5.3: Data Server Login

Fig 5.4: Registration Server

6. Conclusion

In this project, 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. At last, we have shown the feasibility of introducing the semi-honest registration center with detailed theoretical analysis and substantial evaluations. As for further work in data markets, it would be interesting to consider diverse data services with more complex mathematic formulas, e.g., Machine Learning as a Service (MLaaS) [25], [45], [46]. Under a specific data service, it is well-motivated to uncover some novel security problems, such as privacy preservation and verifiability.

.

References

[1] "Microsoft Azure Marketplace," https://datamarket:azure:com/ home/.

[2] "Gnip," https://gnip:com/.

[3] "DataSift," http://datasift:com/.

[4] "Datacoup," https://datacoup:com/.

[5] "Citizenme," https://www:citizenme:com/.

[6] "Gallup Poll," http://www:gallup:com/.

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

[8] "2016 TRUSTe/NCSA Consumer Privacy Infographic - US Edition," https://www:truste:com/resources/privacy-research/ ncsa-consumer-privacy-index-us/.

[9] 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.

[10] 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.

[11] P. Upadhyaya, M. Balazinska, and D. Suciu, "Automatic enforcement of data use policies with datalawyer," in SIGMOD, 2015.

[12] 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.

[13] 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.

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

[15] R. Ikeda, A. D. Sarma, and J. Widom, "Logical provenance in dataoriented workflows?" in ICDE, 2013.

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

[17] 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.

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

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

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