Trading Data in Good Faith: Integrating Truthfulness and Privacy Preservation in Data Markets

Chaoyue Niu, Zhenzhe Zheng, Fan Wu, Xiaofeng Gao, and Guihai Chen
Shanghai Key Laboratory of Scalable Computing and Systems, Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
Email: {rvincency, zhengzhenzhe220}@gmail.com; {fwu, gao-xf, gchen}@cs.sjtu.edu.cn

Introduction

Figure 1. Emerging Data Tradings.

Data Market Model

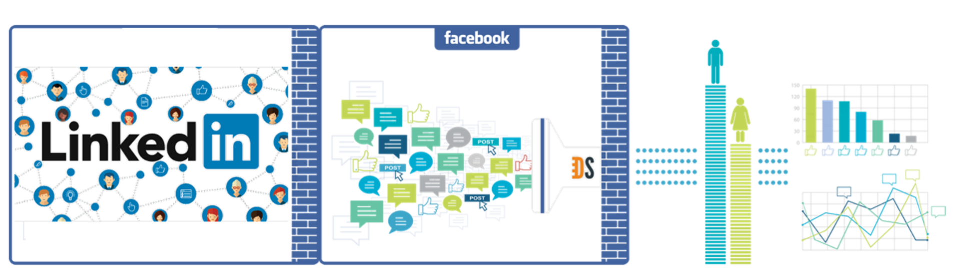
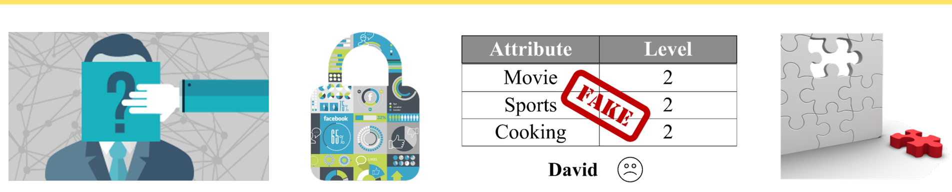


Figure 2. System Architecture of DataSift [5].

Why trade **data services** rather than raw data?

- For data contributors: privacy concerns [1]
- For service provider: value-added data services [2]
- For data consumers: data copyright infringement [12]

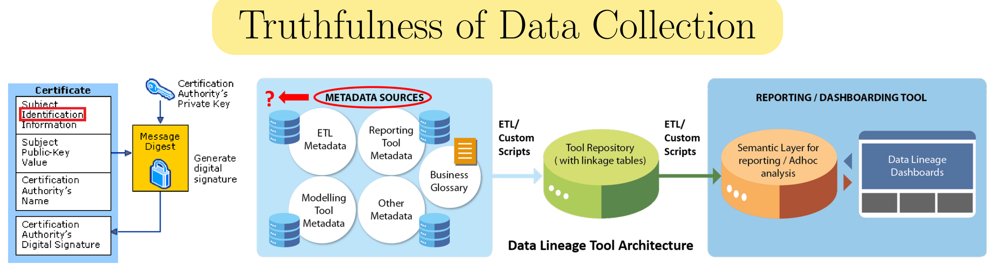
Two Security Issues



- Privacy Preservation
  - Identity Preservation: real identity, *e.g.*, SSN
  - Data Confidentiality: mainly against data consumers
- Data Truthfulness
  - *Partial Data Collection Attack*
  - *No/Partial Data Processing Attack*

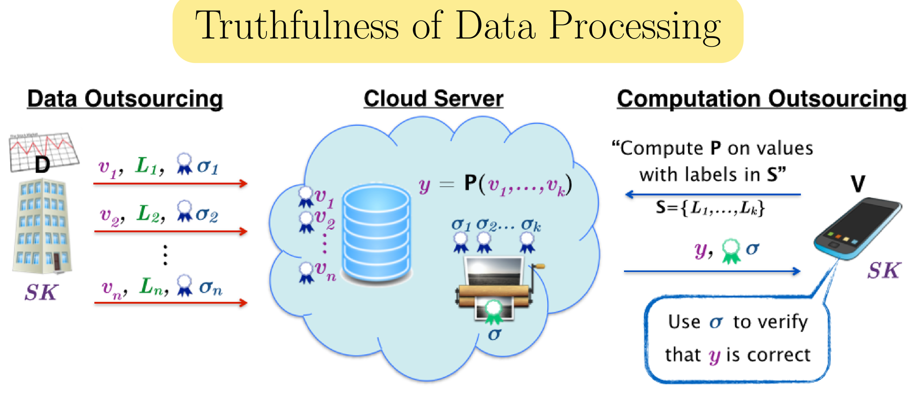
Design Challenges

Truthfulness of Data Collection



- Privacy preservation
  - Digital signature (**non-repudiation**)
  - Message authentication code? secret key sharing
- Data processing
  - Semantic inconsistency [8]
  - Data provenance/lineage [10]

Truthfulness of Data Processing



- Information asymmetry due to data confidentiality
- Differs from verifiable/outsourced computing [7]

Efficiency Requirements

- Large-scale data acquisition (*e.g.*, on DataSift [5] and Gnip [9])
- Data authentication and data integrity? sequential verification, PKI maintenance (computation and communication overheads)

Profile Matching Data Service

Attribute	Level
Movie	3
Sports	0
Cooking	5

Alice

Attribute	Level
Movie	2
Sports	5
Cooking	1

Bob

Attribute	Level
Movie	3
Sports	4
Cooking	2

Charlie $\delta = 2$

Figure 3. Motivating Example: An Illustration of Fine-grained Profile Matching [14].

1. The service provider defines a public attribute set $\mathbb{A} = \{A_1, A_2, \cdots, A_\beta\}$.
2. A data contributor o_i, *e.g.*, a Twitter or OkCupid user, selects an integer u_{ij} to indicate her level of interest in A_j, and thus forms her profile vector $\vec{U}_i = (u_{i1}, u_{i2}, \cdots, u_{i\beta})$.
3. To generate a customized friending strategy, the data consumer also needs to provide her profile vector $\vec{V} = (v_1, v_2, \cdots, v_\beta)$ and an acceptable similarity threshold δ.
4. Without loss of generality, we assume that the service provider employs *Euclidean distance* $f(\cdot)$ to measure the similarity difference, where $f(\vec{U}_i, \vec{V}) = \sqrt{\sum_{j=1}^{\beta}(u_{ij} - v_j)^2}$.

Our Approach: TPDM

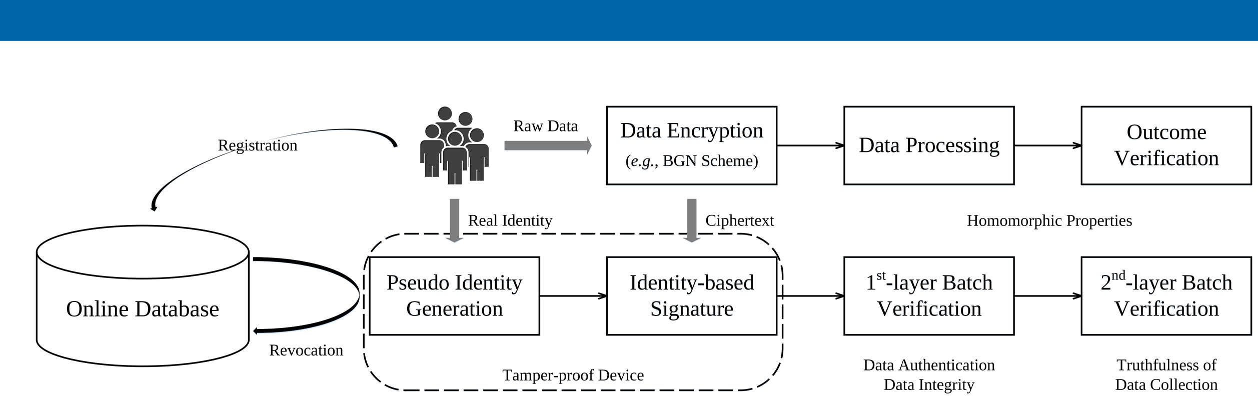


Figure 4. Design Overview of TPDM.

Phase I: Initialization

- The registration center sets up the system parameters at the beginning of data trading:
  - Three multiplicative cyclic groups $\mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_T$ with the same prime order q; g_1, g_2 are generators of $\mathbb{G}_1, \mathbb{G}_2$, respectively; an admissible pairing $\hat{e}: \mathbb{G}_1 \times \mathbb{G}_2 \to \mathbb{G}_T$ [3].
  - Master keys: $s_1, s_2 \in \mathbb{Z}_q^*$; public keys: $P_0 = g_1^{s_1}, P_1 = g_2^{s_1}, \ P_2 = g_2^{s_2}$.
  - BGN cryptosystem [4]: an encryption scheme $E(\cdot)$, a decryption scheme $D(\cdot)$.
  - Registration for a "real" identity $RID_i \in \mathbb{G}_1$ and a password PW_i.

Phase II: Signing Key Generation

- Generate a pair of pseudo identity PID_i and secret key SK_i for registered o_i:

$$PID_i = \langle PID_i^1, PID_i^2 \rangle = \langle {g_1}^r, RID_i \odot {P_0}^r \rangle, \quad (1)$$
$$SK_i = \langle SK_i^1, SK_i^2 \rangle = \langle {PID_i^1}^{s_1}, H(PID_i^2)^{s_2} \rangle, \quad (2)$$

where r is a per-session random nonce, $\odot$ represents the Exclusive-OR (XOR) operation, and $H(\cdot)$ is a MapToPoint hash function [3], *i.e.*, $H(\cdot): \{0,1\}^* \to \mathbb{G}_1$.

Phase III: Data Submission

- Data Encryption: $\vec{D}_i = \left(E(u_{ij}), E({u_{ij}}^2)\right)|_{j \in [1,\beta]}$.
- Encrypted Data Signing:

$$\sigma_i = SK_i^1 \cdot {SK_i^2}^{h(D_i)}, \quad (3)$$

where "$\cdot$" denotes the group operation in $\mathbb{G}_1$, $h(\cdot)$ is a one-way hash function such as SHA-1 [6], and D_i is derived by concatenating all the elements of $\vec{D}_i$ together.

Phase IV: Data Processing and Verifications

- First-layer Batch Verification:

$$\hat{e}\left(\prod_{i=1}^{n} \sigma_i, g_2\right) = \hat{e}\left(\prod_{i=1}^{n} PID_i^1, P_1\right) \hat{e}\left(\prod_{i=1}^{n} H(PID_i^2)^{h(D_i)}, P_2\right). \quad (4)$$

- Data Submission by Data Consumer: $\vec{D}_0 = \left(E({v_j}^2), E(v_j)^{-2} = E(-2v_j)\right)|_{j \in [1,\beta]}$.
- Similarity Evaluation via Homomorphic Multiplication and Homomorphic Addition:

$$R_{ij} = E(1) \otimes E({v_j}^2) \oplus E(u_{ij}) \otimes E(-2v_j) \oplus E({u_{ij}}^2) \otimes E(1) = E\left((u_{ij} - v_j)^2\right),$$
$$R_i = R_{i1} \oplus R_{i2} \oplus \cdots \oplus R_{i\beta} = E\left(\sum_{j=1}^{\beta} (u_{ij} - v_j)^2\right) = E\left(f(\vec{U}_i, \vec{V})^2\right). \quad (5)$$

- Signatures Aggregation: $\sigma = \prod_{i=1}^{m} \sigma_{c_i}$, where $\{o_{c_1}, o_{c_2}, \cdots, o_{c_m}\}$ are matched ones.
- Second-layer Batch Verification:

$$\hat{e}(\sigma, g_2) = \hat{e}\left(\prod_{i=1}^{m} PID_{c_i}^1, P_1\right) \hat{e}\left(\prod_{i=1}^{m} H(PID_{c_i}^2)^{h(D_{c_i})}, P_2\right). \quad (6)$$

- Outcome Verification:
  - Real identity recovery: $PID_{c_i}^2 \odot {PID_{c_i}^1}^{s_1} = RID_{c_i} \odot {P_0}^r \odot {g_1}^{s_1 \cdot r} = RID_{c_i}$.
  - No need of homomorphic multiplications: $R_{ij} = E({u_{ij}}^2) \oplus E(u_{ij})^{-2v_j} \oplus E({v_j}^2)$.
  - Sampling, *e.g.*, p(not evaluating each profile)$= 20\%$, 26 checks, success rate $= 99.70\%$.

Evaluation Result

- Dataset:
  - R1-Yahoo! Music User Ratings of Musical Artists Version 1.0 [13]
  - 11,557,943 ratings of 98,211 artists given by 1,948,882 users
  - Choose β **common artists** as the evaluating attributes
  - Append each user's ratings ranging from 0 to 10
- Evaluation Settings:
  - Pairing-Based Cryptography (PBC) library [11]
  - Identity-based signature scheme
    - SS512: a supersingular curve with a base field size of 512 bits and an embedding degree of 2
    - MNT159: a MNT curve with a base field size of 159 bits and an embedding degree of 6
    - q is 160-bit long; all hashings are implemented in SHA1, considering its digest size closely matches q.
  - BGN cryptosystem: Type A1 pairing, in which the group order is a product of two 512-bit primes.
  - OS: 64-bit Ubuntu 14.04, Intel(R) Core(TM) i5 3.10GHz

Table I. Computation overhead of identity-based signature scheme per data contributor.

Setting	Preparation		Operation
	Pseudo Identity Generation	Secret Key Generation	Signing
SS512	4.698ms (39.40%)	6.023ms (50.53%)	1.201ms (10.07%)
MNT159	1.958ms (57.33%)	1.028ms (30.10%)	0.429ms (12.57%)

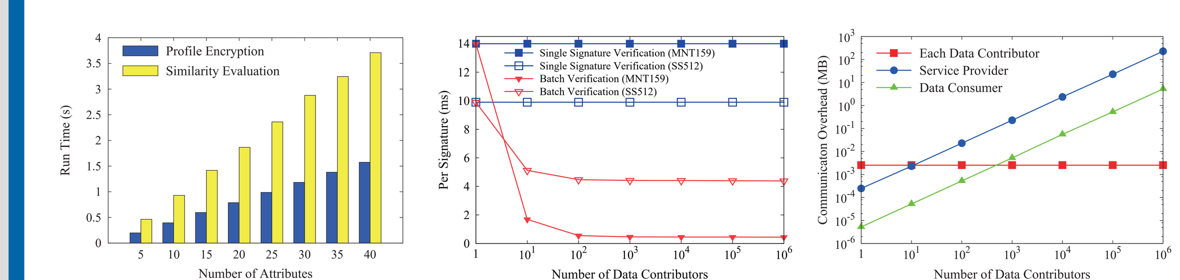


Figure 5. Performance of TPDM on Yahoo! Music Ratings Dataset.

Table I and Figure 5 reveal that TPDM incurs affordable computation and communication overheads, even when supporting as many as 1 million data contributors.

Conclusions

- Consider both data truthfulness and privacy preservation in data markets, and propose TPDM.
- Instantiate TPDM with a profile-matching service, and evaluate its performance on a real-world dataset.
- Evaluation results have demonstrated its scalability, especially from computation and communication overheads.

References

[1] "2016 TRUSTe/NCSA Consumer Privacy Infographic - US Edition," https://www.truste.com/resources/privacy-research/ncsa-consumer-privacy-index-us/.
[2] M. Balazinska, B. Howe, and D. Suciu, "Data markets in the cloud: An opportunity for the database community," in *VLDB*, 2011.
[3] D. Boneh and M. Franklin, "Identity-based encryption from the weil pairing," in *CRYPTO*, 2001.
[4] D. Boneh, E. Goh, and K. Nissim, "Evaluating 2-dnf formulas on ciphertexts," in *TCC*, 2005.
[5] "DataSift," http://datasift.com/.
[6] D. Eastlake and P. Jones, "US Secure Hash Algorithm 1 (SHA1)," *IETF RFC 3174*, 2001.
[7] D. Fiore, R. Gennaro, and V. Pastro, "Efficiently verifiable computation on encrypted data," in *CCS*, 2014.
[8] 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.
[9] "Gnip," https://gnip.com/.
[10] R. Ikeda, A. D. Sarma, and J. Widom, "Logical provenance in data-oriented workflows?" in *ICDE*, 2013.
[11] "PBC Library," https://crypto.stanford.edu/pbc/.
[12] P. Upadhyaya, M. Balazinska, and D. Suciu, "Automatic enforcement of data use policies with datalawyer," in *SIGMOD*, 2015.
[13] "Yahoo! Webscope datasets," http://webscope.sandbox.yahoo.com/.
[14] R. Zhang, Y. Zhang, J. Sun, and G. Yan, "Fine-grained private matching for proximity-based mobile social networking," in *INFOCOM*, 2012.

上海交通大学先进网络实验室
Advanced Network Laboratory