数据租赁——数据流通的新方式

阮雯强[1,2]，徐铭辛[1,2]，涂新宇[1,2]，宋鲁杉[1,2]，韩伟力[1,2]
1. 复旦大学数据分析与安全实验室，上海 200438；
2. 上海市数据科学重点实验室，上海 200438

摘要
数据正成为推动社会发展的新生产要素。以合规的、可审计的方式使数据在多方之间流通对于数据价值的形成至关重要。从隐私保护以及数据利用的角度，提出了一种新的数据流通方式——数据租赁。首先介绍了提出数据租赁的动机，然后明确了数据租赁应当满足的5项需求，最后提出了一种基于秘密共享的数据租赁技术。

关键词
数据流通；秘密共享；数据租赁；隐私保护

中图分类号：TP391 文献标志码：A doi: 10.11959/j.issn.2096-0271.2022071

Data tenancy: a new paradigm for data circulation

RUAN Wenqiang[1,2], XU Mingxin[1,2], TU Xinyu[1,2], SONG Lushan[1,2], HAN Weili[1,2]
1. Laboratory for Data Analytics and Security, Fudan University, Shanghai 200438, China
2. Shanghai Key Laboratory of Data Science, Shanghai 200438, China

Abstract
Data is becoming a new type of factor of production. How to compliantly and audibly circulate data among multiple parties is very important for data value formation. A novel data circulation paradigm, namely data tenancy, was proposed from the perspective of privacy preservation and data utilization. The motivation of data tenancy was discussed, and five requirements that data tenancy should satisfy were identified. Finally, a secret sharing-based data tenancy technique was proposed.

Key words
data circulation, secret sharing, data tenancy, privacy preservation

0 引言

数据已经与资本、土地、劳动力、技术等传统生产要素并列，成为一种新型的生产要素[1]。在数据价值的形成过程中，数据流通扮演着极为重要的角色。当前数据流通的方式主要包括政府部门或企业的数据公开、数据交易等。然而，随着《中华人民共和国网络安全法》（以下简称《网络安全法》）、《中华人民共和国数据安全法》（以下简称《数据安全法》）、《中华人民共和国个人信息保护法》（以下简称《个人信息保护法》）的发布，涉及用户隐私的数据难以直接在各个机构之间流通。此外，许多机构出于商业竞争的目的，可能并不愿意直接将原始数据传输给其他机构。目前得到较多关注的场景是如何使多个机构以隐私保护的方式对数据进行联合利用，即每个机构均贡献数据并且得到数据分析的结果，而如何使某个机构通过“租赁”的方式挖掘其他机构数据中蕴含的价值尚缺乏相应的研究。因此，为了促进数据价值的充分形成，本文提出了一种数据流通的新方式——数据租赁（data tenancy）。

数据租赁使数据租赁方能够通过付费的、隐私保护的以及可审计的方式，利用数据出租方的数据完成预先约定好的计算任务（如机器学习模型训练），并获得计算结果，即通过“租赁”数据获得价值。本文根据与隐私保护相关的法律法规，讨论了提出数据租赁的动机及其定义，并明确了数据租赁需要满足的5项需求。随后，本文提出了一项基于秘密共享的数据租赁技术，使分散在各个机构的数据能够通过“租赁”的方式更好地流通，从而促进数据价值的形成。

1 相关知识与已有研究

1.1 基于秘密共享的安全多方学习技术

安全多方学习即基于安全多方计算的隐私保护机器学习技术[2]。基于秘密共享的安全多方学习技术能够使多个参与方共同训练一个预先约定好的机器学习模型（训练过程由一个布尔电路或者算术电路表示），并保证不泄露除结果模型外的其他任何隐私信息[3-5]。如**图1**所示，其中，D_1、D_2、D_n分别表示参与方1、参与方2、参与方n的隐私数据集，在一个基于秘密共享的n方安全多方学习过程中，参与方i首先将其持有的隐私数据集（D_i）分解为n个秘密份额（$\langle D_i\rangle_1,\langle D_i\rangle_2,\cdots,\langle D_i\rangle_n$），随后将数据集的秘密份额分发给其他参与方。同时，在某些场景中，部分参与方可以不向其他参与方发送秘密份额，而只接收来自其他参与方的秘密份额。数据集的秘密份额分发完成后，所有参与方利用安全多方计算协议共同生成一个随机化的初始模型参数，随后进入一个基于秘密共享的安全多方计算过程，通过本地计算与交互通信，利用数据的秘密份额完成对目标模型的训练，最终每个参与方各自得到一份目标模型的秘密份额。随后，根据具体的场景，参与方可以选择不还原目标模型，但在对数据进行推理时仍然通过交互完成，或者通过交换各自持有的秘密份额，将目标模型还原为明文。目前较为流行的用于安全多方学习的秘密共享技术有两种：加法秘密共享和Shamir秘密共享。其中，加法秘密共享可以支持两方及以上的参与方数量，Shamir秘密共享则支持三方及以上的参与方数量。

基于秘密共享的安全多方学习技术具有以下4个特性：①所有参与方只能得到结果模型，而得不到其他参与方输入的任何信息；②所有参与方共同训练一个预先约定好的、训练过程能够用电路（算术电路或布尔电路）表示的目标模型；③所有参与方都需要参与训练过程；④结果模型可以由所有参与方持有，也可以只由一个或部分参与方持有，即所有参与方将所持有的结果模型的秘密份额发送给有权恢复最终结果模型的参与方。在获得其他参与方的秘密份额后，有权恢复最终结果模型的参与方将还原出最终的结果模型。

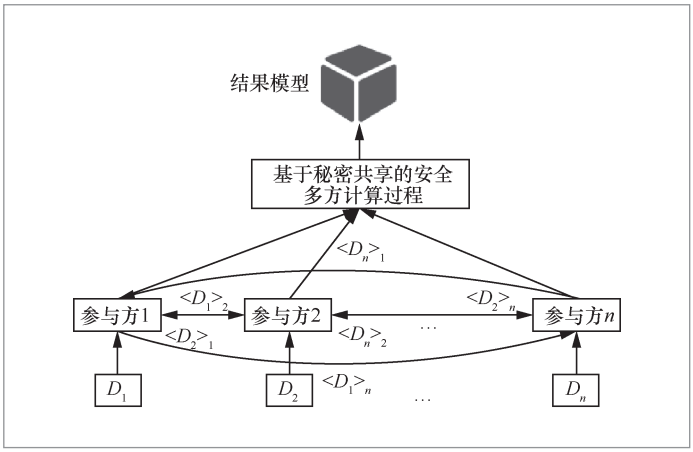

图1 基于秘密共享的安全多方学习过程示例

1.2 安全模型

本文提出的数据租赁技术采用半诚实的安全模型，即每一个参与方均会根据协议规定的步骤进行计算，并向其他参与方发送预先定义好的信息，但参与方会尽量从收到的信息中推断其他参与方的输入信息。由于当前参与方之间使用安全多方学习技术的目的是满足隐私保护法律法规对数据流通的各项要求，在参与方均有共享数据的意愿的前提下，半诚实模型是一个适用于实际场景的安全模型。

1.3 相关研究工作

随着世界各国纷纷发布与个人信息保护相关的法律法规，如欧盟于2018年发布了《通用数据保护条例》、我国于2021年发布了《个人信息保护法》等，涉及用户隐私的数据流通受到了极大的限制。近年来，为了在合规的前提下充分挖掘来自不同机构的数据中潜藏的价值，研究者提出并实现了许多个隐私计算算法与系统，使多个数据出租方能够以隐私保护的方式对分散在各方的数据进行联合建模与分析，实现“数据可用不可见”的目标。当前受到较多关注的隐私计算技术包括安全多方学习技术[6-12]、联邦学习[13-15]等。

Mohassel P等人[6]于2017年提出了第一个支持神经网络模型训练的安全多方学习系统——SecureML。随后研究者提出并实现了许多安全多方学习系统，包括支持更多参与方且更加高效的ABY3[7]、Fantastic-Four[8]等，支持恶意参与方模型的SWIFT[9]、BLAZE[10]等，支持复杂模型训练与推理的CryptGPU[11]、Falcon[12]等。在这些已有的安全多方学习系统中，每个参与方的身份都是对等的，都需要提供数据并且都能在计算完成后得到计算结果。一个机构以隐私保护的、可审计的“租赁”方式对其他机构的数据进行分析的框架和机制尚需要进一步研究。

此外，Google于2015年提出了联邦学习的概念[13]。随后，许多企业推出了基于联邦学习的联合建模系统，例如Google发布的TensorFlow Federated、微众银行推出的FATE（federated AI technology enabler）等。相较于安全多方学习系统，基于联邦学习的系统具有更高的效率，但

是也有更高的隐私风险，例如，参与方之间传输的中间结果很有可能泄露输入数据的相关隐私信息[16-18]。同时，当前并没有一个数学模型对联邦学习系统的隐私风险进行量化分析。此外，基于联邦学习的系统对各方的数据进行联合建模可能对得到的模型精度造成一定的损失，特别是当各方的数据为非独立同分布时，联邦学习会造成较大的精度损失[14]。

2 数据租赁概述

2.1 数据租赁的动机

当前数据流通的主要方式是不同机构之间进行数据交易，即数据买家通过支付一定的费用从数据卖家的手中获得数据。向数据卖家支付一定的费用后，数据买家可以直接得到数据，并对其开展任意的分析操作。目前国内已经产生了许多数据交易平台。尽管数据交易对于促进数据流通发挥着重大的作用，但是它仍然存在两个限制，使得数据在一些场景中无法充分流通，具体如下。

- 需要流通的数据可能包含用户的隐私信息，随着《网络安全法》《数据安全法》以及《个人信息保护法》的陆续出台，直接转让或传输这些数据可能会给售卖数据的机构带来严重的法律风险。
- 出于商业竞争等目的，持有数据的机构或个人可能并不希望直接将数据发送给其他机构，但可以允许其他机构对其所有的数据进行部分特定的、敏感程度较低的计算操作。

当数据较为敏感，无法直接在机构之间进行流通时，数据租赁可以使用一种隐私保护的、可审计的方式，使数据租赁方能够利用数据出租方的数据完成特定的计算任务，从而促进数据价值充分形成。

2.2 数据租赁的定义

参考传统的资产租赁的定义，并考虑数据资产特有的形态以及当前已经发布的各项隐私保护法律，本文对数据租赁的定义如下：数据租赁是指在约定的时间内，数据出租方使用其持有的数据资产完成数据租赁方要求的特定计算任务，最终数据租赁方只获得计算结果、数据出租方获取租金的行为。

由于数据的复制成本几乎为零，并且涉及用户的隐私信息，受到法律保护，当把数据作为租赁标的时，数据出租方无法像传统的资产租赁那样在一段时间内将数据资产直接转让给数据租赁方，只能通过完成数据租赁方指定的计算任务这种方式，获得租赁数据带来的收益。

此外，相较于定义为“让在不同地方使用不同计算机、不同软件的用户能够读取他人数据并进行各种操作、运算和分析”的数据共享，数据租赁有以下3点不同：①数据出租方的数据无法被数据租赁方直接读取，数据租赁方仅能获取计算任务的输出结果；②数据出租方能够根据数据租赁方的计算任务对租金进行定价；③数据出租方和数据租赁方均要对计算过程进行监督，确保数据租赁交易按照事先约定的流程进行。综上所述，相较于数据共享，数据租赁带来了更多的要求，这些要求为实现数据租赁带来了更多、更大的技术挑战。

2.3 数据租赁的特征

根据数据租赁的定义，当设计一种数

据租赁框架时，应当使其能够满足以下5项需求。

- 可计价：根据使用目标计算任务的复杂程度以及使用数据的次数等，能够计算数据租赁方应当支付给数据出租方的租赁费用。
- 隐私性：数据出租方不直接将明文数据传输给其他机构。为了规避潜在的法律风险，数据出租方的数据应当保留在其本地，以防用户隐私信息泄露。
- 有效性：数据租赁方能够利用数据出租方的数据与数据出租方共同完成双方事先约定好的计算任务，并且得到计算结果。在数据租赁的计算过程中，数据租赁方自身的数据也可能参与计算。值得注意的是，可能会有多个数据出租方同时向一个机构租赁数据以完成其目标计算任务。
- 计算过程可监督：数据出租方和数据租赁方应当都能对计算操作进行监督，即数据出租方和数据租赁方都应该能够确保对方对数据执行预先约定好的计算操作。通过确保计算过程的可监督性，数据出租方能够根据计算操作的类型和数量收取相应的租赁费用，而数据租赁方能够确保其能利用其他机构的数据完成特定的计算任务。
- 可审计：数据出租方和数据租赁方对数据所做的计算操作应当能够被第三方审计，从而避免计算任务完成后，双方对于已完成的计算操作的类型和数量无法达成一致意见，导致支付租金时双方发生纠纷。

3 基于秘密共享的数据租赁技术设计

尽管其他隐私计算技术（如联邦学习等）能够实现一定程度的隐私保护，然而，这些技术对于自身提供的隐私保护缺乏理论保障，而安全多方学习使用安全多方计算技术完成底层运算，能够为计算过程提供严格的安全保障。因此，本文提出一种基于秘密共享的数据租赁技术，令数据出租方、数据租赁方共同参与一个基于秘密共享的安全多方学习过程，以完成数据出租方和数据租赁方预先约定好的计算任务。接下来对本文提出的数据租赁技术涉及的角色以及计算过程进行详细的介绍，并分析该技术如何满足隐私性、有效性、计算过程可监督以及可审计这4项需求。对于可计价需求，由于其与后续的计算过程是解耦的，且当前已经有许多与数据定价相关的研究工作[19-20]，如基于博弈论的方法[20]，本文对如何满足此项需求不做过多讨论。相较于已有的基于同态加密[21]的数据安全外包计算方法，本文提出的基于秘密共享的数据租赁技术使数据出租方和数据租赁方能够通过参与计算过程的方式监督对方所做的计算操作。此外，通过引入区块链技术，本文提出的数据租赁技术使第三方能够在交易完成后对交易信息进行审计，可避免出现数据出租方或数据租赁方抵赖的情况。

3.1 角色定义

本文提出的基于秘密共享的数据租赁技术中（如**图2**所示）共有3类角色，即数据租赁方、数据出租方以及租赁平台方，具体如下。

- 数据租赁方。数据租赁方自身可能拥有一部分数据，并希望能够通过支付一笔费用租赁数据出租方的数据，从而通过联合多方的数据挖掘得到更多的有效信息。数据租赁方需要向数据出租方和租赁平台方说明其目标计算任务，并通过基于秘密共享的安全多方学习完成该计算任务。
- 数据出租方。数据出租方对数据租赁方出租其需要的数据，并根据数据租赁

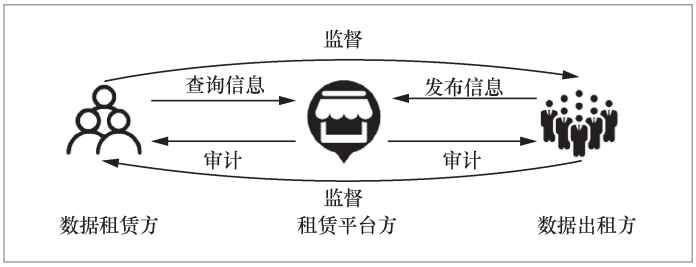

图 2 基于秘密共享的数据租赁技术中的 3 类角色

方利用其数据完成的计算任务的复杂程度和使用数据的次数收取相应的费用。在一次数据租赁中，可能会有多个数据出租方参与。数据出租方通过与数据租赁方共同参与一个基于秘密共享的安全多方学习过程，完成数据租赁方的目标计算任务以及监督数据租赁方对其数据所做的计算操作。

- 租赁平台方。租赁平台方负责提供数据租赁的资讯平台，并审计数据租赁交易。租赁平台方接收并发布来自数据出租方的数据信息，同时响应数据租赁方的数据信息查询请求，促使数据租赁交易的形成。

3.2 学习过程

在数据租赁方和数据出租方对租赁的数据类型和数量、目标计算任务以及租赁费用达成共识后，数据租赁方、数据出租方共同参与一个基于秘密共享的安全多方学习过程，以完成数据租赁交易，具体过程如**图3**所示。在**图3**所示的计算过程中，各方先将自身持有的数据通过秘密共享技术产生秘密份额，然后将秘密份额分发给其他参与方作为输入，随后各方通过一个基于秘密共享的安全多方学习过程完成目标计算任务，最后将计算结果返回数据租赁方。

具体来说，数据租赁方首先将其目标计算任务转化为电路（由与门、或门、非门组成的布尔电路或由乘法门、加法门组成的算术电路）表示，随后将该电路发送给其他参与方作为后续计算过程的输入。同时，数据租赁方需要计算目标电路的数字摘要并将其上传到区块链，使得数据租赁交易完

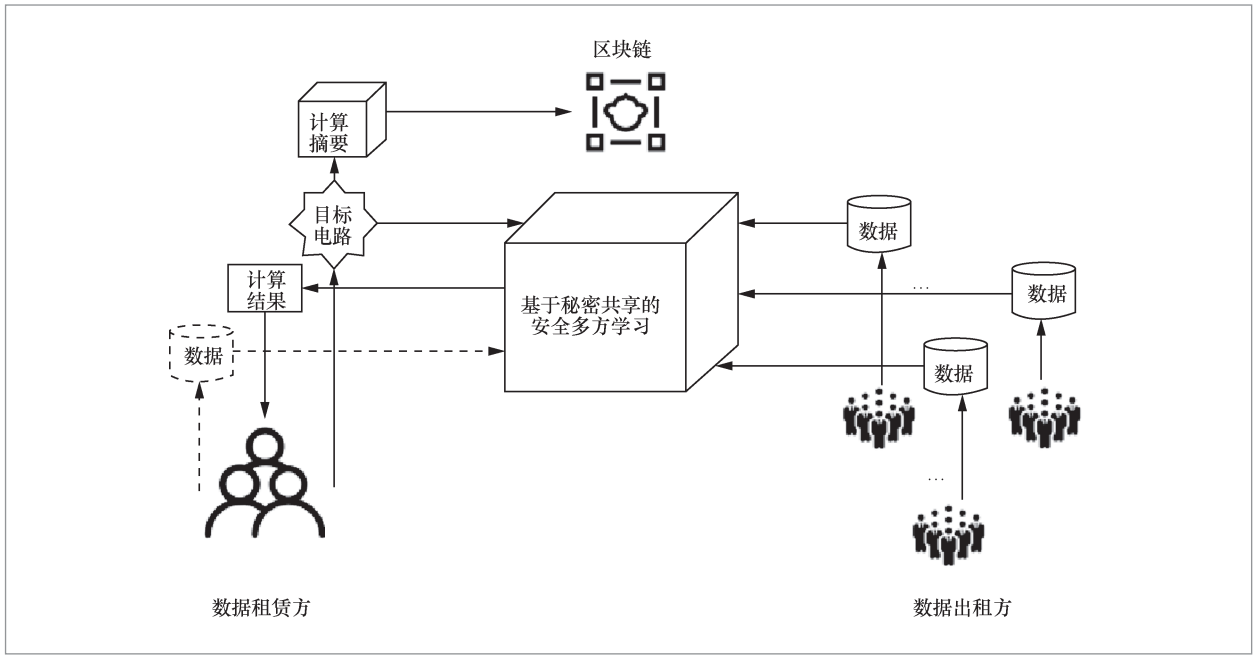

图 3 基于秘密共享的数据租赁技术计算过程

成后第三方能够根据链上的数据对该交易进行审计。倘若数据租赁方自身的数据需要参与计算任务，则将自身数据使用秘密共享技术产生秘密份额后，将相应的秘密份额分发给其他参与方。而数据出租方将自身数据使用秘密共享技术产生秘密份额后，将相应的秘密份额分发给其他参与方作为后续计算过程的输入，完成数据的“出租”。数据出租方与数据租赁方得到输入数据的秘密份额以及计算任务的电路表示后，利用基于秘密共享的安全多方学习技术通过本地计算和通信交互利用自身的秘密份额对目标电路进行计算，该电路的输入即各方持有的秘密份额。在计算目标电路时，各方首先根据门电路之间的依赖关系将目标电路拆解为多个电路层，每个电路层的输入都来自前一个电路层，输出都传向下一个电路层。随后，逐层计算目标电路，即依次对每一层包含的门电路进行计算，最后一个电路层的输出即计算结果的秘密份额。其中，非门与加法门可以在本地完成计算，与门、或门以及乘法门则需要通过各方间的交互完成计算。最后，数据出租方将各自持有的计算结果的秘密份额发送给数据租赁方，数据租赁方使用收到的秘密份额还原得到计算结果，并向数据出租方支付相应的租金，完成数据租赁交易。

3.3 分析

接下来对计算过程进行分析，说明其能够满足数据租赁技术应当满足的隐私性、有效性、计算过程可监督以及可审计这4项需求。

- 隐私性。数据租赁方与数据出租方的数据均使用秘密共享技术产生秘密份额后，将秘密份额分发给其他参与方，并且后续所有的计算都是使用基于秘密共享的安全多方学习技术完成的。根据基于秘密共享的安全多方学习的特性，所有参与方都无法在计算过程中获得其他参与方的数据信息，从而保障了数据出租方数据的隐私性。
- 有效性。基于秘密共享的安全多方学习技术能够支持多个参与方共同计算，使得数据租赁方与数据出租方能够基于多方的输入数据共同完成事先约定的计算任务。最终，数据租赁方得到计算结果，保障了数据租赁交易的有效性。
- 计算过程可监督。基于秘密共享技术的安全多方学习要求所有参与方在计算过程中都知晓计算任务对应的电路，并参与计算。因此，在上述计算过程中，所有计算都需要数据出租方和数据租赁方共同参与，从而数据租赁方与数据出租方能够监督对方所做的计算操作。
- 可审计。如**图3**所示，在计算开始前，数据租赁方将目标电路的摘要上传到区块链。在计算完成后，第三方（如租赁平台方）可以通过查验区块链上的数据摘要的方式对已完成的数据租赁交易进行审计。

4 结束语

基于当前已发布的隐私保护法律法规，本文提出了一种新的数据流通方式——数据租赁，分析了数据租赁应该满足的5项需求，并提出了一种基于秘密共享的数据租赁技术，旨在进一步促进数据的流通与数据价值的形成。在未来，如何使数据租赁方在租赁开始前对数据出租方的数据进行检验或将成为数据租赁技术下一步的发展方向，需要研究者进行更加深入的探索与研究。

参考文献:

[1] XU X. Research prospect: data factor of production[J]. Journal of Internet and Digital Economics, 2021, 1(1): 64–71.
[2] SONG L S, WU H Q, RUAN W Q, et al.

SoK: training machine learning models over multiple sources with privacy preservation[J]. arXiv preprint, 2020, arXiv:2012.03386.
[3] CRAMER R, DAMGÅRD I B, NIELSEN J B. Secure multiparty computation and secret sharing[M]. Cambridge: Cambridge University Press, 2015.
[4] GOLDREICH O, MICALI S, WIGDERSON A. How to play ANY mental game[C]// Proceedings of the 19th Annual ACM Conference on Theory of Computing. New York: ACM Press, 1987: 218–229.
[5] BEN-OR M, GOLDWASSER S, WIGDERSON A. Completeness theorems for non-cryptographic fault-tolerant distributed computation[C]//Proceedings of the 20th Annual ACM Symposium on Theory of Computing. New York: ACM Press, 1988: 1–10.
[6] MOHASSEL P, ZHANG Y P. SecureML: a system for scalable privacy-preserving machine learning[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2017: 19–38.
[7] MOHASSEL P, RINDAL P. ABY^3: a mixed protocol framework for machine learning[C]//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2018: 35–52.
[8] DALSKOV A P K, ESCUDERO D, KELLER M. Fantastic four: honest-majority four-party secure computation with malicious security[C]//Proceedings of 30th USENIX Security Symposium. [S.l.:s.n.], 2021: 2183–2200.
[9] KOTI N, PANCHOLI M, PATRA A, et al. SWIFT: super-fast and robust privacy-preserving machine learning[C]// Proceedings of 30th USENIX Security Symposium. [S.l.:s.n.], 2021: 2651–2668.
[10] PATRA A, SURESH A. BLAZE: blazing fast privacy-preserving machine learning[C]// Proceedings of 2020 Network and Distributed System Security Symposium. Reston: Internet Society, 2020.
[11] TAN S J, KNOTT B, TIAN Y, et al. CryptGPU: fast privacy-preserving machine learning on the GPU[C]// Proceedings of 2021 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2021: 1021–1038.
[12] WAGH S, TOPLE S, BENHAMOUDA F, et al. Falcon: honest-majority maliciously secure framework for private deep learning[J]. Proceedings on Privacy Enhancing Technologies, 2021, 2021(1): 188–208.
[13] LI T, SAHU A K, TALWALKAR A, et al. Federated learning: challenges, methods, and future directions[J]. IEEE Signal Processing Magazine, 2020, 37(3): 50–60.
[14] YANG Q, LIU Y, CHEN T J, et al. Federated machine learning[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1–19.
[15] 王健宗, 孔令炜, 黄章成, 等. 联邦学习隐私保护研究进展[J]. 大数据, 2021, 7(3): 130–149
WANG J Z, KONG L W, HUANG Z C, et al. Research advances on privacy protection of federated learning[J]. Big Data Research, 2021, 7(3): 130–149.
[16] 周传鑫, 孙奕, 汪德刚, 等. 联邦学习研究综述[J]. 网络与信息安全学报, 2021, 7(5): 77–92.
ZHOU C X, SUN Y, WANG D G, et al. Survey of federated learning research[J]. Chinese Journal of Network and Information Security, 2021, 7(5): 77–92.
[17] ZHU L G, LIU Z J, HAN S. Deep leakage from gradients[J]. Advances in Neural Information Processing Systems, 2019, 32.
[18] JERE M S, FARNAN T, KOUSHANFAR F. A taxonomy of attacks on federated learning[J]. IEEE Security & Privacy, 2021, 19(2): 20–28.
[19] NIU C Y, ZHENG Z Z, WU F, et al. Online pricing with reserve price constraint for personal data markets[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(4): 1928–1943.
[20] ZHENG Z J, SONG L Y, HAN Z. Bridging the gap between big data and game theory: a general hierarchical pricing framework[C]//Proceedings of 2017 IEEE International Conference on Communications. Piscataway: IEEE Press, 2017: 1–6.
[21] GENTRY C. Fully homomorphic encryption using ideal lattices[C]// Proceedings of the 41st Annual ACM Symposium on Theory of Computing. New York: ACM Press, 2009: 169–178.

作者简介

阮雯强（1999– ），男，复旦大学计算机科学技术学院博士生，主要研究方向为基于安全多方计算的隐私保护机器学习、差分隐私等。

徐铭辛（1997– ），男，复旦大学软件学院硕士生，主要研究方向为基于安全多方计算的隐私保护机器学习、差分隐私等。

涂新宇（1999– ），男，复旦大学软件学院硕士生，主要研究方向为基于安全多方计算的隐私保护机器学习、秘密共享等。

宋鲁杉（1999– ），女，复旦大学计算机科学技术学院博士生，主要研究方向为基于安全多方计算的隐私保护、机器学习等。

韩伟力（1975– ），男，博士，复旦大学计算机科学技术学院教授，主要研究方向为数据安全、访问控制。

收稿日期： 2022-05-09

通信作者： 韩伟力，wlhan@fudan.edu.cn

基金项目： 国家重点研发计划资助项目（No.2019YFE0103800）；上海市科委“创新行动计划”项目（No.21511101600）

Foundation Items: The National Key Research and Development Program of China (No.2019YFE0103800), Science and Technology Innovation Action Plan of Shanghai (No.21511101600)