联邦学习隐私保护研究进展

王健宗，孔令炜，黄章成，陈霖捷，刘懿，卢春曦，肖京
平安科技（深圳）有限公司，广东 深圳 518063

摘要
针对隐私保护的法律法规相继出台，数据孤岛现象已成为阻碍大数据和人工智能技术发展的主要瓶颈。联邦学习作为隐私计算的重要技术被广泛关注。从联邦学习的历史发展、概念、架构分类角度，阐述了联邦学习的技术优势，同时分析了联邦学习系统的各种攻击方式及其分类，讨论了不同联邦学习加密算法的差异。总结了联邦学习隐私保护和安全机制领域的研究，并提出了挑战和展望。

关键词
联邦学习；联邦学习系统攻击；隐私保护；加密算法

中图分类号：TP311 文献标识码：A doi: 10.11959/j.issn.2096-0271.2020030

Research advances on privacy protection of federated learning

WANG Jianzong, KONG Lingwei, HUANG Zhangcheng, CHEN Linjie, LIU Yi, LU Chunxi, XIAO Jing
Ping An Technology (Shenzhen) Co. Ltd., Shenzhen 518063, China

Abstract
To this end, many laws and regulations on privacy protection have been introduced, and the phenomenon of data-island has become a major bottleneck hindering the development of big data and artificial intelligence technology. Federated learning has received widespread attention to break this phenomenon. Started with the historical development of federated learning, the definition, and architecture and classification of federated learning, the advantages of federated learning in privacy protection domainwere introduced. At the same time, various attack methods and their classification aboutfederated learning were introduced in detail. The classification of various encryption algorithms in federated learning were summarized. In conclusion, the contribution of federated learning in privacy protection and security mechanism were summarized and the new challenges in these domains were proposed.

Key words
federated learning, federated learning system attack, privacy protection, encryption algorithm

1 引言

大数据、人工智能和云产业等的爆发式发展，一方面为传统行业带来升级变革的新机遇，另一方面也给数据和网络安全带来了新挑战。不同行业的公司会收集大量的数据信息，同一企业下不同层级的部门也会收集不同的信息，由于行业间的竞争和垄断，以及同一企业下不同系统和业务的闭塞性与阻隔性，很难实现数据信息的交流与整合。当不同的企业之间，以及同一企业下属不同部门之间需要合作进行联合建模时，将面临跨越重重数据壁垒的考验。这类挑战也被称为数据孤岛问题。

早期的分布式计算试图通过整合不同来源的数据进行分布式的建模，从而解决这类数据孤岛问题。分布式建模将具有庞大计算量的任务部署到多台机器上，提升了计算效率，减少了任务耗能。但是分布式机器学习依旧存在问题，重量级的分布式系统架构通常会产生巨大的沟通成本，影响数据的传输和处理效率。随着人工智能技术的进一步发展和更广泛的应用，数据隐私敏感性问题日益被重视。大规模的数据传输不可避免地会涉及隐私泄露问题，对于异构数据的联合训练和隐私安全问题，依然没有找到一个令人满意的解决方案。

联邦学习（federated learning，FL）给上述难题提供了解决方案。联邦学习是由谷歌公司[1]在2016年率先提出的概念，该技术在数据不共享的情况下完成联合建模共享模型。具体来讲，各个数据持有方（个人/企业/机构）的自有数据不出本地，通过联邦系统中加密机制下的模型参数交换方式（即在不违反数据隐私法规的情况下），联合建立一个全局的共享模型，建好的模型为所有参与方共享使用。相对于分布式计算，联邦学习有更多的优势，例如在隐私保护领域，联邦学习从算法层面上设计并考虑了客户端间传输信息的加密。本文主要从隐私保护和安全加密的角度，对联邦学习进行系统综述。

本文的主要贡献如下。

- 本文对联邦学习的历史进行了详细的叙述，从安全隐私的分布式学习发展到现在的联邦学习系统，总结了联邦学习发展的历程。
- 本文从新的角度阐述了联邦学习的类型。与传统方式不同，本文从面向企业（to business，ToB）和面向客户（to customer，ToC）的应用场景的区别出发，分析了联邦学习的不同。
- 详细地从联邦学习攻击的角度分析联邦系统面临的各种可能的攻击手段，并系统地将联邦学习的攻击手段进行了分类总结。
- 联邦学习的加密机制在一定程度上可以抵御一些联邦学习攻击，或者大大增加攻击的难度。本文从加密算法的角度详细讨论了联邦学习的加密机制。

2 联邦学习概述

2.1 联邦学习的历史

随着人工智能的发展[2-3]，联邦学习可以追溯到分布式学习的诞生，其学习的模式与分布式学习相似，但又有很多不同之处，主要表现在中心控制权、节点稳定性、通信代价、数据分布和数据量级上。联邦学习概念的正式确立得益于谷歌的推动，谷歌在2016年提出了联邦学习[4]概念，联邦学习这一名词开始频繁地出现。

前述的联邦学习与分布式学习的差异

最终体现在对隐私保护的要求上。分布式计算为人工智能和大数据的结合提供了算力基础，保证了大规模的数据能够被有效地使用和学习。但随着社会的发展，无论是机构还是个人对隐私保护的要求越来越严格[5]。不同的机构甚至因为个人隐私问题不愿意共享自己的数据，大数据时代面临着前所未有的挑战。此时，大数据时代迫切地需要一种能够提供隐私保护的技术来弥补分布式学习的不足。如果隐私保护技术能够做到在使用数据联合训练的同时，任何一方都无法知晓他人的数据，就可以激励更多的机构和个人提供数据，促进相关领域发展[6]。联邦学习就是在这一背景下孕育而生的，其特性就是能够保证各参与方在不共享数据的前提下，进行隐私保护下的联邦建模。联邦学习的隐私保护技术不仅体现在机器学习建模上，对于区块链场景下存在的安全问题，也已经有学者考虑用联邦学习的方法来解决[7]。

2.2 联邦学习的定义和分类

对联邦学习的定义和分类有很多，目前还没有统一的标准。在学术上被广泛认可的是由Yang Q等人[8]提出的相关定义和分类。下面简述联邦学习的相关定义及分类。

2.2.1 联邦学习的定义

对于一次联邦学习建模任务，设有N个数据拥有方参与（以下简称参与方）此次建模任务，定义参与方拥有的数据集为$\{D_1,\cdots,D_N\}$。联邦学习的做法不再是将其简单地聚合起来形成一个新的数据集，从而完成下一阶段的训练任务。设在一次联邦建模任务完成后的全局模型为$\mathcal{M}_{\text{Fed}}$，对应的聚合后训练所得模型为$\mathcal{M}_{\text{Sum}}$。一般而言，全局模型$\mathcal{M}_{\text{Fed}}$由于存在参数交换和聚合的操作，在整个训练过程中会出现精度损失，即全局模型$\mathcal{M}_{\text{Fed}}$的表现不如聚合模型$\mathcal{M}_{\text{Sum}}$的表现。为量化这一差异，定义全局模型$\mathcal{M}_{\text{Fed}}$在测试集上的表现为$\mathcal{V}_{\text{Fed}}$，聚合模型$\mathcal{M}_{\text{Sum}}$在测试集上的表现为$\mathcal{V}_{\text{Sum}}$。此时定义模型的δ-精度损失为：

$$\left|\mathcal{V}_{\text{Fed}}-\mathcal{V}_{\text{Sum}}\right|<\delta \tag{1}$$

其中，δ为非负数。但在实际情况下，最终无法获取聚合模型$\mathcal{M}_{\text{Sum}}$，因为联邦学习的基本要求是隐私保护。

2.2.2 联邦学习的分类

对联邦学习的分类，广为人知的是由Yang Q等人[8]提出的横向联邦学习、纵向联邦学习以及联邦迁移学习。这种分类方式是从用户维度和特征维度的重叠情况考虑的。但在实际的生产中，更多的是依据业务场景考虑实际的分类情况。在业务上经常提及的是B端业务和C端业务，对应的联邦学习的分类也与这种业务分类方式有关，定义联邦学习的分类为ToB和ToC两大场景。

对于ToB场景的联邦学习来说，其主要服务对象为机构、公司和政府等。在这种联邦学习的场景下，参与方之间通过新增一个信任第三方作为中心服务器，协作各参与方完成联邦学习的过程，同时可以保证中间传输内容的可审计性。通常中心服务器的作用是控制参数交换、中间计算以及训练流程。

对于ToC场景的联邦学习来说，联邦建模的参与方主要以边缘端计算设备为主，通常这类联邦学习的参与方数量较多、算力较低。针对这样的场景，若仍保留中心服务器，其作为流程控制节点的特性会被弱化，往往联邦模型更新的功能会被集成在每一个参与方的计算节点上。ToC

中的参与方通过获取联合建模的模型，达到提升本地模型的效果的目的。

2.3 联邦学习架构

本节将介绍联邦学习系统的基础架构。由于不同的联邦学习任务具有不同的学习场景，因此联邦学习架构的设计也是不同的。从这些复杂架构中，笔者总结出以下两种基础的架构模式。一种是服务器客户端架构，另一种是端对端架构。根据联邦学习应用场景的复杂度、安全需求，笔者将采用不同的架构。同时，当应用场景特别复杂时，笔者可以根据需求将这两种基础的联邦学习架构进行拼接组合，从而形成一种混合的联邦学习架构。

在很多横向联邦学习应用场景中，参与训练的参与方数据具有类似的数据结构（特征空间），但是每个参与方拥有的用户是不相同的。有时参与方比较少，例如，银行系统在不同地区的两个分行需要实现联邦学习的联合模型训练；有时参与方会非常多，例如，做一个基于手机模型的智能系统，每一个手机的拥有者将会是一个独立的参与方。针对这类联合建模需求，可以通过一种基于服务器客户端的架构来满足很多横向联邦学习的需求，如图1所示。将每一个参与方看作一个客户端，然后引入一个大家信任的服务器来帮助完成联邦学习的联合建模需求。在联合训练的过程中，被训练的数据将会被保存在每一个客户端本地，同时，所有的客户端可以一起参与训练一个共享的全局模型，最终所有的客户端可以一起享用联合训练完成的全局模型。如图1所示，云服务器作为中心的服务器进行联合训练模型参数的聚合，每一个参与方作为客户端通过与服务器之间进行参数传递来参与联合训练。服务器客户端架构的联合训练的过程如下。

步骤1：中心服务器初始化联合训练模型，并且将初始参数传递给每一个客户端。

步骤2：客户端用本地数据和收到的初始化模型参数进行模型训练。具体步骤包括：计算训练梯度，使用加密、差异隐私等加密技术掩饰所选梯度，并将加密后的结果发送到服务器。

步骤3：服务器执行安全聚合。服务器只收到加密的模型参数，不会了解任何客户端的数据信息，实现隐私保护。服务器将安全聚合后的结果发送给客户端。

步骤4：参与方用解密的梯度信息更新

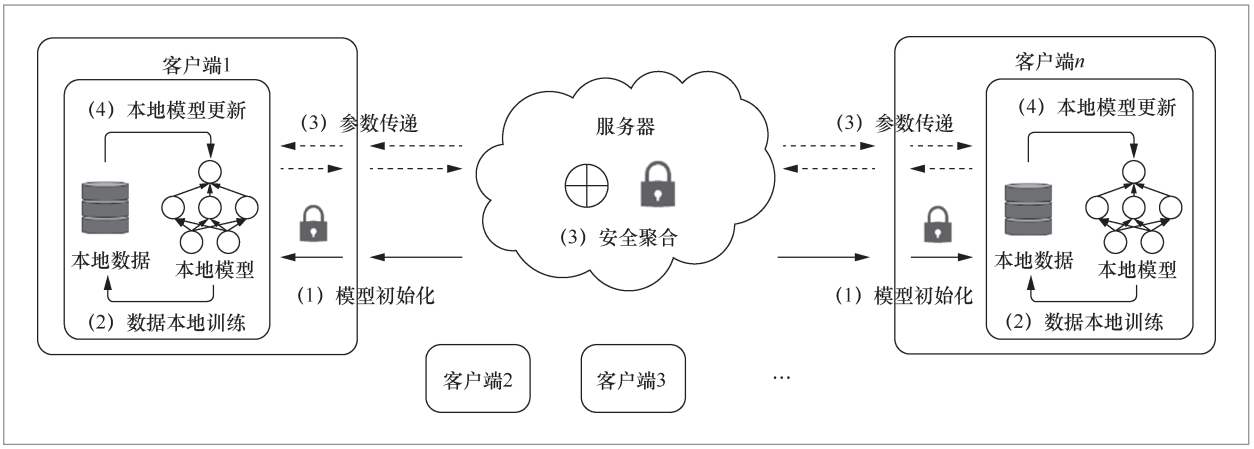

图1 服务器客户端架构

各自的本地模型，具体方法重复步骤2。

这个架构可以保证所有的参与方不会泄露个人的信息，也不会泄露信息给服务器，服务器只负责安全聚合加密的模型参数，并且发送给所有的客户端，保证了大家可以共享联合训练的模型。重复这个过程，直到损失函数可以收敛，就可以完成整个联合模型的训练。

在一些纵向联邦学习的应用场景中，通常参与训练的参与方有很多重叠的用户，但是关于用户的数据结构是不相同的。例如，在同一个城市的银行系统和电商系统，它们的用户群体大部分是本地住户，会有大量的用户重叠，但是两个公司收集到的用户信息是不相同的。双方是不能共享两边的用户信息的，如果需要联合训练一个模型，将会有很大的难度，需要将相同用户的不同特征在加密的状态下进行聚合，从而在增强模型能力的同时保证用户数据隐私。有的情况下，由于安全信任的缘故，公司双方可能不会达成第三方服务器的共识。因此，基于端对端（peer-to-peer）的联邦学习框架被提出。**图2**描述了端对端联邦学习架构。每一个参与方可以通过广播将自己的训练参数传递给其他所有的参与方，或者通过循环传递链往下一个参与方传递参数。端对端的联合训练过程步骤如下。

步骤1：参与方使用本地数据和初始化模型参数进行模型训练。

步骤2：参与方加密传递参数。具体步骤包括：计算训练梯度，使用加密、多方安全计算等方法将加密后的参数结果广播给其他所有的参与方。如果是链式传递模型，就只将模型参数传递给链式下端的参与方。

步骤3：参与方收到其他所有的加密模型参数之后进行安全聚合。

步骤4：解密并继续进行模型训练，更新本地模型。如果是链式传递模型，则接收到链式上端的参与方模型参数数据后进行安全聚合，解密后继续进行模型训练。重复步骤2。

在处理现实中更加复杂的联邦学习应用时，一种单独的框架可能不足以满足所有的需求。因此，需要将两种框架融合起来，形成混合框架，例如在包含中央节点的服务器客户端架构中，一个客户端集群下包含着一个端到端的联邦学习架构子集，以满足现实中的具体应用。

2.4 联邦学习的隐私保护和安全需求

在当前大数据的大环境下，人工智能算法在以机器学习、深度学习等基础算法为基础，为人们生活提供便捷的同时，也面临新的挑战。用户在享受人工智能带来的服务的同时，也越来越注重个人隐私的保护。同时，政府机构也出台了越来越严格的法规来保护机构之间的数据安全和隐私。因此，研究如何在保障安全以及隐私的前提下继续提供优质人工智能服务的大数据架构，成为新时代人工智能研究的新趋势。以智能零售为例，系统需要将用户的银行信息、社交网络信息、电子商城信息结合在一起，组成一个更优质的客户个性化产品推荐服务，但是不同的企业之间不能够暴露各自用户的隐私信息，仅通过传统的机器学习是无法在如此严格的数据障碍下，完成这样的智能零售服务的。联邦学习就是为了解决类似的行业难题而建立起来的关键技术。

在解决用户隐私问题上，联邦学习相对于传统机器学习具有多种优势。首先，联邦学习实现了数据的隔离，客户数据始终被保存在本地，从而满足了用户隐私保护和数据安全的需求。在保证所有参与方数据独立的前提下，联邦学习的模型训练

主要通过信息与模型参数的加密交换完成一个联合模型，为所有人提供服务，在保护隐私的前提下促进了参与方之间的公平合作和共赢。其次，联邦学习满足了市场监管的需求。在欧盟提出《通用数据保护条例》（GDPR），国内提出《中华人民共和国网络安全法》《中华人民共和国电子商务法》的背景下，数据隐私保护的法律法规会越来越严格化、全面化。企业需要保证用户数据的收集必须公开透明，企业之间不能在没有用户授权的基础上私自交换用户数据。过去可行的人工智能算法在这些严格的数据隐私保护前提下变得不太可行。因此需要有更高安全要求和隐私要求的联邦学习来帮助实现大数据产品和服务的提供。

尽管联邦学习这种交换模型参数而不交换具体数据的训练方式可以有效地保护用户的隐私，联邦学习依然面临一些安全性的风险。首先，联邦学习没有对参与方进行检测和校验，例如，没有审核参与方提供的参数模型是否真实。因此，恶意的参与方有可能通过提供虚假的模型参数来攻击和破坏联邦学习训练过程。这些虚假参数未经过校验就与正常的参数进行聚合，将会影响整体模型的最终质量，甚至会导致整个联邦学习过程无法收敛成一个可用的模型，进而导致训练失败。其次，联邦学习需要考虑是否对训练过程中的参数传递和存储进行隐私保护。一些研究表明，恶意的参与方可以依据联邦学习梯度参数在每一轮中的差异，反向推测出用户的敏感数据。因此，不通过加密保护的参数被泄露，在一定的程度上是可以成为攻击目标，从而间接泄露用户隐私数据的。

接下来将深入介绍联邦学习在隐私保护领域的前沿技术，详细探讨每一种隐私保护技术的原理、应用、潜在的挑战和未来的发展方向。联邦学习中常见的隐私保护技术包括安全多方计算（secure multi-praty computation，SMC）和差分隐私。本文将这些隐私保护技术分为降噪隐私保护和加密隐私保护两大类。

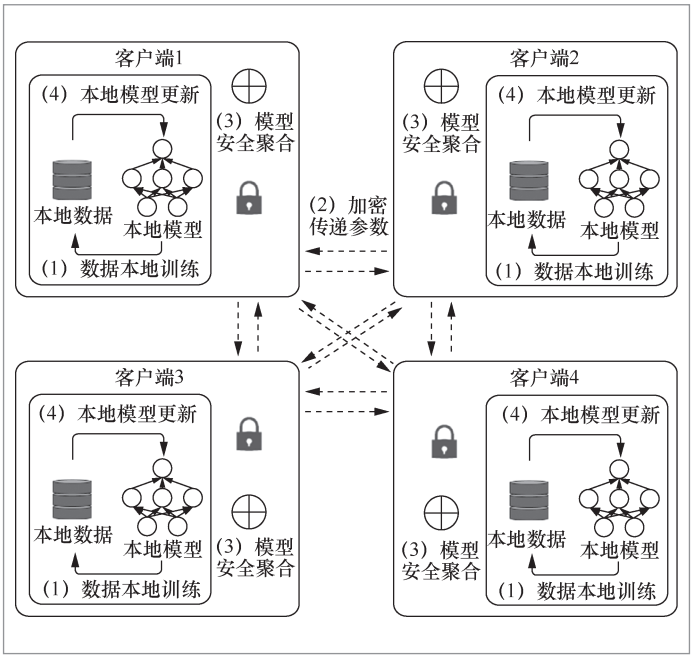

图 2 端到端联邦学习架构

降噪隐私保护主要是通过差分隐私等方法实现的。其主要原理是给数据添加噪声，或者使用归纳方法隐藏参与方的某些敏感属性，直到第三方无法通过差分攻击来区分个人为止，使数据无法还原，从而达到保护用户隐私的目的。但是这类方法会带来模型准确性上的损失，因此这种降噪隐私保护通常需要在参与方隐私与模型准确性之间进行权衡。

加密隐私保护主要是通过安全多方计算、同态加密等方法来实现的。安全多方计算成本较高，为降低数据传输成本，参与方可能需要降低对数据安全的要求来提高训练的效率。同态加密能够对所有数据进行加密处理，参与方接收到的是密文，攻击者无法推理出原始数据信息，从而保障数据层面的安全。因此，加密隐私保护通常需要设计复杂的加密计算协议来隐藏真实的输入和

输出。参与方和服务器之间传递的都是加密以后的参数信息，从而保证了这些加密过的参数信息即使被攻击，也不会泄露模型和用户隐私。但是加密隐私保护在计算量和模型效率上有更高的要求，因此这种加密类型的隐私保护通常需要在参与方计算效率和模型安全性之间进行权衡。

3 联邦学习攻击类型

3.1 联邦学习的隐私保护问题

联邦学习提供了一种可以保护用户数据隐私的训练模型，从而实现了参与方之间数据不共享而模型共享的机制。但是，最近一些工作表明，联邦学习可能并不能保证提供足够的隐私保护能力。例如，在联邦学习训练的参数通信更新的过程中，有可能会泄露一些敏感的信息。这些模型迭代过程中深层次的信息泄露可能由第三方攻击者造成，也可能通过中央服务器泄露。例如，参考文献[9]介绍了一种通过一小部分原始的梯度信息，反推出原始数据信息的隐私泄露方法。参考文献[10]介绍了恶意攻击者通过部分更新的梯队信息窃取原始数据的攻击方法。

在联邦学习被提出之前，机器学习的隐私保护问题一直是一个热门研究课题。作为一个创新科技，联邦学习与之前的隐私保护类型的研究领域也是紧密关联的。例如，隐私保护机器学习领域、基于安全性的分布式的机器学习、隐私安全类型的边缘计算等研究领域的进展，都对联邦学习隐私保护的研究提供了很好的参考和帮助。

尽管联邦学习提供了隐私保护的机制，还是有各种类型的攻击方式可以攻击联邦学习系统，从而破坏联邦学习系统安全和参与方的隐私。本节将讨论关于联邦学习的攻击问题。从参与方的类型来看，可以将联邦学习的威胁模型细分为半诚实模型（semi-honest model）和恶意模型。对于联邦学习系统的攻击，本文按照不同的维度进行不同层次的分类。从攻击方向角度来看，可以将联邦学习的攻击分为从内部发起和从外部发起两个方面。从攻击者的角色角度来看，可以将攻击分为参与方发起的攻击、中心服务器发起的攻击和第三方发起的攻击。从发动攻击的方式角度来看，可以将攻击分为中毒攻击和拜占庭攻击。从攻击发起的阶段角度，可以将攻击分为模型训练过程的攻击和模型推断过程的攻击。

3.2 联邦学习威胁模型类型

联邦学习系统是一种安全模型，需要根据需求设定相应的安全协议以及所需要的安全假设。在密码学领域，基于模型安全的假设通常可以被分为半诚实但好奇（honest but curious）的攻击方假设以及恶意攻击方假设。

3.2.1 半诚实但好奇的攻击方

半诚实但好奇的攻击方假设也被称为被动攻击方假设。被动攻击方会在遵守联邦学习的密码安全协议的基础上，试图从协议执行过程中产生的中间结果推断或者提取出其他参与方的隐私数据。目前联邦学习攻防方面的大部分研究假设模型威胁的类型为半诚实模型，这种模型设定有助于联邦学习理论研究过程中安全方案的设计。而在现实场景中，由于关于数据的法律法规等因素的约束，参与联邦学习模型训练的参与方大部分符合这类半诚实但好奇的攻击方假设，不会尝试进行极端的恶意攻击。

半诚实但好奇的参与方很多时候充当

的是客户端的角色，它们可以检测从服务器接收的所有消息，但是不能私自修改训练的过程。在一些情况下，安全包围或者可信执行环境（trusted execution environment，TEE）等安全计算技术的引入，可以在一定程度上限制此类攻击者的影响或者信息的可见性。半诚实但好奇的参与方将很难从服务器传输回来的参数中推断出其他参与方的隐私信息，从而威胁程度被削弱。

3.2.2 恶意攻击方

在设定联邦学习算法协议时，如果假设参与方为恶意攻击方，模型协议设定将会更加困难，模型的安全性要求也会更高。恶意攻击方也被称为主动攻击方。由于恶意攻击方不会遵守任何协议，为了达到获取隐私数据的目的，可以采取任何攻击手段，例如破坏协议的公平性、阻止协议的正常执行、拒绝参与协议、不按照协议恶意替换自己的输入、提前终止协议等方式，这些都会严重影响整个联邦学习协议的设计以及训练的完成情况。

恶意的参与方可以是客户端，也可以是服务器，还可以是恶意的分析师或者恶意的模型工程师。恶意客户端可以获取联邦建模过程中所有参与方通信传输的模型参数，并且进行任意修改攻击。恶意服务器可以检测每次从客户端发送过来的更新模型参数，不按照协议，随意修改训练过程，从而发动攻击。恶意的分析师或者恶意的模型工程师可以访问联邦学习系统的输入和输出，并且进行各种恶意攻击。在这种恶意攻击方假设的情况下，构造一个安全的联邦学习密码协议将会有很大的难度。通常情况下，需要在每一个可能被攻击的环节中引入安全多方计算协议。因此在相同的需求业务场景下，假设存在恶意攻击，联邦学习为了提升安全性，计算和通信代价会大大增加，并且关于协议的设计和实现也会变得更加困难，甚至会出现实际上无法使用联合训练的模型的情况，影响最终的产品效果和用户体验。

因此，在实际构建联邦学习业务时，大部分情况下系统面临的潜在攻击方来自半诚实但好奇的攻击方的威胁。由于法律法规的约束以及业务场景下强力的监管机制，恶意攻击将会承受严厉的处罚，因此大部分的研究假设参与方为半诚实但好奇的威胁模型，并且在此假设下构建隐私保护技术方案，从而显著地提高系统的安全和隐私保护性能，满足用户需求和提供优质用户体验。

3.3 联邦学习攻击类型分类

内部攻击可以由联邦学习服务器发起，也可以由联邦学习参与方发起。外部攻击（包括偷听者）通过参与方与服务器之间的通信通道发起。外部攻击的发起者大部分为恶意的参与方，例如敌对的客户、敌对的分析者、破坏学习模型的敌对设备或者其组合。在联邦学习中，恶意设备可以通过白盒或者黑盒的方式访问最终模型，因此在防范来自系统外部的攻击时，需要考虑模型迭代过程中的参数是否存在泄露原始数据的风险，这对严格的隐私保护提出了新的挑战。

内部攻击通常比外部攻击更强烈，攻击者更容易通过内部发动攻击。大部分的联邦学习攻击类型属于内部攻击类型。本文讨论的攻击方法主要属于内部攻击。联邦学习内部攻击可以分为以下3种类型。一是中毒攻击（可以细分为模型中毒攻击和数据中毒攻击），以中毒的方式污染或者破坏模型的数据或者模型，从而达到攻击目的。例如Bagdasaryan E等人[11]和 Bhagoji A N等人[12]介绍了一个恶意参与方攻击模型导致分类精度大幅下降的方法。恶意攻击者有时为

了达到攻击目的，会同时使用数据中毒攻击和模型中毒攻击。二是拜占庭攻击[13]，拜占庭恶意参与方会随机或者故意改变自己的输出，致使模型无法正常收敛，同时每次迭代可以输出类似的梯度更新结果，并且使得自己更难被发现。三是女巫攻击，攻击方伪装为参与方[14]攻击联邦学习模型，导致模型效果显著降低。

3.3.1 中毒攻击

一种中毒攻击是通过数据中毒发起的。数据中毒攻击方不能直接攻击发送给服务器的信息，而是通过替换本地数据的标签或特定的特征来操作客户端数据，从而发起攻击。当攻击方只能影响联邦学习系统边缘的数据收集过程，不能直接破坏学习系统中的导出量（例如模型更新）时，这种攻击往往很难被察觉。数据中毒是一种比模型中毒更具限制性的攻击类型，但是这种攻击方式更具有隐秘性。由于联邦学习会假定参与方遵守协议诚实地参与联邦训练，在实际部署中，检测有毒数据是一项很有挑战性的工作。参考文献[15]设计了一种专门针对中毒攻击的数据净化方法，达到移除模型的中毒数据或其他异常数据的目的。参考文献[16]在此基础上使用具有鲁棒性统计的数据净化方法抵御数据中毒，并且证明了该方法在少量异常值下能够保证鲁棒性。该方法在应对有针对性的数据中毒攻击以及无针对性的数据中毒攻击方面都取得了一定程度的成功。

另一种中毒攻击是通过模型中毒发起的。例如参考文献[11]介绍了通过加入后门的方式进行模型中毒攻击来攻击联邦学习系统。联邦学习的任何参与方都可以在联邦全局模型中引入隐藏的后门功能。例如致使图像分类器固定地识别出特定的标签结果，或者语义理解模型固定输出特定的错误结论。这种模型替换技术可以控制一个或多个攻击方“后门攻击”全局模型，使得最终的模型在攻击方选择的输入上表现不正确。攻击可以由一个参与方发起或者由多个参与方一起发起。实验结果显示，模型中毒攻击比只针对训练数据的数据中毒攻击更严重。

部分研究考虑如何从攻击策略上让模型中毒攻击更有效。参考文献[12]研究了联邦学习的模型中毒攻击问题，并且调整联邦学习中不同的攻击策略，从而实现更好的攻击效果。作者使用不同的策略进行攻击，例如，通过加速恶意参与方的更新速度来覆盖其他参与方的更新效果；通过变换最小化策略，变换和最优化训练损失函数以及攻击的目标函数等；通过预测良好节点的更新参数等策略提高模型中毒攻击的成功率。该文献展示了恶意攻击方通过这些攻击策略不仅可以实现模型攻击的目的，同时还能够保证攻击的隐秘性，使攻击不容易被系统发现。该文献的不足是目前的实验只基于服务器客户端的多参与方架构，在多方安全计算的对等网络架构下，攻击难度会更大，进行策略优化后的模型中毒攻击方案能否一样高效还未知。

还有些模型中毒攻击不是为了破坏整个模型，而是为了让模型按照自己的想法表现。参考文献[17]研究了神经网络中的中毒攻击问题。该文献提出一种基于优化的中毒生成方式，有效地证明了对迁移学习的图像分类器的攻击效果。该文献提出的方法是一种不需要控制标签功能的干净标签攻击方法，这使得中毒的训练数据似乎被正确地贴上了标签，从而达到不仅使攻击难以检测，而且为攻击方成功打开大门的攻击目的，并且攻击过程无须访问任何内部数据收集或者标记过程。

3.3.2 拜占庭攻击

拜占庭攻击[11]主要考虑的是多用户的情况。攻击方控制了多个用户，这些用户被称为拜占庭用户。拜占庭用户可以给中心服务器发送任意参数，而不是发送本地更新后的模型参数。这种攻击会导致全局模型在局部最优处收敛，甚至导致模型发散。假设拜占庭客户端拥有了访问联邦学习模型的权限，或者拥有非拜占庭式客户端更新的白盒访问权限，通过正常的模型更新调整输出，难以被系统检测。对于拜占庭类型的攻击，参考文献[12]提出通过冗余和数据洗牌的更新防御机制来防御拜占庭攻击，但是存在的问题是，这些机制通常具有严格的理论保证，并且建立在一些难以实现的假设上。例如，需要假定服务器可以直接访问数据，这些假设与联邦学习的实现存在矛盾，并且会增加通信成本。如何在联邦学习中协调和实现这种基于冗余的防御拜占庭攻击机制是一个很有挑战的问题。

3.3.3 女巫攻击

在联邦学习中，参与联邦训练的参与方需要信任服务器的专用通信渠道，与此同时，服务器也需要以公平和诚实的方式对待客户群。女巫攻击一般指网络中的单一节点可能具有多个身份标识，并且通过其控制系统的大部分节点来削弱网络冗余备份的作用。例如在社交网络中，可以通过少数节点控制多个虚假的身份，然后利用这些身份控制或者影响网络的大量正常节点。女巫攻击方式包括直接通信、伪造或者盗用身份、同时攻击和非同时攻击等方式。在联邦学习的服务器客户端架构的训练模型中，发动恶意攻击的参与方可以控制服务器，并伪造大量的客户端设备或者控制设备池中曾经受到破坏的设备，从而发动女巫攻击。这种攻击破坏了联邦学习协议的安全性，与此同时，有些联邦学习协议出于隐私考虑会将参与方的输入进行混合洗牌，这样会导致难以区分诚实用户和恶意用户，增加了抵御女巫攻击的难度。由于联邦学习系统不需要限制攻击方的数目，不需要训练过程以外的信息，并且对参与方以及他们的数据有更少的设定限制，现有的传统防御策略往往不足以抵御联邦学习过程中的女巫攻击。

Fung C等人[14]提出了一种叫作FoolsGold的抵御方法，用来抵御联邦学习中中毒方式的女巫攻击。具体的方法是根据贡献相似度动态适应客户端的学习速率，从而在分布式学习过程中通过大量的客户端更新信息识别出有毒的女巫攻击。但是不足之处是这种防御方式只能减轻女巫攻击，并且只在同时有很多攻击假设时才有效，例如假设攻击类型为变换标签策略或者后门策略时奏效。

3.4 联邦学习攻击的阶段分类

3.4.1 训练阶段的攻击

在训练过程中，攻击方可以试图学习、影响或者破坏联邦学习模型。在联邦训练的过程中，攻击方可以通过数据中毒攻击的方式改变训练数据集合收集的完整性，或者通过模型中毒攻击改变学习过程的完整性。攻击方可以攻击一个参与方的参数更新过程，也可以攻击所有参与方的参数更新过程。

训练过程中的攻击有时会被用作推理阶段的攻击的初始阶段。许多针对推理阶段的攻击的防御措施是在训练阶段部署的。在训练阶段，攻击者可以通过数据中毒、模型中毒等方式进行对抗攻击。在训练阶段，除了单个对手发动攻击的情况，还存在多个对手配合攻击的情况。不同的对手入侵不同的

客户端并且进行协调配合，从而完成中毒攻击。这种攻击比单独发动的攻击更加高效。多个恶意攻击方在训练阶段通过共谋发动的攻击给联邦学习的防御带来了新的挑战，当参与联邦学习的设备多达成千上万的级别之后，攻击方会在不定期的训练轮次中互相配合发动攻击，对于多参与方的联邦学习模型来说，如何找到这些训练过程中的恶意攻击方是一个很大的考验。

3.4.2 推理阶段的攻击

若联邦学习的参与方想利用各方的数据集合训练一个模型，但是又不想让自己的数据集泄露给服务器，就需要约定联邦建模的模型算法（例如神经网络）和参数更新的机制（例如随机梯度下降（stochastic gradient descent，SGD））。那么在训练前，攻击方就可以获取联邦学习参数更新的机制，从而指定对应的推断攻击策略。在理想条件下，一般假设参与方为两个：一个是被攻击方，另一个是攻击方。

推理攻击也被称作探索攻击（入侵攻击）[18]。通常情况下，推理攻击不会破坏目标模型，而是影响模型，从而使其输出错误的结果（或者攻击方希望的结果）。这种攻击的成功率或者有效性，在很大的程度上依赖于攻击方对整个模型的了解程度。推理攻击可以被分为白盒攻击和黑盒攻击。白盒攻击可以完全使用联邦学习模型；黑盒攻击只能查询联邦学习模型。与机器学习不同，在联邦学习中，服务器维护的联邦学习模型需要与很多的恶意客户端进行参数传递，因此，联邦学习需要更多地考虑如何抵御白盒入侵攻击。现在大部分的攻击类型属于训练阶段的攻击，属于推理阶段的攻击比较少。联邦学习攻击类型见**表1**。

3.5 联邦学习不同场景的攻击

在工业界，联邦学习通常会被分为ToB和ToC两种场景。比较有代表性的ToB场景有微众银行FATE联邦学习平台、平安科技联邦学习平台“蜂巢”系统等。这种模式主要应用于企业之间以隐私保护为目的的联邦学习应用。比较有代表性的ToC场景为谷歌公司，例如移动设备的键盘输入内容建议的应用。针对不同的场景，上述各类型的攻击方式会有不同的攻击效果。

3.5.1 ToB 场景的攻击

针对ToB场景，通常情况下参与联邦

表1 联邦学习攻击类型

对比项	攻击方式分类	主要的方法	描述
攻击方法	中毒攻击	污染或者破坏数据模型	数据中毒：通过替换本地数据的标签或者特定的特征影响联邦学习的过程。该方式虽然具有局限性但更加隐蔽 模型中毒：一个或多个传统参与方通过替换模型来影响模型更新过程。该方式比数据中毒的影响更严重
	拜占庭攻击	控制多个用户影响全局模型更新	攻击方控制多个用户，向中心服务器发送任意参数，由此影响全局模型，导致其偏离正常训练过程
	女巫攻击	直接通信、伪造或者盗用身份、同时攻击和非同时攻击等	单一节点具有多个身份信息，通过少数节点控制多个虚假身份，控制或影响网络中的大量正常节点
攻击阶段	训练阶段的攻击	在训练阶段，从数据或者模型的角度影响联邦学习模型	单个参与方或多个参与方在训练阶段通过入侵不同客户端且配合中毒攻击的方式，在不同的轮次发起攻击
	推理阶段的攻击	在推理阶段，影响模型的预测结果	该方式依赖于攻击方对模型的了解程度，被分为白盒攻击和黑盒攻击

学习训练的是企业，并且参与的企业数目通常是比较少的。例如银行系统内部的不同部门联合建立模型，或者某个银行系统和某个电商系统联合建立模型。企业之间通常在有一定信任基础的前提下才会签订一些安全隐私的合作协议，并且双方都会遵循联邦学习的训练协议，不会恶意攻击模型的收敛方向。但是在有些情况下，企业可能会根据服务器端接收到的模型参数进行反向推理，试图得到合作方的数据信息，目前这类攻击是比较常见的攻击方式。企业参与方不会用中毒数据或者中毒模型恶意攻击和破坏联邦学习的训练过程，而是遵照协议协同训练有益于每个合作企业方的联合模型，但是有可能存在某些企业试图获取别的企业的隐私数据的行为。拜占庭攻击之类的方式需要多个参与方协同配合破坏联邦学习的训练流程，此类攻击方式更适用于参与方很多的情况，很难在ToB场景下应用。由于参与联合训练的企业对联合模型的内部结构十分了解，并且参与了整个训练过程的参数更新，因此这种情况下的攻击多数为白盒攻击。

3.5.2 ToC场景的攻击

针对ToC场景，通常情况下参与联邦学习训练的是普通客户端，并且参与训练的客户端通常是非常多的。例如谷歌推出的一些基于联邦学习的应用，包括移动设备上的应用程序排名、移动设备的键盘输入内容建议、谷歌输入的下个词汇预测等，联邦模型的参与方很可能是来自上百万个用户的键盘输入信息，或者是移动终端使用过程中的隐私信息。这种类型的应用通常也是采用基于服务器客户端的架构。由于很多参与方都是移动终端（例如手机），因此很难约束和确保每一个终端都能遵守某种协议，这种情况下会发生更多的恶意攻击。这种应用情形下，客户端通常不会对其他客户端的隐私数据感兴趣，而是对联合模型更感兴趣。有的客户端会试图破坏联合模型，使联合模型难以收敛完成训练，例如敌对公司组织大量的恶意移动客户终端发动攻击。有的客户端试图改变联合模型，使最后的推理结果更有利于自己，例如在移动键盘输入内容建议中，发动数据中毒攻击，使联合模型多关联有利于自己公司的产品或者关键词。客户端可以通过恶意中毒数据或者恶意中毒模型的方式参与联邦学习训练。因此，这种ToC场景下的联邦学习往往会在客户端选择的角度进行优化，通过一些随机方式筛选优质的训练客户端，从而减轻这类恶意攻击的危害，同时也增加了攻击的难度。拜占庭攻击、女巫攻击等方式非常适合这一类ToC场景的联邦学习，并且有时通过一些恶意客户端的协同配合（例如在不同的训练参数更新轮次选择性地发动攻击），将会有不错的攻击效果，同时也具有很好的隐蔽性，不容易被发现。这类复杂的多参与方协同配合的恶意攻击，是ToC场景的联邦学习面临的比较大的挑战。

4 联邦学习的加密通信

数据的隐私保护是联邦学习的重要特性，增强隐私保护也是联邦学习抵御攻击的一种方式。考虑一个完整的联邦学习过程，在整个过程中联邦学习都应该保证数据持有方的数据隐私性，即保护用户数据本地化。一些中间信息的共享也可能导致信息泄露，例如模型更新或梯度信息[19-21]，因此对隐私的保护应该是多方面的。关于隐私保护的研究[22-23]已经比较充分，在设计联邦学习系统时可以利用这些成果。在实现联邦学习隐私保护的过程中，为了防御攻击，在实际部署时会使用一些加密技术，如

同态加密、哈希表加密等。本节将从安全多方计算、差分隐私和混合加密的角度对联邦学习的隐私保护进行系统性的论述。

4.1 加密隐私保护机制

4.1.1 混淆电路

混淆电路[24]早在1986年就被提出，用来解决百万富翁问题，即两个富翁如何在不暴露自己具体金额的情况下比较谁更富有。在联邦学习系统中，具体思路为：两个客户端分别为A和B，数据集分别为x、y。在联邦学习任务中，两个数据集通过函数$f(x,y)$实现最终的模型训练。首先，客户端A选取标签利用函数$f(x,y)$进行加密，得到$f'=g(f(x,y))$。然后发送f'和x对应的标签到客户端B。加密函数的操作不可逆，所以客户端B无法获得客户端A中的具体数据x。然后运行不经意间传输（oblivious transfer）[25-26]获取与y相关的标签。客户端B将y进行加密得到y'，并带入f'，然后解密f'，得到输出结果。最后将结果发送给客户端A。在运算过程中，客户端A和B无法获得彼此的原始数据，在满足计算要求的情况下，实现了对数据的保障。

4.1.2 同态加密

一般的加密方案关注的是数据存储安全，同态加密[27]是一类基于同态原理的特殊的加密函数，其关注的是数据处理安全，其允许直接对已加密的数据进行处理，而不需要知道任何关于解密函数的信息。也就是说，其他人可对加密数据进行处理，但在处理过程中无法得知任何原始数据信息。同时，基于同态加密的计算结果与直接对未加密数据进行计算的结果是一致的。

同态加密定义x和y是明文空间M中的元素，o是M上的运算，Ek(·)是M上密钥空间为K的加密函数，如果存在一个有效的算法F，使得：

$$F\left(\mathrm{Ek}(x),\mathrm{Ek}(y)\right)=\mathrm{Ek}(xoy) \qquad (2)$$

则称加密函数 Ek(·)对运算 o是同态的。在大多数应用场合中，同态加密方案需要支持两种基本、典型的运算，即加法同态运算和乘法同态运算。同时满足加法同态和乘法同态的函数被称为全同态。云计算和分布式机器学习等是同态加密的典型的应用场景，数据持有方传输数据前先将数据加密，云服务器在接收到数据后照例计算，只不过是在密文上进行的，待得到结果后再将结果的密文返还给数据持有方，数据持有方解密后即可得到最终结果。

同态加密具有较强的隐私保护能力，但是效率很难提升[27]，主要原因在于加密数据的运算量较大，计算速度比较慢，由此带来更多的数据存储空间占用问题。Dai W等人[28]利用CUDA GPU开发cuHE库，用于加速基于多项式的同态加密，并利用该库完成了较快的同态块密码实现。

在联邦学习的场景下，需要利用从多个数据源收集的汇总信息来训练模型，目的是在不披露关于单个数据源的细粒度信息的情况下进行培训。Yuan J W等人[29]提供了一种安全、高效、准确的“双同态”加密算法来支持对密文的灵活操作，从而对电子商务数据进行数值分析。Ho Q R等人[30]使用同态加密实现了横向线性回归的隐私保护协议。Hardy S等人[31]通过使用实体解析和同态加密对纵向分布数据进行具有隐私保护的联邦学习。

4.1.3 差分隐私

虽然同态加密可以通过对加密数据进行计算来保护学习过程，然而这些工具要求每个数据源执行大量的加密操作，并传输大量的密文，这使得它们反而成为整个系统的负担。对于联邦学习系统来说，选

择一个相对简单且不会对性能造成额外负担的算法是至关重要的。在这些不同的隐私方法中，差分隐私[32]由于其强大的信息理论保证、算法的简单性和相对较小的系统开销而得到广泛的应用。

差分隐私机制允许某个参与方共享数据集，并确保共享的形式只会暴露想要共享的那部分信息，保护的是数据源中一点微小的改动，解决例如插入或者删除一条记录导致的计算结果差异进而产生的隐私泄露问题。例如在数据集D发布之前，输出扰动机制使用随机扰动算法F干扰数据集D上的统计信息，这样扰动算法F的输出就不会暴露太多关于数据集D中任何特定数据记录的变量信息。如果一个扰动算法F提供差分隐私保护，那么两个相邻的数据集$D1$和$D2$中只有一个样本不同。对于扰动算法F的任何输出O，一定有：

$$\Pr\{F(D1)=O\} <= e\gamma \cdot \Pr\{F(D2)=O\} \quad (3)$$

其中，Pr为当前情况发生的概率值；e为约束因子；γ反映了扰动算法F的隐私保护水平，γ越小，隐私保护水平越高[33]。即如果该扰动算法作用于任何相邻数据集得到一个特定输出的概率差不多，就说这个扰动算法能达到差分隐私的效果。即观察者通过观察输出结果很难察觉出数据集的微小变化，从而达到保护隐私的目的。实践中通常使用拉普拉斯机制（Laplace mechanism）和指数机制（exponential mechanism）实现差分隐私保护。其中，拉普拉斯机制用于针对数值型结果的保护，指数机制用于针对离散型结果的保护[34-35]。

基于梯度的联邦学习方法，往往通过在每次迭代中随机地扰动中间输出来应用差分隐私[35-37]。也就是说，联邦学习的过程不会暴露是否使用某个特定的样本信息。现在流行的扰动方式有许多，例如，Wu X等人[38]对梯度数据添加了高斯噪声，Luca M 等人[39]采用了拉普拉斯噪声。此外，Bun M等人[40]还提出了一种利用定义一个线性上限$a(\lambda)$的方式对梯度进行剪辑，以限制每个参与方对整体更新的影响的方法。增加更多的噪声和扰动会提供更好的隐私保护，但可能会严重损害精度。因此，需要很细心地调整差分隐私和模型精度之间的平衡。Choudhury O等人[41]成功地将差分隐私部署在联邦学习框架内，用来分析与健康相关的数据，但是试验证明，差分隐私可能会带来较大的函数损失值。Geyer R C等人[42]证明了差分隐私对于保障数据持有方的数据隐私的有效性，同时认为大量的数据持有方会使带有差分隐私的联邦学习表现得更加稳定，准确率更高。

4.1.4 秘密分享

秘密分享[43]指将原本要传递的数据划分为多个部分，然后将它们依次发送到每个参与方。而仅通过一个或少部分参与方无法还原出原始数据，只有较大部分或者所有参与方将各自的数据凑在一起时，才能还原出原始数据[44]。例如，若参与方C要将数字c分发到其他参与方$A_1,A_2,\cdots,A_n$中，那么C首先生成$n-1$个随机数$c_1,c_2,\cdots,c_{n-1}$，然后计算第n个随机数$c_n=c-\sum_{i=1}^{n-1}c_i$，最后将$c_1,c_2,\cdots,c_n$发送给$A_1,A_2,\cdots,A_n$。因为$c_i$是随机数，所以单独一个或不多于$n-1$个随机数时，不会泄露任何信息，只有在$c_1,c_2,\cdots,c_n$全部出现时，才能得出$c$，因为$c=\sum_{i=1}^{n}c_i$。

上面是一个简单的秘密分享的例子，只有收到数据的所有参与方同时出现，才能恢复数据。根据实际情况需要，还可以选用阈值秘密分享[45]来确定至少需要多少个参与方才能恢复数据。

4.1.5 混合加密

根据上述的联邦学习隐私性相关技

术，一个自然的想法是能否将这些技术进行结合，即使用混合技术进行加密。基于上述想法，Pettai M等人[46]将安全多方计算与差分隐私技术结合，用来保护来自不同数据持有方的数据。类似地，Jeong E等人[47]也设计了一种结合了安全多方计算与差分隐私技术的联邦学习隐私保护系统，这种系统结合降噪差分隐私与加性同态加密，有效地保障了联邦学习系统的隐私性。Bonawitz K等人[48]将故障共享协议中的秘密共享技术与经过验证的加密技术结合，以安全地聚合高维技术。除此之外，Xu R H等人[49]提出了一种新的加密方法HybridAlpha，将差分隐私技术和基于功能加密的安全多方计算结合，该方法被证明拥有很好的通信效率。加密保护机制对比见**表2**。

4.2 加密计算环境

4.2.1 安全多方计算

在当前互联网场景下，各个公司拥有海量的数据，但是尚不能完成数据之间的安全流转，安全多方计算针对一组互不信任的参与方之间的协同计算问题提出隐私保护方案，安全多方计算要确保输入的独立性、计算的正确性以及去中心化等特征不受影响，同时不泄露各输入值给参与计算的其他成员[49-51]。安全多方计算主要针对的是在无可信第三方的情况下，如何安全地计算一个约定函数的问题，同时要求每个参与主体除了计算结果，不能得到其他实体的任何输入信息，在整个计算协议执行过程中用户对个人数据始终拥有控制权，只有计算逻辑是公开的。

现在针对多方安全计算的研究越来越多，包括秘密分享、同态加密、混淆电路以及差分隐私在内的各种加密保护机制被运用到该框架中。较为常见的是，利用混淆电路将需要计算的函数转化为布尔加密电路进行数据和标签传送，双方在此基础上无法通过标签反推输入信息，最终利用该电路完成计算并解密获取结果。在一些与金融相关的安全多方计算框架中，利用差分隐私对参与方数据添加噪声，以保护个体隐私。

4.2.2 可信计算环境

可信计算（trusted computing，TC）是可信计算组织（trusted computing group，TCG）推出的一项研究，希望通过专用的安全芯片（TPM/TCM）增强各种计算平台的安全性[52]。相较于可信计算，TEE更有利于便携设备的使用。因为该环

表 2 加密保护机制对比

加密算法	特点	性质	应用
混淆电路	非对称加密	以布尔函数的观点构造安全函数进行计算	常用来构建安全多方计算环境
同态加密	非对称加密	对密文进行代数运算解密后的结果与对明文进行相同的代数运算的结果相同	数据拥有方需要进行大量的运算，但本身算力不足时，常常使用同态加密
差分隐私	函数加密	通过向聚合查询结果添加随机噪声实现	保护个人条目，最大限度地减少记录识别机会
秘密分享	一般为对称加密	将秘密进行拆分，并分配给不同的参与方，单个参与方无法恢复秘密信息	防止秘密过于集中，实现风险分散和一定的入侵容忍性
混合加密	对称加密和非对称加密混合	对称加密对信息进行加密，非对称加密对密钥进行加密	结合对称加密和非对称加密的优点，保证消息机密性

境中的安全性可以被验证，可以将联邦学习过程中的一部分放到可信计算环境中。

与TEE相对应，传统移动设备的普通运行环境（rich execution environment，REE）技术具有开放性、可扩展性和通用性。但是在数据隐私安全的场景下，需要隔离的可信的环境来处理密钥和隐私数据。这里TEE与REE之间是相互隔离的，只能通过特定的端口互相通信。可信计算环境硬件机制保护的特性充分保障了数据的隐私安全性。

可信计算环境对联邦学习系统起到了很好的数据保护作用，为隐私等敏感数据提供了远程安全计算的保障。TEE已在较多的产品中推广应用，在阿里云Link TEE系列产品中，针对密码算法和密钥管理，利用国密加密算法，进行密钥层级的架构和管理；在英特尔的SGX（software guard extensions）指令集[53]中也应用了TEE，用于保护敏感数据。

5 结束语

联邦学习系统的客户端与服务器之间往往存在很多的通信节点，信息在节点间进行传输时，由于端口开放等因素，一旦被监听就容易产生信息泄露的情况。为了进一步保障数据的隐私安全，需要对联邦学习系统通信过程进行加密。这方面的安全保障可以从网络编码的角度进行考虑，防止监听的数据被解析。除了客户端设备的异构性，数据的隐私敏感性也是联邦学习的重要特点。为此，联邦学习系统需要在多环节从多角度考虑数据的安全问题。这是联合训练模型的基础要求，更是激励用户广泛参与的前提保障。目前，联邦学习正逐渐发展为一个综合性的研究领域，但主体的要求始终是隐私保护。本文从隐私保护的角度，对目前联邦学习的发展情况进行了综合性的阐述。首先从隐私保护的角度讨论了联邦学习的发展历史、定义以及在业务场景下的分类；然后对联邦学习场景下隐私保护面临的问题进行了综述，整理了联邦学习中的攻击类别。

根据本文的分析总结，联邦学习目前面临的瓶颈和未来的研究方向可以归纳为以下几点。

- 对于联邦学习的中心服务器来说，在联邦建模的过程中，其抗恶意节点攻击的能力较弱，不能完全保证参与方的贡献均为正向。如何识别恶意节点以及减少恶意节点带来的影响都是值得研究的问题。
- 联邦学习理想状态是实现一种完全去中心化的联合建模框架，但就目前发展的情况而言，完全去中心化仍然存在困难，且许多业务场景也确实需要中心服务器。链式联邦的构想会是一个解决该问题的方向。
- 本文在第3节描述了联邦学习可能面临的攻击情况，按照联邦学习目前的研究进展，这些攻击是无法被完全抵御的，这将会导致产业落地困难，因此，针对联邦学习系统鲁棒性和对抗攻击等方向的研究是非常重要的。

参考文献:

[1] KONEČNÝ J, MCMAHAN H B, RAMAGE D, et al. Federated optimization: distributed machine learning for on-device intelligence[J]. arXiv preprint, 2016, arXiv:1610.02527.

[2] GOODFELLOW I, YOSHUA B, AARON C. Deep learning[M]. Massachusetts: MIT Press, 2016.

[3] 王健宗, 黄章成, 肖京. 人工智能赋能金融科

技[J]. 大数据, 2018, 4(3): 114–119.
WANG J Z, HUANG Z C, XIAO J. Artificial intelligence energize Fintech[J]. Big Data Research, 2018, 4(3): 114–119.

[4] KONEN J, MCMAHAN H B, YU F X, et al. Federated learning: strategies for improving communication efficiency[J]. arXiv preprint, 2016, arXiv:1610. 05492.

[5] 刘雅辉, 张铁赢, 靳小龙, 等. 大数据时代的个人隐私保护[J]. 计算机研究与发展, 2015, 52(1): 229–247.
LIU Y H, ZHANG T Y, JIN X L, et al. Personal privacy protection in the era of big data[J]. Journal of Computer Research and Development, 2015, 52(1): 229–247.

[6] 孟绪颖, 张琦佳, 张瀚文, 等. 社交网络链路预测的个性化隐私保护方法[J]. 计算机研究与发展, 2019, 56(6): 1244–1251.
MENG X Y, ZHANG Q J, ZHANG H W, et al. Personalized privacy preserving link prediction in social networks[J]. Journal of Computer Research and Development, 2019, 56(6): 1244–1251.

[7] 韩璇, 袁勇, 王飞跃. 区块链安全问题: 研究现状与展望[J]. 自动化学报, 2019, 45(1): 206–225.
HAN X, YUAN Y, WANG F Y. Security problems on blockchain: the state of the art and future trends[J]. Acta Automatica Sinica, 2019, 45(1): 206–225.

[8] YANG Q, LIU Y, CHEN T J, et al. Federated machine learning: concept and applications[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1–19.

[9] PHONG L T, AONO Y, HAYASHI T, et al. Privacy–preserving deep learning via additively homomorphic encryption[J]. IEEE Transactions on Information Forensics and Security, 2018(5): 1.

[10] ZHU L, LIU Z, HAN S. Deep leakage from gradients[C]//Proceedings of the Advances in Neural Information Processing Systems. [S.l:s.n.], 2019: 14774–14784.

[11] BAGDASARYAN E, VEIT A, HUA Y, et al. How to backdoor federated learning[C]// Proceedings of the International Conference on Artificial Intelligence and Statistics. [S.l.:s.n.], 2020.

[12] BHAGOJI A N, CHAKRABORTY S, MITTAL P, et al. Analyzing federated learning through an adversarial lens[C]// Proceedings of the International Conference on Machine Learning. [S.l.:s.n.], 2019.

[13] CHEN L J, WANG H Y, CHARLES Z, et al. DRACO: byzantine–resilient distributed training via redundant gradients[J]. arXiv preprint, 2018, arXiv:1803. 09877.

[14] FUNG C, YOON C J M, BESCHASTNIKH I. Mitigating sybils in federated learning poisoning[J]. arXiv preprint, 2018, arXiv: 1808.04866.

[15] ABHISHEK B, JOHN D, JULIEN F, et al. Protection against reconstruction and its applications in private federated learning[J]. arXiv preprint, 2018, arXiv: 1812.00984.

[16] CARLINI N, LIU C, KOS J, et al. The secret sharer: measuring unintended neural network memorization & extracting secrets[J]. arXiv preprint, 2018, arXiv: 1802.08232.

[17] FREDRIKSON M, JHA S, RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures[C]//Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2015: 1322–1333.

[18] BARRENO M, NELSON B, SEARS R, et al. Can machine learning be secure[C]// Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security. New York: ACM Press, 2006.

[19] 孙慧中, 杨健宇, 程祥, 等. 一种基于随机投影的本地差分隐私高维数值型数据收集算法[J]. 大数据, 2020, 6(1): 1–11.
SUN H Z, YANG J Y, CHENG X, et al. A high–dimensional numeric data collection algorithm for local difference privacy based on random projection[J]. Big Data Research, 2020, 6(1): 1–11.

[20] 王平, 张玉书, 何兴, 等. 基于安全压缩感知的大数据隐私保护[J]. 大数据, 2020, 6(1): 12–22.

WANG P, ZHANG Y S, HE X, et al. Big data privacy protection based on secure compressive sensing[J]. Big Data Research, 2020, 6(1): 12–22.

[21] 卢文雄, 王浩宇. 基于同源策略的移动应用细粒度隐私保护技术[J]. 大数据, 2020, 6(1): 23–34.
LU W X, WANG H Y. Same origin based fine-grained privacy protection for mobile applications[J]. Big Data Research, 2020, 6(1): 23–34.

[22] 孟小峰, 王雷霞. 人工智能时代的数据隐私、垄断与公平[J]. 大数据, 2020, 6(1): 35–46.
MENG X F, WANG L X. Data privacy, monopoly and fairness for AI[J]. Big Data Research, 2020, 6(1): 35–46.

[23] 李政, 洪莹. 基于隐私保护的政府大数据治理研究[J]. 大数据, 2020, 6(2): 69–82.
LI Z, HONG Y. Study on big data management for government based on privacy protection[J]. Big Data Research, 2020, 6(2): 69–82.

[24] YAO C C. How to generate and exchange secrets[C]//Proceedings of the Symposium on Foundations of Computer Science. Piscataway: IEEE Press, 2008.

[25] NAOR M, PINKAS B. Efficient oblivious transfer protocols[C]//Proceedings of the 20th Annual Symposium on Discrete Algorithms. [S.l.:s.n.], 2001.

[26] RABIN M O. How to exchange secrets with oblivious transfer[J]. IACR Cryptol. ePrint Arch., 2005(187).

[27] HALEVI S, SHOUP V. Design and implementation of a homomorphic-encryption library[J]. IBM Research (Manuscript), 2013, 6: 12–15.

[28] DAI W, SUNAR B. A homomorphic encryption accelerator library[C]// Proceedings of the Springer International Publishing, [S.l.:s.n.], 2015.

[29] YUAN J W, YU S C. Privacy preserving back-propagation neural net-work learning made practical with cloud computing[J]. IEEE Transactions on Parallel and Distributed Systems, 2013, 5(1): 212–221.

[30] HO Q R, CIPARJ, CUI H G, et al. More effective distributed ml via a stale synchronous parallel parameter server[J]. Advances in Neural Information Processing Systems, 2013: 1223–1231.

[31] HARDY S, HENECKA W, IVEY-LAW H, et al. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption[J]. arXiv preprint, 2017, arXiv:1711.10677.

[32] DWORK C. A firm foundation for private data analysis[J]. Communications of the ACM, 2011, 54(1): 86–95.

[33] ABADIM, CHUA, GOODFELLOW I, et al. Deep learning with differential privacy[C]//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2016: 308–318.

[34] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]//Proceedings of the Theory of Cryptography Conference. [S.l.:s.n.], 2006: 265–284.

[35] DWORK C, ROTH A. The algorithmic foundations of differential privacy[J]. Foundations and Trends® in Databases, 2014, 9(3–4): 211–407.

[36] BASSILY R, SMITH A, THAKURTA A. Private empirical risk minimization: efficient algorithms and tight error bounds[C]//Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science. Piscataway: IEEE Press, 2014: 464–473.

[37] PAPERNOT N, SONG S, MIRONOV I, et al. Scalable private learning with pate[J]. arXiv preprint, 2018, arXiv:1802. 08908.

[38] WU X, LI F G, KUMAR A, et al. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics[C]// Proceedings of the 2017 ACM International Conference on Management of Data. New York: ACM Press, 2017: 1307–1322.

[39] LUCA M, GEORGE D, EMILIANO DE C. Efficient private statistics with succinct

sketches[J]. arXiv preprint, 2015, arXiv: 1508.06110.

[40] BUN M, STEINKE T. Concentrated differential privacy: simplifications, extensions, and lower bounds[C]//Proceedings of the Theory of Cryptography Conference. Berlin: Springer, 2016: 635–658.

[41] CHOUDHURY O, GKOULALAS-DIVANIS A, SALONIDIS T, et al. Differential privacy-enabled federated learning for sensitive health data[J]. arXiv preprint, 2019, arXiv:1910.02578.

[42] GEYER R C, KLEIN T, NABI M. Differentially private federated learning: a client level perspective[J]. arXiv preprint, 2017, arXiv:1712.07557.

[43] TIAN X X, SHA C F, WANG X L, et al. Privacy preserving query processing on secret share based data storage[C]// Proceedings of the International Conference on Database Systems for Advanced Applications. Berlin: Springer, 2011: 108–122.

[44] BONAWITZ K, IVANOV V, KREUTER B, et al. Practical secure aggregation for federated learning on user-held data[J]. arXiv preprint, 2016, arXiv:1611.04482.

[45] TASSA T. Hierarchical threshold secret sharing[J]. Journal of Cryptology, 2007, 20(2): 237–264.

[46] PETTAI M, PEETER L. Combining differential privacy and secure multiparty computation[C]//Proceedings of the 31st Annual Computer Security Applications Conference. New York: ACM Press, 2015.

[47] JEONG E, OH S, KIM H, et al. Communication-efficient on-device machine learning: federated distillation and augmentation under non-iid private data[J]. arXiv preprint, 2018, arXiv:1811.11479.

[48] BONAWITZ K, IVANOV V, KREUTER B, et al. Practical secure aggregation for privacy-preserving machine learning[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2017: 1175–1191.

[49] XU R H, BARACALDO N, ZHOU Y, et al. HybridAlpha: an efficient approach for privacy-preserving federated learning[C]//Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security. New York: ACM Press, 2019.

[50] CHAUM D. The dining cryptographers problem: unconditional sender and recipient untraceability[J]. Journal of Cryptology, 1988, 1(1): 65–75.

[51] SLAWOMIR G, LI X. A comprehensive comparison of multiparty secure additions with differential privacy[J]. IEEE Transactions on Dependable and Secure Computing, 2015, 14(5): 463–477.

[52] SADEGH RM, CHRISTIAN W, OLEKSANDR T, et al. Chameleon: a hybrid secure computation framework for machine learning applications[C]//Proceedings of the 2018 on Asia Conference on Computer and Communications Security. New York: ACM Press, 2018: 707–721.

[53] FENG D G, QIN Y, FENG W, et al. The theory and practice in the evolution of trusted computing[J]. Chinese Science Bulletin, 2014, 59(32): 4173–4189.

作者简介

王健宗（1983– ），男，博士，平安科技（深圳）有限公司副总工程师，资深人工智能总监，联邦学习技术部总经理。美国佛罗里达大学人工智能博士后，中国计算机学会高级会员，中国计算机学会大数据专家委员会委员，主要研究方向为联邦学习和人工智能等。

孔令炜（1995– ），男，平安科技（深圳）有限公司联邦学习团队算法工程师，中国计算机学会会员，主要研究方向为联邦学习系统和安全通信等。

黄章成（1990– ），男，平安科技（深圳）有限公司联邦学习团队资深算法工程师，人工智能专家，中国计算机学会会员，主要研究方向为联邦学习、分布式计算及系统和加密通信等。

陈霖捷（1994– ），男，平安科技（深圳）有限公司联邦学习团队算法工程师，主要研究方向为联邦学习与隐私保护、机器翻译等。

刘懿（1994– ），女，平安科技（深圳）有限公司联邦学习团队算法工程师，主要研究方向为联邦学习系统等。

卢春曦（1994– ），女，平安科技（深圳）有限公司联邦学习技术团队产品经理，负责联邦学习系统研发与应用落地。

肖京（1972– ），男，博士，中国平安集团首席科学家，2019年吴文俊人工智能杰出贡献奖获得者，中国计算机学会深圳分部副主席，主要研究方向为计算机图形学、自动驾驶、3D显示、医疗诊断、联邦学习等。

收稿日期： 2020–09–29

基金项目： 国家重点研发计划资助项目（No.2018YFB1003503，No.2018YFB0204400，No.2017YFB1401202）

Foundation Items: The National Key Research and Development Program of China (No.2018YFB1003503, No.2018YFB0204400, No.017YFB1401202)