安全多方计算与联邦学习技术、应用与监管框架

安全多方计算的数学描述为，“ 有n个参与者P1,P2，…Pn，要以一种安全的方式共同计算一个函数，这里的安全是指输出结果的正确性和输入信息、输出信息的保密性。具体地讲，每个参与者P1，有一个自己的保密输入信息X1，n个参与者要共同计算一个函数f(X1,X2, … ,Xn)=(Y1,Y2, … ,Yn),计算结束时，每个参与者Pi只能了解Yi,不能了解其他方的任何信息。”[[1]](#footnote-0)其起源为20世纪80年代姚期智教授提出的百万富翁问题。在成熟的应用中，多方安全计算更多是集合中的逻辑运算与域上的加法运算，解决数据传输中的信任问题。

+多方安全计算的应用

多方安全计算技术作为联邦学习技术，在数据安全与信息加密领域必备的前置技术，在联邦学习中发挥出更大的价值，在金融领域更具有革命性意义。其影响范围深远，可能对金融行业产生意料之外的冲击。因此我们将联邦学习技术作为关注的重点进行研究。

联邦学习是一种全新的数据挖掘模式，根据其用户特征与用户的重叠程度，分为横向纵向与迁移学习。两个数据集的用户特征（X1,X2,…）重叠部分较大，而用户(U1, U2…)重叠部分较 小；两个数据集的用户(U1, U2…)重叠部分较大，而用户特征（X1,X2,…）重叠部分较小；两个数据集的用户(U1, U2…)与用户特征重叠（X1,X2,…）部分都比较小这两种情况分别对应了横向联邦，纵向联邦与迁移学习。利用不同的技术，联邦学习提出了对数数据孤岛问题的解决方案。其一改传统数据挖掘所需的数据收集过程，转变为本地建模并收集模型。我们利用现有工业级框架FATE分析了联邦学习建模的特点。

传统机器学习过程中，数据将会被集中集中于数据寡头手中进行集中建模。在我们的试验中，我们利用数据库基数估计数据模拟了该过程。我们收集了13个数据库，每一个数据库数据量从过百到8000不等，共计1万条数据。利用传统学习建模框架pytorch进行建模，需要读入全部数据后才能进行分析，这样会导致中心化节点——数据寡头企业掌握大量明码信息，产生数据安全问题。在FATE的框架下，同样利用pytorch进行建模，该框架所设计的算法将单独对13个数据库，每一个数据库分别进行建模，得到一个image为格式的矩阵，该矩阵不包含任何明码信息，且其不是加密技术，而是特征提取技术，无法对其还原为明码信息。但同时，该模型包含了所关注标签的特征信息，并上传给中心节点——数据寡头企业。中心节点利用联邦技术，对本地模型进行整合，学习出一个包含本地模型特征的联邦模型，该模型强调，在不知晓各节点信息的情况下拟合出一个提取了各节点信息特征的模型。我们屏幕后发现联邦学习训练的联邦模型与传统机器学习方法训练的模型相差不大，这证明了联邦学习在技术上已趋近成熟，它解决了数据安全问题，将会对金融为代表的行业产生冲击。

所谓数据安全问题，我们认为包含以下三个方面：第一，对于企业来说，非公开的数据涉及到隐私民生等敏感问题的数据数据，企业无法进行收集；但这类数据往往包含巨大的数据信息价值；第二，对于用户来讲，企业收集数据过程中可能造成意外的数据泄露，这将对用户的隐私安全造成影响；第三也是最令人担心的，企业收集数据后，可能拿收集的数据进行见不得人的事情，比如被国际上广泛讨论的监控问题。相应的国际国内有诸多法律法规对其进行节制，例如欧盟出台的《数据治理法》，以及我国的隐私保护法，网络安全法等。联邦学习技术避免了数据传输与收集，改为收集本地模型，对于用户来说，无疑是解决了隐私安全方面的担忧，而对于企业来说，本地建模也为其收集更多数据特征，特别是当下被认为是应当保密的数据的特征，例如企业的税务特征，提供了一种可能，这将极大丰富数据企业收集数据特征规模，极大优化其建立的模型，有着广阔的应用前景。

区块链相关

联邦学习本身尚没有大规模应用于工业界，但对于其联盟框架的讨论，学界已有较为统一的声音。联邦学习技术作为一种脱胎于分布式机器学习的算法，天然具有联盟属性，能够与区块链特别是联盟链技术相结合，完善联邦学习框架，创造出一种较为成熟的商业应用模式。区块链作为一个去中心化的、不可变的、共享的分布式账本和数据库，为联邦学习提供了较好的架构与底层支撑。[[2]](#footnote-1)基于联盟链的区块链联邦学习（BlockFL），用区块链网络来替代传统机器学习中的中央服务器，区块链网络允许交换设备的本地模型更新，同时验证和提供相应的激励机制。利用区块链的不可篡改性保证了模型的正确性。[[3]](#footnote-2)更进一步的，基于梯度采集与参数更新的正确性，以及数据的隐私保护问题，深度链被提出用于联邦学习。深度链是一种基于区块链激励机制的安全分散框架和用于隐私保护的分布式深度学习的密码原语来提供数据机密性、计算可审计性和参与联邦训练的激励机制。

与区块链类似的是，出于技术信任，联邦学习算法在整个联盟中起支配性作用，其存储了不断迭代的模型，控制着基于模型的定价，分配由过程模型产生市场价值，并将它的部分返还给数据提供者。从而鼓励参与者对模型的训练，进一步加强模型的效果。[[4]](#footnote-3)我们发现在联邦学习目前的应用场景中基本遵循了这一框架。

+应用介绍（胡银东部分）

但同时，通过对目前应用模式的分析，我们发现其确有不完善之处，将会给监管带来挑战。在区块链中，不同节点间地位平等，而在联邦学习联盟链上，数据提供者与数据收集者的地位天然就不平等。在以比特币为代表的区块链上，技术作为超然于经济利益的存在，能够参与公平的分配。但在联邦学习模型中，技术由数据收集者所拥有，被少数数据寡头企业所掌控，这导致本应由技术分合理分配的联邦学习总收益变成了企业利润，导致了数据寡头对数据提供者的剥削。[[5]](#footnote-4)

我们注意到，[[6]](#footnote-5)数据本是用户的一项私人资源，数据所有权也就是一项私人权力。在当前大数据平台尚采用分布式系统，收集私人数据并被存储于中心化数据库的过程中，受到隐私保护法等法律的节制，用户特征信息提取尚不完善的情况下，依然使科技公司产生了大数据支配优势，用户在平台上留下的任何结构性的和非结构性的数据，经过科技公司的数学模型分析后，变得具有预测性。隐秘在用户深处的欲望、需求、情绪、情感可能被算法洞悉，科技公司可借此推送信息，引导消费，改变甚至控制人们的思想及行为，造成了严重的数据剥削，已经值得当前社会警惕；而如前文所述，联邦学习技术解决了数据安全问题，在大数据企业收集数据中不再有隐私问题的困扰，其对用户特征的提取，将会在用户不知情且不违法的情况下变本加厉的进行，进一步加剧社会矛盾。

同时联邦学习技术为数据寡头间的数据交换提供了可能。在联盟链中，数据提供方与数据需求方均为大数据企业时，双方均有提取对方数据特征的需求，此时，提供与收集方才在链上处于同样的地位，在形式上达成了联盟的公平，这种公平的联盟正是数据寡头间进行数据合作的基础，在不泄露己方数据核心机密的情况下，分享数据特征，形成合作。而数据寡头间由数据竞争变为数据合作，也将进一步剥削普通用户的数据价值。[[7]](#footnote-6)

因此针对这些新兴的联盟，必须要有与之相称的监管措施。但联邦学习算法无疑是数据企业最为核心的价值所在，若继承传统工商监管的思路，直接对企业行为进行监管的话，监管方势必将插手联邦学习建模过程，对于企业来说也是一种不公平。这样的问题同样存在于区块链的监管当中。区块链作为一种基于技术信任所搭建的平台，若监管方直接监控区块链的运行，将破坏区块链去中心化的公平特征，使其失去原有的价值。

共票相关

共票是区块链治理的新方案，同样作为具有联盟链结构的联邦学习模型，也可以由共票进行规范。“共票”是基于区块链等技术对数据进行确权、定价、交易、开放、共享、赋能,实现集政府、劳动者、投资者、消费者与管理者多位一体的数据共享分配机制。[[8]](#footnote-7)前文也提到，基于隐私保护的数据监管，在联邦学习体系下已经不适用；而共票所代表的数据价值的回报正是联邦学习监管的方向。共票是一种全新的，基于区块链架构的利益分配机制，强调数据集中利用之后发挥更大价值，并合理进行分配。但在技术上，要求数据集中利用，存在隐私保护与数据安全问题。而联邦学习从技术上解决了数据安全问题，但就目前来看，需要对处于弱势地位的数据提供者的数据价值进行保护。我们发现共票与联邦学习在是优缺点互补的。

当然联邦学习所在的联盟链与传统区块链结构有所不同，以比特币为代表的区块链，利用发代币作为其利益的分配，这在联盟链中显然是不适用的。而结合共票在贵州省的实践来看，其次取得积分制或许更适合联邦学习联盟。更进一步的，基于联邦学习没有数据安全问题的技术信任，可以构建一个数据特征的交易市场，数据企业提供算法，数据提供者利用企业的算法在本地进行建模，并在市场上售卖本地模型，其交易的媒介正是共票。如此一来避免了数据企业抽取用户数据所造成的剥削，使得本在联邦学习联盟中处于弱势地位的数据提供者，利用供销市场能够与数据企业进行平等数据交易，保障了数据价值。其中，监管方监管市场行为，对企业所提供的算法进行验证，由监管方监管方背书，保证其不泄露用户隐私，确保供销市场健康运行。对于该共票数据市场的监管，可以类比于传统金融市场的监管，极大降低了监管难度。

从目前来看，联邦学习在工业上的应用还处于企业间合作阶段，仍未面向大众进行数据收集；同样的供票理论也正处于萌芽状态，需要进行进一步的实践。两者如果能够相结合并推广，借助新兴数据企业扩张所必需的数据需求，利用联邦学习技术在特征提取上的优势，带动共票理论落地，实现社会的数据公平，其意义无疑是重大的。我们也将在我们独立设想的CFS框架中与共票相结合。

+CFS

1. 陈杰，《可调节安全等级的多方安全计算》，《软件》2012年 第2期 [↑](#footnote-ref-0)
2. 《On-Device\_Federated\_Learning\_via\_Blockchain\_and\_it》 https://arxiv.org/abs/1808.03949v1​arxiv.org [↑](#footnote-ref-1)
3. 《DeepChain: Auditable and Privacy-Preserving Deep Learning with Blockchain-based Incentiv》，https://www.researchgate.net/publication/336361007\_DeepChain\_Auditable\_and\_Privacy-Preserving\_Deep\_Learning\_with\_Blockchain-based\_Incentive​www.researchgate.net [↑](#footnote-ref-2)
4. 翁翕，丛明舒，于涵：《联邦学习之激励机制设计》，https://dl.ccf.org.cn/lecture/lectureDetail?id=4841322416883712 [↑](#footnote-ref-3)
5. 田锋 吕金伟，《西方资本主义的时弊:从数据资本到数据剥削》，《自然辩证法研究》 2020年 第8期 [↑](#footnote-ref-4)
6. 《算法，即剥削》，清和，《智本社》 [↑](#footnote-ref-5)
7. 钟伟，《警惕数据寡头终极垄断》，《新金融》 2020年 第1期 [↑](#footnote-ref-6)
8. 杨东：《“共票”：区块链治理新维度》，《东方法学》，2019年第三期 [↑](#footnote-ref-7)