arXiv:2304.02932v1 [cs.CR] 6 Apr 2023

联邦知识图谱嵌入中的隐私威胁量化与防御

Yuke Hu 浙江大学, HIC-ZJU 中国杭州
yukehu@zju.edu.cn

Wei Liang 浙江大学, HIC-ZJU 中国杭州
liangwei23@zju.edu.cn

Ruofan Wu 蚂蚁集团 上海，中国
ruofan.
wrf@antgroup.com

Kai Xiao 蚂蚁集团 上海, 中国
xiaokai.xk
@antgroup.com

Weiqiang Wang 蚂蚁集团 杭州，中国
w
eiqiang.wwq@antgroup.com

Xiaochen Li 浙江大学, HIC-ZJU 杭州，中国
xiaochenli@zju.edu.cn

Jinfei Liu 浙江大学, HIC-ZJU 中国杭州
jinfeiliu@zju.edu.cn

Zhan Qin*浙江大学, HIC-ZJU 杭州，中国
qinzhan@zju.edu.cn

摘要

知识图谱嵌入（KGE）是一项基础技术，可从知识图谱（KG）中提取富有表现力的表示，以支持各种下游任务。新兴的联邦知识图谱嵌入（FKGE）通过协作训练客户端持有的分布式知识图谱，同时避免交换客户端的敏感原始知识图谱。然而，正如其他联邦模型训练（如神经网络）中所证实的那样，它仍可能面临隐私威胁。尽管如此，对于具有独特特性、不同于以往研究模型的 FKGE，如何量化并防御此类隐私威胁仍未得到探索。本文从攻击和防御两个角度，首次对 FKGE 面临的隐私威胁进行了全面研究。针对攻击，我们提出三种新的推理攻击来量化隐私威胁；这些攻击能够成功推断受害客户端中知识图谱三元组是否存在，揭示了巨大的隐私风险。针对防御，我们提出 DP-Flames，这是一种采用隐私选择的新型差分隐私 FKGE。它利用 FKGE 的实体绑定稀疏梯度特性，实现更优的隐私-效用权衡，并结合最先进的私有选择技术，提供紧致的隐私会计方法。我们进一步提出一种自适应隐私预算分配策略，以在训练过程中动态调整防御强度。全面评估结果表明，所提出的防御方法能够有效降低推理攻击的成功率，平均从 83.1% 降至 59.4%，而效用仅有小幅下降。

*通讯作者

允许为个人或课堂使用而免费制作本作品的全部或部分数字或纸质副本，但前提是不得为营利或商业利益制作或分发副本，并且副本首页须附有本声明及完整引用信息。必须尊重本作品中由作者以外的其他方拥有版权的组成部分。在注明出处的情况下，允许撰写摘要。以其他方式复制、重新发表、发布到服务器或重新分发到列表，均须事先获得特定许可和/或支付费用。请通过 permissions@acm.org申请许可。*WWW '23*，*2023*年*4*月*30*日至*5*月*4*日，美国得克萨斯州奥斯汀 © 2023年 版权所有归所有者/作者所有。出版权授权给 ACM。ACMISBN 978-1-4503-9416-1/23/04...$15.00https://doi.org/10.1145/3543507.3583450

CCS 概念

• **安全与隐私 → 隐私保护**。

关键词

知识图谱嵌入、联邦学习、成员推断攻击、差分隐私

ACM 参考格式:

Yuke Hu、Wei Liang、Ruofan Wu、Kai Xiao、Weiqiang Wang、Xiaochen Li、Jinfei Liu 和 Zhan Qin。2023年。联邦知识图谱嵌入中的隐私威胁量化与防御。载于 *2023*年 *ACM Web* 大会论文集（*WWW '23*），*2023*年*4*月*30*日至*5*月*4*日，美国得克萨斯州奥斯汀。ACM，美国纽约州纽约市，12页。https://doi.org/10.1145/3543507.3583450

1 引言

近年来，知识图谱（KGs）取得了长足发展。许多大规模知识图谱，例如 Freebase [4],、NELL [8],、YAGO [33], 和 Wikidada [40]，已被构建并应用于各种下游应用中，例如推荐系统[49],、语义网 [42],、知识推理 [6], 和问答 [5]。知识图谱嵌入（KGE）通过将知识图谱的组成部分表示为连续空间（也称为嵌入空间）中的稠密向量[6]，旨在解决知识图谱固有的符号性质所带来的阻碍其广泛应用的问题。各类组织（例如电子商务公司和金融机构）拥有的大多数知识图谱都不完整，并且由于隐私顾虑和监管压力（例如《通用数据保护条例》 [1] 和《加州消费者隐私法案》[2]），无法直接交换。为了在不损害隐私的情况下利用不同知识图谱之间的互补性，研究人员提出了若干联邦知识图谱嵌入（FKGE）模型 [9, 30, 50] ，通过仅交换模型参数（即 FKGE 中的嵌入矩阵），在保护客户端本地敏感知识图谱的同时，协作训练全局知识图谱嵌入模型，如图 1 所示。

然而，越来越多的人达成共识：隐私信息仍可能通过交换的模型参数被泄露，这一点在其他联邦模型训练（例如联邦神经网络训练）中已有见证 [16, 26, 29]。请参阅附录 C 以了解更多

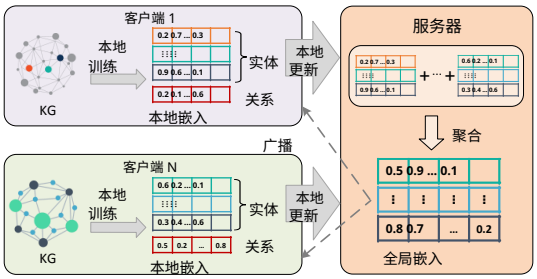

图 1：联邦知识图谱嵌入训练。

相关工作。尽管FKGE 在涉及敏感知识图谱的各个领域中的部署日益增多，但其隐私漏洞仍未得到探索，这需要对以下关键问题进行深入研究：Q1.客户端的敏感知识图谱的隐私在多大程度上可以被隐私攻击推断出来？Q2. 如何通过严格的隐私保证来防御此类攻击?

因此，本文旨在通过首次对联邦知识图谱嵌入中的隐私威胁开展全面研究，填补这一空白，从隐私攻击和防御两个角度显著加深我们对此问题的理解。所提出的攻击和防御机制针对联邦知识图谱嵌入的独特属性量身设计，因此不同于文献中针对神经网络等以往研究模型设计的机制。

攻击。 我们通过成员推断攻击（MIAs）量化联邦知识图谱嵌入的隐私风险，以回答问题 Q1。具体而言，我们提出了三种新型成员推断攻击，旨在推断目标知识图谱三元组是否存在于受害客户端的训练知识图谱中，包括服务器发起的推断攻击、客户端发起的被动推断攻击和客户端发起的主动推断攻击。为全面考察联邦知识图谱嵌入的隐私脆弱性，这些攻击分别针对不同的攻击设置，对手在以下方面有所不同：1）身份：服务器或客户端；2）能力：被动遵循标准的联邦知识图谱嵌入训练，或主动偏离标准流程以获得更强的攻击能力；3）辅助数据集：是否具备辅助知识图谱数据集。我们的攻击能够以较高的 F1分数成功推断出受害客户端中的目标知识图谱三元组，这表明若不采取进一步防御，联邦知识图谱嵌入可能会遭受严重的隐私泄露。此外，我们在联邦知识图谱嵌入训练过程的不同轮次发起攻击，并发现攻击效果会动态变化。这一现象表明，防御机制应根据情况调整防御力度，以更好地权衡防御与效用。

防御。 为回答问题 Q2，我们提出首个针对 FKGE 的严格差分隐私 [11] 保护方案，以降低 FKGE 的隐私风险。差分隐私 [11] 是一种严格的数学概念，可用于防御具有任意先验知识的对手。差分隐私随机梯度下降 [3] 是适用于 FKGE 的一种候选技术，我们将其直接应用于 FKGE，作为基线方法。然而，该基线方法无法维持良好的隐私-效用权衡，根本原因既在于模型规模庞大，也在于未能利用 FKGE 的实体绑定稀疏梯度特性。众所周知，将差分隐私随机梯度下降应用于神经网络时，它只能在小规模模型上 [3, 31, 48] 提供可接受的准确率，却难以为大规模模型维持良好的效用。不幸的是，知识图谱嵌入模型通常规模非常庞大。例如，在嵌入维度为 128 的 Fb15k- 237 数据集 [37] 上，TransE 模型 [6] 约有两百万个模型参数。尽管模型参数规模庞大，但由于实体绑定稀疏梯度特性，FKGE 中随机梯度的“活跃规模”通常要小得多。这一特性是 FKGE 所特有的，而一般的神经网络模型并不具备。

受上述两项局限的启发，我们提出*DP-FLames*（**D**ifferentially **P**rivate **F**ederated **k**now**L**edge **g**raph **e**mbedding with private **s**election，带私有选择的差分隐私联邦知识图谱嵌入），利用联邦知识图谱嵌入的实体绑定稀疏梯度特性，以实现更好的隐私-效用权衡。具体而言，我们设计了私有活跃梯度元素选择机制，避免扰动会引入大量噪声的非活跃梯度元素，从而将“活跃规模”保持在相对较小的范围内。同时，我们确保选择过程本身满足差分隐私，并以紧致的方式核算其隐私损失，避免为保护选择过程而额外添加的噪声超过梯度规模缩减所节省的噪声。此外，鉴于联邦知识图谱嵌入训练过程中攻击强度各不相同，我们设计了自适应隐私预算分配策略，以相应调整防御力度。最后，尽管已有研究探讨了在集中式 [17]和联邦式设置 [30],下通过差分隐私保护知识图谱嵌入的隐私，但现有尝试的隐私分析不够准确，导致其防御机制无法提供严格的差分隐私保证。更多讨论请参阅附录 H.3。此外，与现有工作相比，*DP-FLames*采用了截然不同的差分隐私机制设计，并进行了更深入的隐私分析。

总而言之，我们作出以下贡献：

- 我们针对联邦知识图谱嵌入中的隐私威胁开展了首项全面研究，涵盖攻击和防御两个方面。
- 我们提出了三种新的推理攻击，以全面量化联邦知识图谱嵌入的隐私风险，这些攻击针对不同的攻击设置，涵盖对手的身份、能力和辅助信息。
- 我们提出了首个严谨的差分隐私 FKDE机制。该机制利用 FKGE 训练的实体绑定稀疏梯度属性，实现了更优的隐私-效用权衡；同时，通过引入最先进的私有选择技术，配备了更严格的隐私损失会计器。
- 我们在三个真实世界的知识图谱数据集和四种联邦知识图谱嵌入模型上进行了大量实验评估，以证明所提出的攻击会对联邦知识图谱嵌入构成重大隐私威胁，并验证 *DP-FLames* 能够有效保护联邦知识图谱嵌入免受推理攻击，同时保持良好的效用。

2 预备知识

2.1 联邦知识图谱嵌入

我们首先介绍知识图谱和知识图谱嵌入模型的形式化定义，如下所示，

定义 *2.1*。**(知识图谱G 及其嵌入)**。知识图谱 G 定义为 G ={ E、R、T }，其中 E、R、T 分别是实体、关系和三元组的集合。一个三元组

表示一个事实 $(h, r, t) \in \mathrm{E} \times \mathrm{R} \times \mathrm{E}$，这意味着头实体 h和尾实体 t 通过关系 r相连。知识图谱嵌入模型将实体 h、t和关系 r表示为嵌入空间中的稠密向量 $\mathbf{h}$、$\mathbf{t}$、$\mathbf{r}$，其逐三元组损失函数如下：

$$\mathcal{L}(\mathbf{h},\mathbf{r},\mathbf{t}) = -\log\sigma(f_r(\mathbf{h},\mathbf{t}) - \gamma) - \sum_{i=1}^{n} p_i \log\sigma(\gamma - f_r(\mathbf{h},\mathbf{t}'_i))),$$

其中，$f_r\mathbf{h}$，$\mathbf{t}$是评分函数，用于衡量$(\mathbf{h}, \mathbf{r}, \mathbf{t})$的合理性；$\gamma$是间隔，$(\mathbf{h}, \mathbf{r}, \mathbf{t}\prime) \notin \mathrm{T}$是通过将原始尾实体替换为随机实体而生成的负三元组，n是负三元组的数量，p_i是权重。

常见的知识图谱嵌入模型包括 TransE [6],RotateE [35], DisMult [47], 和 ComplEx [38], ，它们的评分函数形式各不相同。评分较高的三元组更可能有效。这些模型的评分函数 $f_r(h, t)$ 在附录 A 中进行了描述。

联邦知识图谱嵌入学习可以利用各参与知识图谱之间的重叠实体，进一步改善它们的嵌入；也就是说，一个知识图谱中的实体通常与其他知识图谱中的实体重叠 [9, 30]。联邦知识图谱嵌入的正式定义如下。

定义 *2.2.* **(联邦知识图谱嵌入)** 有 m 个知识图谱 $\{\mathrm{G}_i\}_{i=1}^m$，它们的实体集合可能存在重叠，且每个知识图谱由一个客户端持有。在联邦知识图谱嵌入训练的第 κ_{th} 轮中，第 i 个客户端在其本地知识图谱 G_i、本地关系嵌入 $\mathbf{R}^i_{\kappa-1}$ 和本地实体嵌入 $\mathbf{E}^i_{\kappa-1}$ 上执行若干次本地更新步骤，然后将更新后的本地嵌入 $\mathbf{E}^i_\kappa$ 发送给服务器；服务器接收来自 $i \in \{1, ..., m\}$ 的 $\{\mathbf{E}^i_\kappa\}_{i=1}^m$，并进行聚合，然后将全局嵌入 $\mathbf{E}_\kappa$广播给所有客户端。这样的通信轮次会重复进行，直到收敛。

2.2 差分隐私

作为一种应对具有任意先验知识的对手的数学框架，差分隐私 [11] 能够定量分析并界定数据依赖算法的隐私损失。

定义 *2.3.* **(差分隐私** [11]) 若随机算法 M 满足以下条件，则称其满足 (ϵ, δ)-差分隐私：对于任意两个仅相差一个条目的相邻数据库D 和 D'，以及所有可能的结果集合，均有 O $\in$ *Range* [$D \in$] $\leq$ e^ϵ [$D \in$] $+ \delta$.
(M): Pr M( ) O Pr M(') O

差分隐私确保即使拥有任意辅助信息，对手也无法可靠地推断某个特定样本是否在数据集中。Mironov引入了 Rényi 差分隐私 (RDP) [27] 的概念，该概念基于 α-Rényi散-度 $\mathbb{D}_\alpha(\cdot\|\cdot)$，从而实现更紧的隐私损失会计器。我们在下文介绍近似 Rényi 差分隐私的定义，可以将其视为对 [27]中引入的原始 Rényi 差分隐私的推广。

定义 *2.4.* **(近似 Rényi差分隐私**[54])。随机化算法 M 满足 δ-近似-(α，$\epsilon(\alpha)$)- Rényi 差分隐私：对于任意两个相邻数据库 D 和 $D\prime$，存在事件 E（取决于 M(D)）和 $E\prime$（取决于M($D\prime$)），使得 $\Pr[E] > 1-\delta \wedge \Pr[E\prime] > 1-\delta$，并且 $\mathbb{D}_\alpha(\mathrm{M}(D)|E \| \mathrm{M}(D\prime)|E\prime) \leq \epsilon(\alpha)$ 对任意 $\alpha \geq 1$ 均成立，其中

$$\mathbb{D}_\alpha(\mathcal{M}(D)\|\mathcal{M}(D')) := \frac{1}{\alpha-1}\log\mathbb{E}_{o\sim\mathcal{M}(D)}[(\frac{\Pr[\mathcal{M}(D)=o]}{\Pr[\mathcal{M}(D')=o]})^\alpha].$$

当 $\delta = 0$ 时，上述 δ-近似-(α，ϵ)-RDP 可化简为原始的 (α，ϵ)-RDP。其他性质见附录 B。

3 攻击

为量化 vanillaFKGE 面临的隐私威胁，我们提出了三种攻击，用于推断受害客户端持有的特定知识图谱三元组是否存在。我们首先描述攻击设置，然后介绍这些新攻击的具体机制。为避免混淆，重要符号列于附录 D 的表 5 中。

3.1 攻击设置

对手目标。 FKGE 三元组推断攻击 A 的目标，是在给定的 FKGE 学习过程中，判断目标知识图谱三元组 (h, r, t) 是否属于受害客户端的训练集，即对手使用 FKGE 三元组推断攻击 A(h, r, t)，判断 (h, r, t)存在 还是不存在 于该客户端的知识图谱训练集中。由于三元组包含受害客户端的敏感信息，A(h, r, t) 成功实施将对受害客户端造成严重的隐私威胁。

对手的背景信息与能力。 为了全面考察 FKGE 中所有潜在的隐私威胁发起者，我们将服务器和客户端都视为可能的对手，并根据其在 FKGE 中扮演的角色，描述其背景知识和能力。

• 服务器作为对手：对手拥有完整的实体集 E（包括受害客户端的 E_v）以及所有客户端（包括受害客户端的 $\mathbf{E}_v$）定期上传的实体嵌入矩阵 $\mathbf{E}$ 。此外，为了了解这种设置下推断能力的上限，我们尽可能为对手提供充足的辅助信息。因此，对手可以访问来自与 FKGE 学习相同知识领域的辅助知识图谱数据集 D_{aux} 。在实践中，这类辅助知识图谱数据集可以来自公开资源（例如维基百科），也可以基于经验常识构建（例如，患者与某种疾病 之间的关系是确诊患有）。我们假设服务器广播未经修改的正常全局嵌入。

• 客户端作为对手：对手拥有其本地知识图谱数据集，以及服务器在每个通信轮次广播的全局实体嵌入矩阵。恶意客户端可以遵循标准 FKGE 协议，将正常更新后的嵌入发送给服务器（即被动攻击）；也可以篡改嵌入，以便更好地推断受害客户端的三元组（即主动攻击）。

表1：所提出的三种攻击概述。

攻击类型	身份	能力	辅助数据集
$\mathcal{A}_{SI}$	服务器	被动	✓
$\mathcal{A}_{CIP}$	客户端	被动	-
$\mathcal{A}_{CIA}$	客户端	主动	-

攻击分类。表1根据身份、能力和辅助信息等方面的不同攻击设置，总结了三种攻击，分别称为服务器发起的推断攻击（A_{SI}）、客户端发起的被动推断攻击（A_{CIP}）和客户端发起的主动推断攻击（A_{CIA}）。

3.2 服务器发起的推理攻击

在服务器发起的推断攻击中，给定辅助数据集，对手只需推断目标头部和尾部实体之间关系的存在，随后推断三元组（即，特定的

（已推断关系的类型）可以通过查询辅助数据集来推断。该攻击分三个步骤进行。

步骤 1：计算实体嵌入。在攻击轮次期间，恶意服务器会从受害客户端接收完整的实体嵌入矩阵 $\mathbf{E}_v$。它根据知识图谱嵌入模型指定的嵌入关系，枚举所有可能的关系嵌入。以 TransE 为例，它通过 $\mathbf{r} = \mathbf{t} - \mathbf{h}$ 计算所有可能的 $\mathbf{r}$。更一般地，计算方式如下：

$$\mathbf{R}_{SI} = \{\mathbf{r}_i\}_{i=1}^{n_v^2 - n_v}, \mathbf{r} = f_{e-r}(\mathbf{e}_j, \mathbf{e}_k), 0 \le j, k < n_v,$$

其中，n_v 是受害客户端中的实体数量，f_{e-r} 是实体和关系的表示。$\mathbf{e}_j, \mathbf{e}_k$ 是受害客户端 $\mathbf{E}_v$ 中的实体嵌入。当 f_{e-r} 过于复杂、不适合进行 $\mathbf{r}$ 转换时，为简化计算，我们使用 $\mathbf{r}$ 通过 $\mathbf{r} = \mathbf{t} - \mathbf{h}$ 进行近似。

步骤 2：关系推断。由于并非所有枚举得到的关系嵌入都对应于真实存在的关系，因此我们还需要识别“真实”关系。我们的核心直觉是，真实关系的嵌入往往会在嵌入空间中集中在一起，而伪关系的嵌入通常较为分散。因此，我们将关系嵌入聚类为 $2n_r$ 个簇，并找出关系嵌入中高度集中的簇中心。给定一个用于推断关系是否存在的候选三元组 (h, r, t)，如果对应的 $f_{e-r}(\mathbf{h},\mathbf{t})$ 属于某个高度集中的簇，我们便认为存在关系 (h, x, t)，其具体关系类型 x 未知。

步骤 3：三元组推断。借助辅助数据集 D_{aux}，对手可以推断关系 x 的具体类型。也就是说，对手会搜索辅助三元组 $(\mathrm{E}_1, r_{aux}, \mathrm{E}_2)$，利用 $h \in \mathrm{E}_1$ 和 $t \in \mathrm{E}_2$ 找到与 x 匹配的关系类型。一旦对手将 r_{aux} 与 x 匹配，即 $x = r_{aux}$，便成功推断出目标三元组 (h, r, t)。

3.3 客户端发起的被动推断攻击

在客户端发起的被动推断攻击中，对手是一个“好奇但诚实”的客户端，它会推断受害客户端的特定三元组，但仍遵循标准的联邦知识图谱嵌入训练过程。

步骤 1：计算实体嵌入。对手首先比较自己上传的嵌入与服务器广播回来的全局实体嵌入，以识别重叠实体。由于聚合时只会更新重叠实体的嵌入，对手可以利用这一点识别与受害客户端重叠的实体。然后，对手根据 $\mathbf{E}_{CIP}$，$\mathbf{E}_{CIP} \leftarrow (N \times \mathbf{E}_b - \mathbf{E}_u)/(N-1)$，从全局嵌入中提取目标实体的嵌入。其中，$\mathbf{E}_b$ 是服务器广播回来的全局实体嵌入，$\mathbf{E}_u$ 是对手上传的嵌入，N 是联邦知识图谱嵌入中的客户端数量。

步骤 2：计算近似关系嵌入。接下来，对手使用其本地关系嵌入近似目标三元组的关系嵌入。其直觉在于，随着联邦知识图谱嵌入训练的推进，本地关系嵌入将逐渐包含更多受害者知识图谱的关系信息，而对手可以利用这些信息，以越来越高的接近度近似目标嵌入。

步骤 3：三元组推断。利用目标三元组的实体嵌入和关系嵌入，对手通过对评分函数的值设置阈值来推断该三元组是否存在，如下所示：

$$\mathcal{A}_{CIP}(h, r, t) = \begin{cases} exist, & f_r(\mathbf{h}_1, \mathbf{t}_1)/f_r(\mathbf{h}_2, \mathbf{t}_2) \ge \tau_{CIP}, \\ non-exist, & f_r(\mathbf{h}_1, \mathbf{t}_1)/f_r(\mathbf{h}_2, \mathbf{t}_2) < \tau_{CIP}, \end{cases}$$

其中 $\mathbf{h}_1$、$\mathbf{t}_1$ 是 h、t 在攻击模型的嵌入 $\mathbf{E}_{CIP}$ 中的对应嵌入，$\mathbf{h}_2$、$\mathbf{t}_2$ 是 h、t 在对手的本地嵌入 $\mathbf{E}_u$ 中的对应嵌入，而 τ_{CIP} 是可调的决策阈值。A(h、r、t)表示对手对样本（h、r、t）的预测。其原理是，评分函数量化了三元组的可信度，因此较低的评分表示目标三元组更有可能存在。

3.4 客户端发起的主动推断攻击

在客户端发起的主动推断攻击中，对手是一个恶意客户端，通过修改其本地嵌入来偏离标准训练过程，以获得更强的攻击能力。

步骤 1：反转实体嵌入。在攻击轮次中，对手反转目标三元组（h，r，t）的尾实体嵌入，故意将其分数提高到 s_1。经过若干轮联邦知识图谱嵌入训练后，如果三元组（h，r，t）确实存在于受害客户端的知识图谱中，其分数将降低至 s_2。

步骤 2：三元组推断。对手通过对 s_1/s_2 的分数比值进行阈值判定，推断目标三元组是否存在。

$$\mathcal{A}_{CIA}(h, r, t) = \begin{cases} exist, & s_1/s_2 \ge \tau_{CIA}, \\ non-exist, & s_1/s_2 < \tau_{CIA}, \end{cases}$$

其中 τ_{CIA} 是可调决策阈值。

我们将在附录H.1 中进一步讨论上述攻击的局限性，以及服务器与客户端串通时将会发生什么。

4 防御机制

在本节中，我们提出*DP-FLames*，为联邦知识图谱嵌入模型提供严格的隐私保护。我们首先介绍一种基线方法，即直接将差分隐私随机梯度下降应用于联邦知识图谱嵌入，并指出以下挑战：1）如何避免隐私保护过程中模型效用显著下降？2）鉴于隐私攻击的攻击强度各不相同，如何在整个训练过程中动态调整保护强度？*DP-FLames* 正是受到这两项挑战的启发，旨在同时实现严格的差分隐私保护和良好的模型效用。

4.1 基线防御：用于联邦知识图谱嵌入的差分隐私随机梯度下降

为知识图谱嵌入确定差分隐私粒度。由于现有的两种图数据邻接定义，即节点邻接和边邻接，不适用于知识图谱嵌入模型（如附录 H.2 所述），因此我们引入了以下针对知识图谱嵌入差分隐私保护的知识图谱三元组邻接定义。

定义 *4.1*。（知识图谱嵌入的三元组级邻接数据集）。如果 $E_1 = E2$、$R_1 = R2$，并且 $\exists (h, r, t) \in T_1 \cup T2$，使得 $(T1 \mid T2) \cup (T2 \mid T1) = \{(h, r, t)\}$，则两个知识图谱 $G_1 = \{E1, R1, T1\}$ 和 $G_2 = \{E2, R2, T2\}$ 互为邻接数据集。

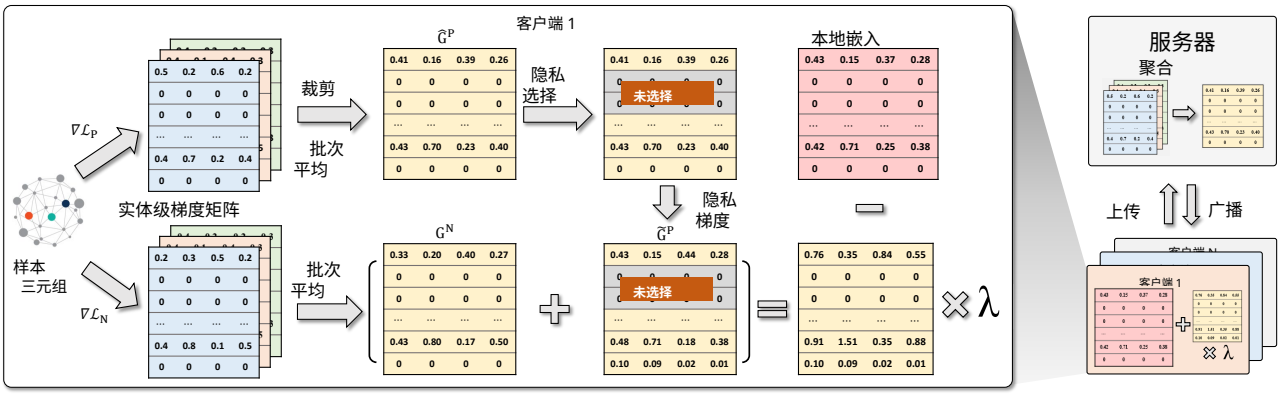

图 2：*DP-FLames* 概览。

基线算法。 我们通过在联邦学习设置下将差分隐私随机梯度下降应用于知识图谱嵌入，展示基线防御机制；该机制基于差分隐私深度学习领域中著名的差分隐私随机梯度下降方法 [3]。算法 1 未 蓝色线条 对基线算法进行了总结。它在客户端侧包含三个关键步骤：1）采样一个随机批次，其中包含抽样概率 $q = B/N$ 并计算每个样本的梯度$\mathbf{G}^i = \nabla \mathbf{L}(\mathbf{h}, \mathbf{r}, \mathbf{t})$（算法 1 第 11-12 行）；2）通过将每个样本梯度的 ℓ_2 范数裁剪到裁剪阈值 C_1，限制任意样本的影响，即 $\hat{\mathbf{G}}^i = \mathbf{G}^i / \max\{1, \frac{\|\mathbf{G}^i\|_2}{C_1}\}$（算法 1 第 13 行）；3）注入按噪声乘数 σ 校准的高斯噪声（算法 1 第 16 行）。随后，客户端定期将本地更新上传到服务器，并接收服务器聚合后的全局广播。

算法 1 *DP-FLames*

输入：三元组集$\{(h, r, t)\}$，损失函数 $\mathrm{L}^{p/n}(\mathrm{h,r,t})$。参数：学习率 λ，批次大小 B，噪声分配参数 σ, η, Δ，裁剪界限 C_1, C_2，隐私选择参数 $\epsilon_r, \sigma_p, \delta_t$ { 蓝色线条表示 DP-FLames 的高级算法设计。} 服务器端流程：1：为所有客户端对齐实体2：初始化并将E_0 广播给所有客户端 3：对于轮次 $\kappa \in [K]$ 执行4：从每个客户端收集 E^c_κ c5：通过对重叠实体的嵌入进行平均来计算 $\mathrm{E}_{\kappa+1}$ 6：将 $\mathrm{E}_{\kappa+1}$ 广播给所有客户端 7：结束循环 客户端 (c) 端流程：8：对于轮次 $\kappa \in [K]$ 执行9：从服务器接收 E_κ 10：对于 $\tau \in [T]$ 执行11：以抽样概率 $q = B/N$抽取随机样本集 B 12：对于每个 $i \in [B]$，计算$\mathrm{G}^p_i \leftarrow \nabla \mathrm{L}^p(\mathrm{h}_i, \mathrm{r}_i, \mathrm{t}_i)$13：$\hat{}\mathrm{G}^p_i \leftarrow \mathrm{G}^p_i / \max\{1, \frac{\|\mathrm{G}^i_p\|_2}{C_1}\}$ 14：对于 $\hat{}\mathrm{G}^p_i$ 中的行，计算 $\mathrm{G}^p_i(j) \leftarrow \hat{}\mathrm{G}^p_i(j) / \max\{1, \|\hat{}\mathrm{G}^p_i(j)\|_2\}$ }15：I ←隐私选择$(\sum_i \mathrm{G}^p_i)$16：$\tilde{}\mathrm{G}^p(\mathrm{I}) \leftarrow \frac{1}{B}(\sum_i \mathrm{G}^p_i(\mathrm{I}) + \mathrm{N}(0, \frac{C_2}{\sigma^2} C_1^2 I_{k\times k}))$17：$\mathrm{G}^p \leftarrow \frac{1}{N}\sum_i \mathrm{G}^p_i$ 18：$\mathrm{G}^n \leftarrow \frac{1}{N}\sum_i \nabla \mathrm{L}^n(\mathrm{h}'_i, \mathrm{r}'_i, \mathrm{t}'_i)$~19：$\mathrm{E}^c_{\kappa_{\tau+1}} \leftarrow \mathrm{E}^c_{\kappa_\tau} - \lambda(\mathrm{G}^p + \mathrm{G}^n)$；$\mathrm{R}^c_{\kappa_{\tau+1}} \leftarrow \mathrm{R}^c_{\kappa_\tau} - \lambda(\mathrm{G}^p + \mathrm{G}^n)$ 20：结束循环21：将 $\mathrm{E}^c_{\kappa+1} \leftarrow \mathrm{E}^c_{\kappa_T}$上传至服务器 22：如果 κ 模验证间隔 $= 0 \wedge \mathrm{MRR}_t - \mathrm{MRR}_{t-1} < \Delta$ 则 23：$\sigma = \eta \cdot \sigma$ { 此行指自适应隐私预算分配} 24：结束条件 ~~25：~~结束循环 输出：E_K，$\{\mathrm{E}^c_K\}^m_{c=1}$

基线方法的局限性分析。 局限性源于模型规模庞大，以及未能利用联邦知识图谱嵌入的实体绑定稀疏梯度特性。知识图谱嵌入模型通常规模很大。例如，Fb15k-237 数据集上的 TransE 模型有两百万个模型参数。尽管模型参数规模庞大，但由于实体绑定稀疏梯度特性，联邦知识图谱嵌入模型的随机梯度“活跃规模”通常要小得多。实体绑定稀疏梯度特性 源自知识图谱嵌入模型的损失函数 [35] $\mathbf{L}(\mathbf{h,r,t})$（如第 2.1 节所示）。每个三元组的实体嵌入梯度仅将$\mathbf{h}$ 和$\mathbf{t}$ 作为活跃元素，而所有其他元素都保持非活跃状态。对于 Fb15k-237 数据集上批次大小为 64 的 TransE 模型，只有 1% 的梯度元素处于活跃状态。直观而言，以差分隐私的方式让 99% 的零梯度元素保持为零，可以避免大量不必要的扰动并提升效用。

4.2 高级防御：*DP-FLames*

有多种技术可用于筛选出活跃梯度，例如稀疏向量技术（SVT） [13, 23, 25] 和 k次指数机制 [10]。然而，这些技术无法提供紧致的隐私损失会计器。我们基于以下两个原因，利用最先进的带有提议-测试-发布（PTR）的嘈杂最大值报告机制，对数据依赖的前k 个隐私选择 [54]（如算法 2 所示）：（1）PTR 一次性隐私选择 k 个向量，而 SVT 和指数机制需要进行 k 次组合并消耗隐私预算；（2）尽管 PTR 中的测试存在失败的概率，这本应是该方案的主要瓶颈，但隐私损失处于可控范围内，因为在这种情况下，被丢弃的隐私梯度步骤不会额外消耗隐私预算。下面我们介绍隐私选择（算法 2）的详细信息。

数据相关的 k选择。 第一步是确定活跃梯度行的数量 k。假设梯度矩阵 G 中有 n 行，每行对应其实体嵌入的梯度向量。只有与采样批次对应的梯度向量才是活跃的。然而，我们无法预先设定一个统一的k，因为每个批次涉及的实体数量都不同。因此，我们采用数据相关的方法来确定 k。具体而言，我们首先按照经 C_2裁剪的 ℓ_2 范数对 G 中的行进行降序排序（算法 1 第 14 行）（第 1 行），然后利用 Gumbel 噪声 [10] 寻找任意两个相邻向量之间的最大间隔（第 2 行）。第 2 行中的正则化项 $r(j)$ 编码了实际的 k 应位于 $[B, 2B]$ 范围内这一先验知识，其形式如下，

$$r(j) = \begin{cases} c, & \text{if } j \in [B, 2B], \\ -\infty, & \text{if } j \notin [B, 2B], \end{cases}$$

其中，c 是任意常数，可根据额外的先验知识进一步缩小。$\|G(k)\|_2$ 与 $\|G(k+1)\|_2$ 之间最大的 ℓ_2 范数差记为 d_k，用于后续的 PTR 过程（第 3 行）。

算法 2隐私选择

输入： 梯度矩阵 G，裁剪界限 C_2，Rényi 差分隐私参数 $\sigma_r, \sigma_p, \delta_t$

1：按 ℓ_2 范数对 G 中的行降序排序：$\|G(1)\|_2$，$\|G(2)\|_2$，$\ldots, \|G(n)\|_2$

2: $k \leftarrow Argmax_{j\in[n-1}\{\|G(j)\|_2 - \|G(j+1)\|_2 + r(j) + \text{Gumbel}(2C_2\sigma_r)$

3: $d_k \leftarrow \|G(k)\|_2 - \|G(k+1)\|_2$

4: ^ $d_k \leftarrow \max\{C_2, d_k\} + \text{N}(0, \sigma_p^2 C_2^2) - \sigma_p C_2\sqrt{2\log(1/\delta_t)}$

5: 如果 $\hat{d}_k > C_2$ **那么**

6: 我 $\leftarrow \{i_{(1)}, \ldots, i_{(k)}\}$

7: **否则**

8: $\mathcal{I} \leftarrow \perp$

9: **结束 if**

输出： 前 k 个梯度的索引集合 I

主动梯度索引发布。 确定 k和 d_k后，我们隐私发布排名前k的梯度向量索引。我们不再调用 SVT 或指数机制 k次，而是采用基于 PTR 的方法 [12, 54]，一次性发布 k个索引。首先，我们对 d_k和 C_2 的最大值添加高斯噪声，将其作为 PTR 框架中的“建议”界限（第 4 行）。然后，我们以隐私方式“测试”^ d_k是否为有效的上界（第 5 行）。如果通过测试，则“发布”^选定的排名前k的向量索引集 d_k；否则不报告任何内容（第 6–8 行）。

负采样。 为了充分利用梯度的稀疏性，我们对自对抗负采样进行了修改。原始方法仅将有效三元组（h, r, t）的尾实体 t替换为随机的 t'，作为负样本 [35]。负样本集通常由数百个负三元组组成，这使得梯度变得更加稠密。相反，我们生成完全随机的负样本（$h' r' t'$），而不是（h, r, t'），从而无需扰动即可释放负梯度。然而，这样做会损害模型效用，因为随机生成的头实体-关系对通常在语义上没有意义。有两种解决方法：引入来自公共数据集的真实头实体-关系对，或者从隐私数据集与差分隐私数据合成技术[15, 24, 52]。为简单起见，我们采用第一种方法。损失函数可以分为正损失和负损失两部分，正损失部分的梯度仍然高度稀疏。

$$\mathcal{L}_p(\mathbf{h}, \mathbf{r}, \mathbf{t}) = -\log\sigma(f_r(\mathbf{h}, \mathbf{t}) - \gamma), \tag{1}$$

$$\mathcal{L}_n(\mathbf{h}, \mathbf{r}, \mathbf{t}) = -\sum_{i=1}^{n} p(\mathbf{h}, \mathbf{r}, \mathbf{t}')\log\sigma(\gamma - f_r(\mathbf{h}, \mathbf{t}'_i))). \tag{2}$$

公式 (1) 的梯度由 *DP-FLames* 处理，而公式 (2) 的梯度则由原始随机梯度下降处理（算法 1 第 18 行）。

4.3 自适应隐私预算分配

在我们的主动攻击实验中，成功率在训练的最初几轮达到峰值，并随着模型收敛而下降。这促使我们在最初几轮提供更强的防御，并在后续轮次逐步降低防御，以获得更好的效用。

我们提出在训练过程期间根据验证准确率自适应地调整噪声尺度 σ（算法1第23-25行）。具体而言，我们将初始的 σ 设置得相对较大，以防御推理攻击，并每几轮次计算一次验证准确率。如果验证准确率不再提升，我们将 σ与系数 $\eta < 1$ 相乘来降低其值，直至收敛。更多细节请参见附录 E。

4.4 隐私分析

在本部分中，我们针对*DP-FLames* 提供隐私分析。我们首先给出每次迭代的隐私分析，其中包括隐私选择部分和隐私梯度部分的 Rényi 差分隐私隐私会计器。然后，我们对 T次迭代中的每次迭代 Rényi 差分隐私隐私会计器进行组合，并将总体Rényi 差分隐私结果转换为差分隐私保证。所有证明均推迟至附录 F。

隐私梯度分析。 RDP 记账器能够紧密追踪隐私损失，已被许多主流软件包用于实现差分隐私随机梯度下降，例如 [14, 36, 41]。下文的引理 4.2 给出了隐私梯度部分每次迭代基于 RDP 记账器的隐私保证，其中包括算法 1 的第 11、13、16 行。

引理 *4.2*。算法 *1* 中的隐私梯度步骤满足 *(α, ϵg(α))-RDP*，其中 $\epsilon_g(\alpha) = \frac{2q^2\alpha}{\sigma}$ *for* $1 < \alpha \le \min(\frac{1}{2}\sigma^2 L - 2\log\sigma, 1/2\sigma^2 L^2 - \log 5 - 2\log L + \frac{1}{2\sigma^2}\log(q\alpha)/(\ )\ )$，其中采样率为 $q = B/N$，且 $L = \log(1 + q^{1}(\alpha - 1))$。

隐私选择分析。 隐私选择步骤分为两个阶段：Report-Noisy-Max 和 Propose-Test-Release，可使用近似 RDP进行分析。定理 4.3 给出了隐私选择部分在进行子采样时基于 RDP 记账器的每次迭代隐私保证，该部分也见算法 2 和算法 1 的第 11 行。

定理4.3。 算法 *2* 中的隐私选择在采样率 q满足 δ_t-近似-（α, ϵ_s（α）*)-Rényi* 差分隐私，其中

$$\epsilon_s(\alpha) = \frac{1}{\alpha - 1}\log(p^2\binom{\alpha}{2}\min\{4(e^{\epsilon(2)} - 1), e^{\epsilon(2)}\min\{2, (e^{\epsilon(\infty)} - 1)^2\}\}$$

$$+ \sum_{j=3}^{\alpha} q^j\binom{\alpha}{j}e^{(j-1)\epsilon(j)}\min\{2, (e^{\epsilon(\infty)} - 1)^j\} + 1) \text{ and } \epsilon(\alpha) = \frac{\alpha}{8\sigma_r^2} + \frac{\alpha}{2\sigma_p^2}.$$

表 2：攻击的 F1分数。

数据集	客户端发起的被动推理				客户端发起的主动推理				服务器发起的推理	
	TransE	RotatE	DistMult	ComplEx	TransE	RotatE	DistMult	ComplEx	TransE	RotatE
Fb15k-237	0.724	0.8211	0.7809	0.7749	0.8779	0.8403	0.7346	0.7949	0.878	0.7827
nell-955	0.8243	0.8728	0.8686	0.8429	0.8817	0.8462	0.7337	0.7848	0.9023	0.748

客户端需要在算法1中进行多次迭代，以训练本地嵌入。在 Rényi 差分隐私下，T 次迭代组合的隐私保证如下方定理 4.4 所示。

定理 *4.4*。算法 *1* 中 T 次更新的组合满足 $T\cdot\delta_t$-近似-α, $\sum_{i=1}^{T}(\mathbb{I}_i\cdot\epsilon_g^i(\alpha)+\epsilon_s^i(\alpha)))$-RDP 和 $(\sum_{i=1}^{T}(\mathbb{I}_i\cdot\ \epsilon_g^i(\alpha)\ +\epsilon_s^i(\alpha))\ +log(1/\delta)\,\alpha-1, T\cdot\delta_t+\delta)$-DP，其中 $\mathbb{I}_i$ 表示 *PTR* ~~中的“~~测试”是否通过，α 满足引理 *4.2* 中的约束，$\epsilon^i(\alpha)$ 是第 i_{th} 次迭代的隐私预算。

5 评估

在本节中，我们通过实验回答三个研究问题：**RQ1**：所提出的攻击对联邦知识图谱嵌入是否有效？**RQ2**：*DP-FLames* 能否缓解这些隐私威胁？**RQ3**：隐私-效用权衡表现如何？

5.1 实验设置

数据集。我们使用了两个真实世界的知识图谱基准数据集 FB15k-237 [37] 和 NELL-995 [46]。这些数据集的详情见附录 G.1。

目标联邦知识图谱嵌入模型。我们采用了四种常见的知识图谱嵌入模型：TransE [6], RotateE [35], DisMult [47] 和 ComplEx [38]。它们的评分函数见表 4。随后，联邦知识图谱嵌入按照以下设置协同训练所有四种知识图谱嵌入模型：我们随机生成 1000 个三元组，其中 500 个作为正样本纳入受害者的训练数据，其余作为负样本。通过攻击者能否正确区分正样本和负样本来衡量攻击有效性。超参数设置详情见附录 G.2。

评估指标。我们使用精确率 和召回率的调和平均数（即 F1分数）来衡量攻击成功率。我们报告平均倒数排名（*MRR*）和 *Hits@N*（*Hits@N*）来评估模型性能，这遵循了知识图谱嵌入文献中的常见做法。这些指标的详情见附录G.3。

5.2 攻击评估（RQ1）

所提出的攻击的整体性能总结如下。

表 2. 攻击轮次设为每五轮一次，表中报告了整个训练过程中的最高 F1分数。CIP 在全部四种知识图谱嵌入算法上表现良好，F1分数约为 0.8。CIA 在 TransE 和 RotatE 上通常优于 CIP，F1分数约为 0.85，但在 DisMult 和 ComplEx 上表现不如 CIP。原因在于，DistMult 和 ComplEx 的评分函数决定了三元组对反向尾实体嵌入并不敏感。SI 仅适用于平移距离模型，即 TransE 和 RotatE，因为这些模型中的关系嵌入表示加性平移，并且可以在嵌入空间中很好地聚类。而双线性模型（即 DisMult 和 ComplEx）中的关系嵌入则是乘法运算，几乎无法聚类。

为了展示这些攻击的性能在整个训练过程中的变化情况，我们在图 3 左侧子图中绘制了各攻击在不同轮次的 F1分数。我们以 TransE 和 RotatE 为例，因为三种攻击都适用于这两种模型。可以看到，在训练初期，随着轮次增加，CIP 和 SI 的 F1分数不断上升。在模型训练后期，由于模型逐渐收敛，这两种攻击的分数可能不再上升。CIA 的 F1分数在训练开始时达到峰值，随后随着模型收敛而下降，因为这种攻击关注的是相邻两轮之间的更新。F1分数会随着阈值而变化，因此我们在图 3 右侧子图中绘制相应攻击最高分数（左侧）的 ROC曲线，以展示阈值的影响。

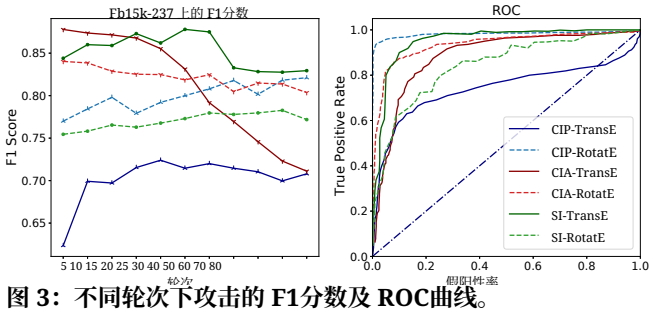

图 3：不同轮次下攻击的 F1分数及 ROC曲线。

5.3 防御评估（RQ2）

随后，我们研究算法 1 中*DP-FLames* 的有效性，并在图 4 中展示结果。实线表示在不同 ϵ 下采用防御时的攻击得分，虚线则表示未采用防御时的得分。采用不同 ϵ 的差分隐私方法均以相同的初始噪声幅度 σ开始，但会在 ϵ 耗尽时的不同轮次结束。我们的防御机制通过大幅降低攻击的 F1分数，有效抵御了这些攻击。用于联邦知识图谱嵌入的基线 DPSGD 通常能取得最低的 F1分数，但代价是效用下降难以接受，后文将对此进行展示。对于 TransE 和 RotateE，*DP-FLames-Adp* 抵御 CIA 的效果优于*DP-FLames*；而在抵御 CIP 和 SI 时，*DP-FLames* 优于 *DP-FLames-Adp*。对于 DistMult 和 ComplEx，*DP-FLames-Adp* 与*DP-FLames* 的防御效果相当。

5.4 模型效用评估（RQ3）

为研究差分隐私防御对模型效用的不利影响，我们将差分隐私模型与非差分隐私模型进行比较，并采用表3所报告的 MRR、Hits@1 和 Hits@10 进行衡量。隐私预算设置为 $(16, 10^{-5})$ 或$(32, 10^{-5})$，这已显示出足够的防御能力。用于联邦知识图谱嵌入的基线 DPSGD 在几乎所有联邦知识图谱嵌入模型上都出现了效用的大幅下降。*DP-FLames* 显示出比基线更高的效用，但仍然不够理想。*DP-FLames-Adp* 取得了

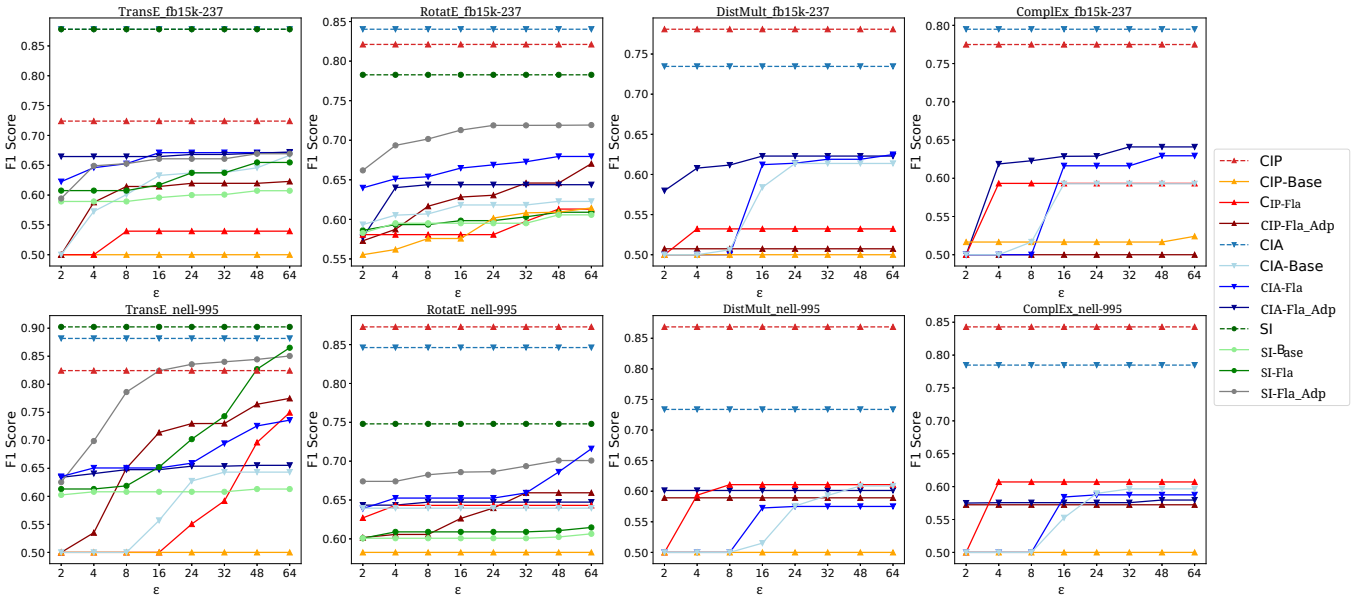

图4：防御后的攻击F1分数。

表3：模型效用。

数据集	设置	TransE			RotatE			DisMult			ComplEx		
		MRR	Hits@1	Hits@10	MRR	Hits@1	命中率@10	MRR	命中率@1	Hits@10	MRR	命中率@1	Hits@10
FB15K-237	FKGE	0.3606	0.2582	0.5720	0.3676	0.2671	0.5745	0.3138	0.2345	0.4718	0.3010	0.2277	0.4444
	基线 (ϵ = 16)	0.1296	0.0492	0.2524	0.0506	0.0197	0.1180	0.0069	0.0060	0.0074	0.0067	0.0058	0.0068
	DP-Flame (ϵ = 16)	0.1843	0.0719	0.3949	0.0901	0.0477	0.1800	0.2342	0.1750	0.3533	0.2638	0.1999	0.3974
	DP-Flame-Adp (ϵ = 16)	0.3193	0.2233	0.5107	0.2946	0.2362	0.4548	0.2764	0.2074	0.4136	0.2874	0.2136	0.4285
	基准（ϵ = 32）	0.1248	0.0430	0.2575	0.0371	0.0138	0.0830	0.0084	0.0068	0.0098	0.0078	0.0066	0.0087
	DP-Flame（ϵ = 32）	0.2964	0.1962	0.4942	0.2740	0.1968	0.4228	0.2881	0.2137	0.4438	0.2784	0.2084	0.4226
	DP-Flame-Adp (ϵ = 32)	0.3239	0.2285	0.5143	0.3044	0.2241	0.4660	0.2728	0.2046	0.4072	0.2850	0.2114	0.4338
NELL-955	FKGE	0.6866	0.6230	0.7976	0.7299	0.6824	0.7993	0.4120	0.3508	0.5183	0.3685	0.3072	0.4852
	基线 (ϵ = 16)	0.1516	0.0035	0.3455	0.0265	0.0052	0.0576	0.0001	0	0	0.0001	0	0
	DP-Flame (ϵ = 16)	0.1419	0	0.3473	0.0151	0.0052	0.0209	0.1975	0.1396	0.2949	0.3340	0.3002	0.3962
	DP-Flame-Adp (ϵ = 16)	0.4687	0.2723	0.7330	0.3822	0.3438	0.4538	0.3745	0.2984	0.5166	0.3620	0.2932	0.4921
	基线 (ϵ = 32)	0.1076	0	0.2618	0.0169	0.0017	0.0297	0.0002	0	0	0.0003	0	0
	DP-Flame (ϵ = 32)	0.3589	0.1658	0.6038	0.1099	0.0803	0.1623	0.3786	0.3054	0.4904	0.3747	0.2967	0.5061
	DP-Flame-Adp (ϵ = 32)	0.4814	0.2757	0.7365	0.5031	0.4555	0.6021	0.3912	0.3141	0.5201	0.3569	0.2740	0.4817

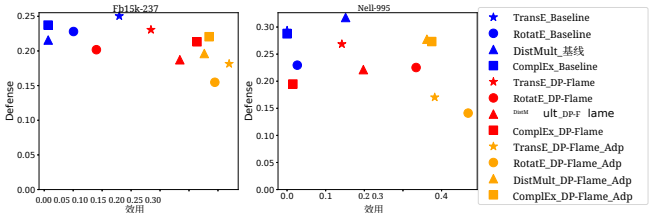

图 5：效用-防御权衡。

在差分隐私模型中效用最佳，在大多数情况下，其效用甚至接近非差分隐私模型，但 NELL-955 上的 RotatE 模型除外。

此外，我们在图 5 中展示了三种防御机制在效用与防御之间的权衡。防御能力通过三种攻击在防御前后平均 F1分数的降幅来衡量，而效用则通过 MRR 来衡量。*DP-FLames* 和 *DP-FLames-Adp* 均实现了更好的效用—防御权衡。*DP-FLames-Adp* 在效用方面始终优于*DP-FLames* 而*DP-FLames* 在 TransE 和 RotatE 上表现出更好的防御效果。

6 结论

本文从攻击和防御两个角度，对联邦知识图谱嵌入（FKGE）的隐私风险进行了全面研究。在攻击方面，我们针对FKGE 提出了三种知识图谱（KG）三元组推断攻击，以揭示其显著的隐私漏洞。在防御方面，提出了*DP-FLames*，为 FKGE 提供严格的差分隐私保护；该方法通过设计隐私主动梯度选择策略，利用 FKGE 的梯度稀疏性。进一步引入自适应隐私预算分配策略，以在整个训练过程中动态调整针对不均衡隐私风险的防御强度。实验结果表明，所提出的防御方法能够以适度的效用下降，有效抵御推理攻击。

致谢

本研究得到国家重点研发计划（2021年YFB3100300）、国家自然科学基金（U20A20178和62072395）以及中国计算机学会-人工智能安全与治理论坛研究基金的资助。

参考文献

[1] 2016。《通用数据保护条例》（GDPR）。（2016年）。https://eur-lex.europa. eu/eli/reg/2016/679/oj [2] 2018。《加州消费者隐私法案》（CCPA）。（2018年）。https://www.oag.ca.gov/ privacy/ccpa [3] Martín Abadi、Andy Chu、Ian J. Goodfellow、H. Brendan McMahan、Ilya Mironov、Kunal Talwar 和 Li Zhang。2016年。差分隐私深度学习。载于*2016* 年 *ACM SIGSAC* 计算机与通信安全会议论文集，奥地利维也纳，*2016*年*10*月*24*日至*28*日，Edgar R. Weippl、Stefan Katzenbeisser、Christopher Kruegel、Andrew C. Myers 和 Shai Halevi（编）。ACM，308–318。[4] Kurt D. Bollacker、Colin Evans、Praveen K. Paritosh、Tim Sturge 和 Jamie Taylor。2008年。Freebase：协作创建的图数据库，用于组织人类知识。载于*ACM SIGMOD* 数据管理国际会议论文集，*SIGMOD 2008*，加拿大不列颠哥伦比亚省温哥华，*2008*年*6*月*10*日至*12*日，Jason Tsong-Li Wang（编）。ACM，1247–1250。[5] Antoine Bordes、Sumit Chopra 和 Jason Weston。2014年。利用子图嵌入进行问答。载于*2014*年自然语言处理实证方法会议论文集，*EMNLP 2014*，*2014*年*10*月*25*日至*29*日，卡塔尔多哈，*ACL* 特殊兴趣组 *SIGDAT* 会议，Alessandro Moschitti、Bo Pang 和 Walter Daelemans（编）。ACL，615–620。[6] Antoine Bordes、Nicolas Usunier、Alberto García-Durán、Jason Weston 和 Oksana Yakhnenko。2013年。用于多关系数据建模的嵌入转换。载于神经信息处理系统进展，Christopher J. C. Burges、Léon Bottou、Zoubin Ghahramani 和 Kilian Q. Weinberger（编）。2787–2795。[7] Nicholas Carlini、Florian Tramèr、Eric Wallace、Matthew Jagielski、Ariel Herbert-Voss、Katherine Lee、Adam Roberts、Tom B. Brown、Dawn Song、Úlfar Erlingsson、Alina Oprea 和 Colin Raffel。2021年。从大型语言模型中提取训练数据。载于第*30*届 *USENIX*安全研讨会，*USENIX*安全 *2021*，*2021*年*8*月*11*日至*13*日，Michael Bailey 和 Rachel Greenstadt（编）。USENIX协会，2633–2650。[8] Andrew Carlson、Justin Betteridge、Bryan Kisiel、Burr Settles、Estevam R. Hruschka Jr. 和 Tom M. Mitchell。2010年。迈向永不停歇的语言学习架构。载于第二十四届 *AAAI* 人工智能会议论文集，*AAAI 2010*，美国佐治亚州亚特兰大，*2010*年*7*月*11*日至*15*日，Maria Fox 和 David Poole（编）。AAAI出版社。[9] Mingyang Chen、Wen Zhang、Zonggang Yuan、Yantao Jia 和 Huajun Chen。2021年。FedE：在联邦设置中嵌入知识图谱。载于*IJCKG'21*：第十届国际知识图谱联合会议，线上会议，泰国，*2021*年*12*月*6*日至*8*日。ACM，80–88。[10] David Durfee 和 Ryan M. Rogers。2019年。实用的差分隐私 Top-k 选择与按需付费式组合。载于神经信息处理系统进展 *32*：*2019*年神经信息处理系统年会，*NeurIPS 2019*，*2019*年*12*月*8*日至*14*日，加拿大不列颠哥伦比亚省温哥华，Hanna M. Wallach、Hugo Larochelle、Alina Beygelzimer、Florence d'Alché-Buc、Emily B. Fox 和 Roman Garnett（编）。3527–3537。[11] Cynthia Dwork。2006年。差分隐私。载于自动机、语言与编程，*2006*年*6*月*14*日，第*33*届国际计算机科学理论研讨会（*ICALP*），意大利威尼斯，*2006*年*7*月，论文集第二部分（计算机科学讲义，第 卷），Michele Bugliesi、Bart Preneel、Vladimiro Sassone 和 Ingo Wegener（编）。施普林格，1–12。[12] Cynthia Dwork 和 Jing Lei。2009年。差分隐私与稳健统计。载于第*41*届 *ACM* 理论计算年会论文集，*STOC 2009*，美国马里兰州贝塞斯达，*2009*年*5*月*31*日至*6*月*2*日，Michael Mitzenmacher（编）。ACM，371–380。[13] Cynthia Dwork 和 Aaron Roth。2014年。差分隐私的算法基础。理论计算机科学基础与趋势9，3-4（2014年），211–407。[14] facebook-github bot。2022年。Opacus。（2022年）。https://github.com/pytorch/opacus/tree/main/opacus [15] Chang Ge、Shubhankar Mohapatra、Xi He 和 Ihab F. Ilyas。2021年。Kamino：具备约束感知能力的差分隐私数据合成。*VLDB* 学报14，10（2021年），1886–1899。[16] Jonas Geiping、Hartmut Bauermeister、Hannah Dröge 和 Michael Moeller。2020年。反演梯度——在联邦学习中攻破隐私有多容易？载于神经信息处理系统进展 *33*：*2020*年神经信息处理系统年会，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上，Hugo Larochelle、Marc'Aurelio Ranzato、Raia Hadsell、Maria-Florina Balcan，以及Hsuan-Tien Lin（编）。[17] Xiaolin Han、Daniele Dell'Aglio、Tobias Grubenmann、ReynoldCheng和Abra-ham Bernstein。2022年。差分隐私知识图谱嵌入框架。*J. Web Semant.* 72（2022年），100696。[18] Michael Hay、Chao Li、Gerome Miklau和David D. Jensen。2009年。私有网络度分布的准确估计。载于*ICDM2009*，第九届*IEEE*国际数据挖掘会议，美国佛罗里达州迈阿密，*2009*年*12*月*6*日至*9*日，Wei Wang、Hillol Kargupta、Sanjay Ranka、Philip S. Yu和Xindong Wu（编）。IEEE计算机学会，169–178。[19] JamieHayes、Luca Melis、George Danezis和Emiliano De Cristofaro。2019年。LOGAN：针对生成模型的成员推断攻击。*Proc. Priv. Enhancing Technol.* 2019年，1（2019年），133–152。[20] Xinlei He、Jinyuan Jia、Michael Backes、Neil Zhenqiang Gong和Yang Zhang。2021年。从图神经网络中窃取链接。载于第*30*届*USENIX*安全研讨会，*USENIX*安全 *2021*，*2021*年*8*月*11*日至*13*日，Michael Bailey和RachelGreenstadt（编）。USENIX协会，2669–2686。[21] James Jordon、JinsungYoon和Mihaela vander Schaar。2019年。PATE-GAN：利用差分隐私保证生成合成数据。载于第*7*届国际学习表征会议，*ICLR 2019*，美国路易斯安那州新奥尔良，*2019*年*5*月*6*日至*9*日。OpenReview.net。[22] Diederik P. Kingma和Jimmy Ba。2015年。Adam：一种随机优化方法。载于第*3*届国际学习表征会议，*ICLR 2015*，美国圣迭戈，*2015*年*5*月*7*日至*9*日，会议论文集，Yoshua Bengio和Yann LeCun（编）。[23] JaewooLee和ChristopherW.Clifton。2014年。通过差分隐私FP树获取前k个频繁项集。载于第*20*届*ACM SIGKDD*知识发现与数据挖掘国际会议，*KDD '14*，美国纽约州纽约市，*2014*年*8*月*24*日至*27*日，Sofus A. Macskassy、Claudia Perlich、Jure Leskovec、WeiWang和Rayid Ghani（编）。ACM，931–940。[24] Ninghui Li。2022年。差分隐私数据合成：现状与挑战。载于*ASIA CCS'22*：*ACM*亚洲计算机与通信安全会议，日本长崎，*2022*年*5*月*30*日至*6*月*3*日，YujiSuga、Kouichi Sakurai、XuhuaDing和Kazue Sako（编）。ACM，815。[25] MinLyu、Dong Su和Ninghui Li。2017年。理解差分隐私的稀疏向量技术。*Proc. VLDB Endow.* 10，6（2017年），637–648。[26] Luca Melis、Congzheng Song、Emiliano DeCristofaro和VitalyShmatikov。2019年。利用协作学习中非预期的特征泄露。载于*2019IEEE*安全与隐私研讨会，*SP 2019*，美国加利福尼亚州旧金山，*2019*年*5*月*19*日至*23*日。IEEE，691–706。[27] Ilya Mironov。2017年。Rényi 差分隐私。载于第*30*届*IEEE*计算机安全基础研讨会，*CSF2017*，美国加利福尼亚州圣巴巴拉，*2017*年*8*月*21*日至*25*日。IEEE计算机学会，263–275。[28] Viraaji Mothukuri、Reza M. Parizi、Seyedamin Pouriyeh、Yan Huang、Ali De- hghantanha和Gautam Srivastava。2021年。联邦学习的安全与隐私综述。*Future Gener. Comput. Syst.* 115（2021年），619–640。[29] Milad Nasr、Reza Shokri和Amir Houmansadr。2019年。深度学习的全面隐私分析：针对集中式和联邦学习的被动与主动黑盒推理攻击。载于*2019IEEE*安全与隐私研讨会，*SP 2019*，美国加利福尼亚州旧金山，*2019*年*5*月*19*日至*23*日。IEEE，739–753。[30] Hao Peng、Haoran Li、Yangqiu Song、VincentW.Zheng和Jianxin Li。2021年。差分隐私联邦知识图谱嵌入。载于*CIKM'21*：第*30*届*ACM*信息与知识管理国际会议，线上会议，澳大利亚昆士兰，*2021*年*11*月*1*日至*5*日，Gianluca Demartini、Guido Zuccon、J. Shane Culpepper、ZiHuang和Hanghang Tong（编）。ACM，1416–1425。[31] VenkatadheerajPichapati、Ananda Theertha Suresh、Felix X. Yu、Sashank J.Reddi和SanjivKumar。2019年。AdaCliP：用于隐私随机梯度下降的自适应裁剪。*CoRR* abs/1908 07643 2019 1908 07643 。（）。http://arxiv.org/abs/ 。[32] Reza Shokri、Marco Stronati、Congzheng Song和VitalyShmatikov。2017年。针对机器学习模型的成员推断攻击。载于*2017IEEE*安全与隐私研讨会，*SP 2017*，美国加利福尼亚州圣何塞，*2017*年*5*月*22*日至*26*日。IEEE计算机学会，3–18。[33] RichardSocher、Danqi Chen、Christopher D.Manning和Andrew Y. Ng。2013年。利用神经张量网络进行知识库补全。载于神经信息处理系统进展*26*：第*27*届神经信息处理系统年度会议*2013*。论文集，会议于*2013*年*12*月*5*日至*8*日在美国内华达州太浩湖举行，Christopher J. C. Burges、Léon Bottou、ZoubinGhahramani和Kilian Q. Weinberger（编）。926–934。[34] CongzhengSong和Ananth Raghunathan。2020年。嵌入模型中的信息泄露。载于*CCS '20*：*2020 ACM SIGSAC*计算机与通信安全会议，线上会议，美国，*2020*年*11*月*9*日至*13*日，Jay Ligatti、XinmingOu、Jonathan Katz和Giovanni Vigna（编）。ACM，377–390。[35] ZhiqingSun、Zhi-HongDeng、Jian-Yun Nie和Jian Tang。2019年。RotatE：通过复数空间中的关系旋转实现知识图谱嵌入。载于第*7*届国际学习表征会议，*ICLR 2019*，美国路易斯安那州新奥尔良，*2019*年*5*月*6*日至*9*日。OpenReview.net。[36] tensorflower gardener。2022年。Tensorflow Privacy。（2022年）。https://github.com/tensorflow/privacy/tree/master/tutorials

[37] Kristina Toutanova、Danqi Chen、Patrick Pantel、Hoifung Poon、Pallavi Choud- hury 和 Michael Gamon。2015年。表示文本以实现文本与知识库的联合嵌入。载于*2015*年自然语言处理实证方法会议论文集，*EMNLP 2015*，*21*，*2015*年，*17* 葡萄牙里斯本，*9*月 -Lluís Màrquez、Chris Callison-Burch、Jian Su、Daniele Pighin 和 Yuval Marton（编）。计算语言学协会，1499–1509。[38] Théo Trouillon、Johannes Welbl、Sebastian Riedel、Éric Gaussier 和 Guillaume Bouchard。2016年。用于简单链接预测的复数嵌入。载于第*33*届国际机器学习会议论文集，*ICML 2016*，美国纽约州纽约市，*2016*年*6*月*19*日至*24*日（*JMLR*研讨会与会议论文集，第*48*卷），Maria-Florina Balcan 和 Kilian Q. Weinberger（编）。JMLR.org，2071–2080。[39] Stacey Truex、Ling Liu、Mehmet Emre Gursoy、Lei Yu 和 Wenqi Wei。2021年。揭开机器学习即服务中的成员推断攻击之谜。*IEEE*服务计算汇刊 14，6（2021年），2073–2089。[40] Denny Vrandecic 和 Markus Krötzsch。2014年。维基数据：一个免费的协作式知识库。*ACM*通讯 57，10（2014年），78–85。[41] Chris Waites。2020年。PyVacy：适用于 PyTorch 的隐私算法。（2020年）。https: //github.com/ChrisWaites/pyvacy [42] Meng Wang、Ruijie Wang、Jun Liu、Yihe Chen、Lei Zhang 和 Guilin Qi。2018年。迈向 SPARQL 中的空答案：使用 RDF 嵌入近似查询。载于语义网——*ISWC 2018*——第*17*届国际语义网会议，*2018*年*10*月*8*日至*12*日，美国加利福尼亚州蒙特雷，论文集，第一部分（计算机科学讲义，第*11136*卷），Denny Vrandecic、Kalina Bontcheva、Mari Carmen Suárez-Figueroa、Valentina Presutti、Irene Celino、Marta Sabou、Lucie-Aimée Kaffee 和 Elena Simperl（编）。施普林格，513–529。[43] Yu-Xiang Wang、Borja Balle 和 Shiva Prasad Kasiviswanathan。2019年。子采样 Rényi 差分隐私与解析矩会计器。载于第*22*届人工智能与统计国际会议，*AISTATS 2019*，*2019*年*4*月*16*日至*18*日，日本冲绳县那霸市（机器学习研究论文集，第*89*卷），Kamalika Chaudhuri 和 Masashi Sugiyama（编）。PMLR，1226–1235。[44] Yu Wang 和 Lichao Sun。2021年。针对知识图谱的成员推断攻击。*CoRR* abs/2104.08273（2021年）。[45] Fan Wu、Yunhui Long、Ce Zhang 和 Bo Li。2022年。LINKTELLER：通过影响力分析从图神经网络中恢复私有边。载于第*43*届 *IEEE*安全与隐私研讨会，*SP 2022*，*26*，*2022*年*22*日2005 2024美国加利福尼亚州旧金山，*5*月 -。IEEE，– 。[46] Wenhan Xiong、Thien Hoang 和 William Yang Wang。2017年。DeepPath：一种用于知识图谱推理的强化学习方法。载于论文集 *2017*年自然语言处理实证方法会议（*EMNLP 2017*），丹麦哥本哈根，*2017*年*9*月*9*日至*11*日，Martha Palmer、Rebecca Hwa 和 Sebastian Riedel（编）。计算语言学协会，564–573。[47] Bishan Yang、Wen-tau Yih、Xiaodong He、Jianfeng Gao 和 Li Deng。2015年。在知识库中嵌入实体和关系以进行学习与推理。载于 第三届国际学习表征会议，*ICLR 2015*，美国加利福尼亚州圣迭戈，*2015*年*5*月*7*日至*9*日，会议论文集，Yoshua Bengio 和 Yann LeCun（编）。[48] Lei Yu、Ling Liu、Calton Pu、Mehmet Emre Gursoy 和 Stacey Truex。2019年。深度学习的差分隐私模型发布。载于 *2019*年*IEEE*安全与隐私研讨会，*SP 2019*，美国加利福尼亚州旧金山，*2019*年*5*月*19*日至*23*日。IEEE，332–349。[49] Fuzheng Zhang、Nicholas Jing Yuan、Defu Lian、Xing Xie 和 Wei-Ying Ma。2016年。面向推荐系统的协同知识库嵌入。载于 第*22*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，美国加利福尼亚州旧金山，*2016*年*8*月*13*日至*17*日，Balaji Krishnapuram、Mohak Shah、Alexander J. Smola、Charu C. Aggarwal、Dou Shen 和 Rajeev Rastogi（编）。ACM，353–362。[50] Kai Zhang、Yu Wang、Hongyi Wang、Lifu Huang、Carl Yang 和 Lichao Sun。2022年。通过隐私保护关系嵌入聚合实现知识图谱上的高效联邦学习。*CoRR* abs/2203.09553（2022年）。[51] Zhikun Zhang、Min Chen、Michael Backes、Yun Shen 和 Yang Zhang。2021年。针对图神经网络的推理攻击。*CoRR* abs/2110.02631（2021年）。[52] Zhikun Zhang、Tianhao Wang、Ninghui Li、Jean Honorio、Michael Backes、Shibo He、Jiming Chen 和 Yang Zhang。2021年。PrivSyn：差分隐私数据合成。载于 第*30*届*USENIX*安全研讨会，*USENIX*安全 *2021*，*2021*年*8*月*11*日至*13*日，Michael Bailey 和 Rachel Greenstadt（编）。USENIX协会，929–946。[53] Ligeng Zhu、Zhijian Liu 和 Song Han。2019年。梯度深度泄漏。载于 神经信息处理系统进展*32*：神经信息处理系统 *2019*年年度会议，*NeurIPS 2019*，*2019*年*12*月*8*日至*14*日，加拿大不列颠哥伦比亚省温哥华，Hanna M. Wallach、Hugo Larochelle、Alina Beygelzimer、Florence d'Alché-Buc、Emily B. Fox 和 Roman Garnett（编）。14747–14756。[54] Yuqing Zhu 和 Yu-Xiang Wang。2022年。具有自适应K的自适应私有K选择及其在多标签PATE中的应用。载于 国际人工智能与统计学会议，*AISTATS 2022*，*2022*年*3*月*28*日至*30*日，线上会议（机器学习研究论文集，第*151*卷），Gustau Camps-Valls、Francisco J. R. Ruiz 和 Isabel Valera（编）。PMLR，5622–5635。

A 评分函数

四种典型知识图谱嵌入方法的评分函数 $f_r(h, t)$ 见表4。注意，$\mathbf{h}$、$\mathbf{r}$、$\mathbf{t}$是 h、r、t 的嵌入，$\circ$ 表示Hadamard积，Re(·) 表示复向量的实向量分量。

表4：典型知识图谱嵌入方法的评分函数 f_r（h, t）

模型	空间	评分函数 $f_r(h,t)$
TransE	$h, r, t \in \mathbb{R}^d$	$-\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|$
RotatE	$h, r, t \in \mathbb{C}^d$	$-\|\mathbf{h}\circ\mathbf{r}-\mathbf{t}\|$
DistMult	$h, r, t \in \mathbb{R}^d$	$\mathbf{h}^\top$对角矩阵(r)t
ComplEx	$h, r, t \in \mathbb{C}^d$	$\mathrm{Re}(\mathbf{h}^\top \mathrm{diag}(\mathbf{r})\mathbf{t})$

B Rényi 差分隐私的性质

Rényi 差分隐私的一项优良性质是其组合的简洁性。

引理 *B.1*。（*RDP*的自适应组合）若 *M1(D)* 满足 *$(\alpha, \epsilon_1(\alpha))$-Rényi* 差分隐私，且 *M2(D, M1(D))* 满足 *$(\alpha, \epsilon_2(\alpha))$-Rényi* 差分隐私，则 *M1* 和 *M2* 的组合满足 *$(\alpha, \epsilon_1(\alpha)+\epsilon_2(\alpha))$-Rényi* 差分隐私。

Rényi 差分隐私可以轻松转换为差分隐私。

引理 B.2。*(*RDP到差分隐私的转换*)* 若随机化算法$\mathcal{M}$满足 $(\alpha, \epsilon(\alpha))$-*Rényi* 差分隐私，则$\mathcal{M}$对任意 $\delta \in (0, 1)$也满足 $(\epsilon(\alpha)+1$ og$(1/\delta)\alpha-1, \delta)$- 差分隐私。

近似RDP 也具有组合性质。

引理 *B.3.*（近似*RDP*的自适应组合）如果 *M1(D)* 满足 δ_1-近似-*$(\alpha, \epsilon_1(\alpha))$-RDP*，*M2(D, M1(D))* 满足 δ_2-近似-*$(\alpha, \epsilon_2(\alpha))$-RDP*，那么 *M1* 和 *M2* 的组合满足 *$(\delta_1+\delta_2)$*-近似-*$(\alpha, \epsilon_1(\alpha)+\epsilon_2(\alpha))$-RDP*。

近似RDP也可以转换为差分隐私。

引理 B.4。*(*近似RDP到差分隐私的转换*)* 如果一个随机化算法 $\mathcal{M}$满足 δ_1-$(\alpha, \epsilon(\alpha))$-*Rényi* 差分隐私，那么 $\mathcal{M}$也满足$(\epsilon(\alpha)$ + 对数 $(1/\delta)\alpha-1, \delta+\delta_1)$- 差分隐私，对于任意 $\delta \in (0, 1)$。

C 相关工作

成员推断攻击。成员推断攻击（MIA）旨在推断某条数据记录是否曾用于训练机器学习模型。在 Shokri 等人提出首个 MIA[32] 后，大量研究已成功针对各种数据领域发起攻击，例如图像数据集 [16, 53], 、语言数据集[19, 34], 或图结构数据集 [20, 45, 51]。尤其是，Nasr 等人 [29] 和 Hayes 等人 [19] 通过获取梯度更新，提出了针对联邦学习模型的成员推断攻击。针对集中式图神经网络 [20, 45] 或知识图谱 [44] 的成员推断攻击表明，图结构数据也面临隐私泄露风险。

联邦知识图谱嵌入与隐私保护。FedE [9] 被提出作为首个联邦知识图谱嵌入框架，它聚合本地计com-算的实体嵌入更新，以训练全局模型。FedR [50] 聚合关系嵌入，而非实体嵌入。尽管在这些联邦框架中，训练数据保存在本地，但传输的梯度更新可能会编码至少部分训练数据的细节 [3, 28]，并可被利用来危害训练数据的隐私 [7, 29, 39]。差分隐私已被用于研究知识图谱嵌入隐私保护。Han 等人 [17] 提出了一种基于差分隐私随机梯度下降的差分隐私知识图谱嵌入算法，用于集中式学习中的单个知识图谱。Peng等人 [30] 利用现成的PATE-GAN技术 [21]，提出了一种基于 PATE-GAN的差分隐私知识图谱嵌入算法，用于去中心化学习中的多个知识图谱。然而，这两项工作在差分隐私分析方面都不够严谨，因此其差分隐私保证仅适用于部分算法，详见附录H.3。

D 符号表

表 5：重要符号。

变量	描述
$\mathcal{G}$、$\mathcal{E}$、$\mathcal{R}$、$\mathcal{T}$	知识图谱以及实体、关系和三元组的集合
h, r, t	头实体、尾实体和关系
$\mathbf{E}, \mathbf{R}$	表示实体和关系的嵌入矩阵
$\mathbf{h}$、$\mathbf{t}$、$\mathbf{r}$	表示 h, r, t 的嵌入向量
$\mathcal{L}(\mathbf{h}, \mathbf{r}, \mathbf{t})$	嵌入模型的损失函数
$f_r(\mathbf{h}, \mathbf{t})$	用于衡量（$\mathbf{h}$，$\mathbf{r}$，$\mathbf{t}$）合理性的评分函数
B, N, λ	批次大小，训练集中的三元组数量，学习率
q	抽样概率
$\mathbf{G}$	梯度矩阵
C_1, C_2	差分隐私随机梯度下降和隐私选择中的裁剪阈值
σ, η, Δ	噪声分配参数
$\epsilon r, \sigma_p, \delta_t$	p隐私选择参数

E 自适应隐私预算分配的详细信息

在主动攻击实验中，成功率在训练的前几轮达到峰值，并随着模型收敛而下降。这促使我们在前几轮提供更强的防御，并在后续轮次逐步减弱防御，以获得更好的效用。

我们提出在训练过程中（算法 1 第 23–25 行）根据验证准确率自适应地调整噪声尺度 σ。具体而言，我们将初始 σ 设得相对较大，以防御客户端发起的主动推断攻击，并每隔几轮计算一次验证准确率。随着学习过程接近局部最优，较大的 σ 可能会阻碍模型收敛，这会反映在验证准确率的波动上。如果验证准确率不再提高，我们就将 σ 乘以系数 $\eta < 1$，从而减小其值：

$$\sigma = \begin{cases} \eta \cdot \sigma, & \text{if } \mathrm{MRR}_t - \mathrm{MRR}_{t-1} < \Delta, \\ \sigma, & \text{if } \mathrm{MRR}_t - \mathrm{MRR}_{t-1} \geq \Delta, \end{cases}$$

其中，MRR 是衡量知识图谱补全效果的指标，Δ 是一个阈值。验证可以在公开数据集上进行，否则，非隐私验证可能会引发隐私顾虑。如果没有可用的公开数据集，我们会采用预定义的噪声衰减方案，以经验方式定期降低噪声。

F 定理证明

定理 F.1. 算法 *2* 中采样率为 q 的隐私选择满足 δ_t-近似-$(\alpha, \epsilon_s(\alpha))$-*Rényi* 差分隐私，其中

$$\epsilon_s(\alpha) = \frac{1}{\alpha-1}\log(p^2\tbinom{\alpha}{2}\min\{4(e^{\epsilon(2)}-1), e^{\epsilon(2)}\min\{2, (e^{\epsilon(\infty)}-1)^2\}\} + \sum_{j=3}^{\alpha} q^j \tbinom{\alpha}{j} e^{(j-1)\epsilon(j)}\min\{2, (e^{\epsilon(\infty)}-1)^j\} + 1) \; and \; \epsilon(\alpha) = \frac{\alpha}{8\sigma_r^2} + \frac{\alpha}{2\sigma_p^2}.$$

证明。第一阶段使用的指数机制采用 Gumbel 噪声实现方式 [10]，我们利用"有界范围"性质通过 Rényi 差分隐私对其进行分析，从而得到更紧的界。该指数机制的实现满足 $(\alpha, \frac{\alpha}{8\delta_r^2}) - RDP$。

我们在第二阶段采用带有高斯噪声 [54] 的PTR 变体。由于PTR中的检验可能失败的概率不超过 δ_t，该算法满足 δ_t-近似-$(\alpha, \frac{\alpha}{2\sigma_p^2})$-Rényi 差分隐私。

因此，通过组合，不进行子采样的隐私选择满足 δ_t-近似-$(\{v2, (\frac{\alpha}{8\sigma_r^2} + \frac{\alpha}{2\sigma_p^2})\}$-Rényi 差分隐私。

由于算法2的输入是从原始数据集的子采样批次中计算得到的（算法 1 中的第 11行），因此可以通过带有子采样机制 [43] 的 Rényi 差分隐私隐私放大定理来增强隐私保证。□

G 实验设置

G.1 数据集

我们使用了两个真实世界的知识图谱基准数据集：

- FB15k-237 [37]：这是一个子集，包含 237 种关系、14505 个实体和 310116 个三元组，提取自大规模知识图谱 Freebase [4]。
- NELL-995 [46]：这是 NELL [8] 的一个子集，包含 14505 个实体、237 种关系和 154213 个三元组。这些数据由名为 Never-Ending Language Learner 的智能体从网络中构建而成。

对于每个客户端，我们随机选取一个实体子集，并根据该子集分配三元组和关系。我们删除一些三元组，以确保不同客户端中与重叠实体对应的三元组集各不相同。我们还使用相同的方法，从源数据集中划分出验证集和测试集中的三元组。

G.2 超参数设置

对于联邦知识图谱嵌入训练，我们将批次大小设为 64，实体和关系的嵌入维度均设为 128。使用 Adam [22]，初始学习率为 0.001，用于随机梯度下降。我们将负采样数量设为 256，自对抗负采样温度 α 设为 1，间隔 γ 设为 10。

使用差分隐私训练知识图谱嵌入模型时，我们将 $\sigma, \sigma_r, \sigma_p$ 设为 1，整体 δ 设为 10^{-5}。裁剪界限 C_1 和 C_2 分别为 1.2 和 0.8。采样率 q 可根据批次大小 B 和训练数据规模 N 计算，例如 $q = B/N$。对于自适应分配方法，衰减系数 η 为 0.95，阈值 Δ 为 0.001。在训练过程中，可使用上述参数跟踪整体隐私损失（ϵ, δ）。对于预设的 ϵ，当整体隐私损失达到 ϵ 时，我们会中止训练。我们将预设的 ϵ 设为 {2, 4, 8, 16, 24,32, 48,64}。

G.3 指标

攻击。 我们使用 *F1*分数 来衡量攻击的成功率，

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}.$$

效用。 知识图谱嵌入模型的效用取决于其在链接预测任务上的性能，该任务旨在预测三元组（h, r, ?）中缺失的部分。预测过程会给出一个包含所有可能实体的排序列表，排名越高的实体越有可能是正确答案。我们使用平均倒数排名（*MRR*）和*N* 命中率（*Hits@N*）来评估模型在链接预测上的性能。

$$MRR = \frac{1}{|T|} \sum_{i=1}^{|T|} \frac{1}{rank_i},$$

$$Hits@N = \frac{1}{|T|} \sum_{i=1}^{|T|} \mathbb{I}(rank_i \leq n),$$

其中 T 是测试三元组集，$rank_i$ 是真实答案的排名。

H 补充讨论 S

H.1 攻击局限性

我们想指出，上述攻击仍存在一些局限性。服务器与客户端串通时，其中一些局限性可以被突破。例如，由于串通客户端提供了关系信息，A_{SI} 不再需要辅助数据集；而串通服务器可以提供从受害者处收集的完整实体嵌入，因此 A_{CIA} 和 A_{CIP} 不再局限于重叠实体。不过，一些局限性仍然存在。例如，A_{SI} 仅适用于跨国距离类型的知识图谱嵌入模型，其实体嵌入计算更为简单。尽管如此，所提出的攻击足以达到我们的目的，揭示联邦知识图谱嵌入面临着严重的隐私威胁，需要采取防御措施。

H.2 差分隐私粒度

现有的两种图数据邻接定义，即节点邻接和边邻接，都不适用于知识图谱嵌入模型。前者将仅相差一个节点的两个图定义为邻接图。然而，知识图谱嵌入模型会为每个节点（即知识图谱中的每个实体）输出一个嵌入向量，因此节点邻接定义并不适用，因为输出的嵌入已经揭示了对应实体的存在。后者将仅相差一条边的图视为邻接图，这只能反映知识图谱关系的存在，却无法捕捉知识图谱关系中的类型信息。

H.3 现有工作的差分隐私分析不准确

韩等[λ7] 提出了一种基于差分隐私随机梯度下降的差分隐私知识图谱嵌入算法，用于集中式学习设置下的单个知识图谱 [3]。尽管他们也试图只扰动与采样实体绑定的活跃梯度元素，但未能以差分隐私的方式进行选择。因此，整体差分隐私保证并不像论文所声称的那样成立，因为若不进行差分隐私扰动，更新实体的位置就会准确揭示哪些实体存在于训练数据集中，以及哪些实体在迭代过程中被选中进行更新。

Peng等人 [30]提出了一种利用现成PATE-GAN技术、适用于去中心化设置下多个知识图谱的基于PATE-GAN的差分隐私知识图谱嵌入算法 [21]。不同于传统的PATE-GAN用法（即向生成器输入随机数据），[30] 向生成器输入从训练知识图谱中提取的嵌入，而这些训练知识图谱本应通过差分隐私得到保护。