arXiv:2603.15609v1 [stat.AP] 16 Mar 2026

Di网络连通性指数的差分隐私

TOM A. RUTTER、YUXIN LIU、M. AMIN RAHIMIAN

研究人员越来越多地使用社会和经济网络数据来研究一系列社会科学问题，但发布从网络中得出的统计量可能引发重大的隐私担忧。我们展示了如何在边相邻差分隐私下发布网络连通性指数，以量化节点属性之间的同配混合。标准隐私技术在这一场景下表现不佳，原因有二：一是连通性指数具有较高的全局敏感度；二是单个节点的属性可能成为数千个单元中连通性计算的输入，从而导致不佳的组合效果。我们的方法易于应用：首先向节点属性添加噪声，然后对下游统计量进行解析去偏，最后再添加第二层噪声，以保护单条边的存在或缺失。我们证明了对于离散标签和连续标签，我们的估计量均具有一致性和渐近正态性，并且展示了该方法在模拟以及真实网络中的良好表现；这些真实网络由社会科学家收集，规模最小的仅包含 200 个节点。

其他关键词和短语：差分隐私、网络统计量、社交网络、连通性指数、隐私—效用权衡

目录

1 引言 1 2 文献综述 3 3 隐私与连通性指数 4 3.1 实现边相邻差分隐私的组合定理 8 4 二元标签差分隐私：高低地位连通性 8 4.1 网络连通性去偏私有估计的一致性 9 5 连续标签差分隐私：以自身排名回归朋友排名 12 6 模拟网络上的准确性 16 6.1 Erdős-Rényi 和随机块模型中的二元标签 16 6.2 图子上的连续标签 17 7 实证应用 18 7.1 印度农村扩散网络 18 7.2 Twitch 上的相互关注 22 8 讨论 22 A 证明 26 A.1 定理1的证明：(ε, δ) 边邻接下的组合 26 A.2 命题 1 的证明：个体连通性的最小方差无偏估计量 26 A.3 定理 2 的证明：去偏私有估计量的一致性 27 A.4 定理 3 的证明：二元标签的边相邻差分隐私保证 30 A.5 定理 4 的证明：渐近正态性与扰动阶 31 A.6 命题 2 的证明：私有回归估计量的一致性 38 A.7 定理 5 的证明：回归估计的边相邻差分隐私保证 42 A.8 定理6的证明：私有去偏回归估计量的渐近正态性 43 B 补充材料 46 B.1 亚马逊上经常一起购买的商品 46 B.2 符号列表 50

Rutter：斯坦福大学经济学系；*Liu*、*Rahimian*：匹兹堡大学斯旺森工程学院。联系邮箱：*rahimian@pitt.edu*。我们感谢 *Jordan Awan*、*Sahana Subramanyam*、*SamThau*、三位匿名审稿人，以及参加 *2025*年差分隐私理论与实践会议、*2025*年谷歌研究广告隐私会议、匹兹堡大学统计学系研讨会和隐私与公共政策会议的与会者所提供的宝贵意见。*Rutter* 感谢 *Knight-Hennessy* 奖学金的资助。用于复现我们所有分析的代码可在以下网址获取：*https://github.com/TomRutter42/Privacy-for-Connectedness-Indices*。

1 引言

研究各种社会现象的社会科学家越来越多地使用有关个体、企业、金融机构和国家所嵌入的网络的数据。（例如，Easley 和 J. Kleinberg（2010）、Jackson（2019）以及 Goyal（2023）提供了相关综述。）网络数据的使用为传染病传播、同伴和导师对个人生活结果的影响、全球供应链的韧性以及金融系统中系统性风险的性质提供了宝贵见解。然而，使用此类数据也引发了对数据集中个体和组织隐私保护的担忧。天真地公布这些数据集中的聚合统计量，可能会泄露数据集中是否存在特定连接的信息，甚至可能泄露数据所代表的人或企业的个体特征。

为回应人们日益关注如何保障社会科学家所用数据集中涉及的个体和组织的隐私这一问题，自 Dwork 等人（2006）以来，差分隐私领域逐步发展出一系列工具，使人们能够发布数据集中的特定统计量，同时提供数学保证：无论他人可能掌握何种外部知识，仅凭查看聚合统计量，任何人都不可能过多地了解数据中的某个个体或组织。

这一领域的研究源于传统数据保护技术未能充分保护数据中有记录的个体，例如从数据集中删除个人可识别信息等做法。举例来说，Narayanan 和 Shmatikov（2008）将匿名化的 Netflix 电影评分与公开的 IMDB 用户资料交叉比对，成功识别出 Netflix 数据库中的特定用户，尽管 Netflix 电影评分数据集本身并不包含任何个人可识别信息。即使是聚合数据集，如果只采用掩盖小单元或在单元之间调换部分个体等简单的数据保护技术，也可能泄露数据集中某些特定个体的大量信息，令人意外。例如，Dick 等人（2023）针对 2010 年美国人口普查提出了重构攻击，仅凭聚合数据就能可靠地推断出许多个体的特征。他们的攻击之所以有效，是因为人口普查发布了许多包含汇总统计量的表格；对特定表格中的统计量进行差分（这些表格汇总的是不同但并非互斥的个体集合），便能逐步锁定特定个体的数据。计算能力的不断提升以及人工智能的普及，将大幅降低未来针对公开数据发起攻击的成本，同时也会降低成功实施此类攻击所需的数学专业知识门槛。

差分隐私文献中的核心思想，是向聚合统计量中注入随机统计噪声，从而在数学上保证任何可能的攻击都无法揭示数据集中任何一个数据点的大量信息；这样一来，外部人员便无法区分某个特定个体对统计量的影响与随机噪声的影响。通常，这种噪声是根据统计量对数据集中任意特定观测值变化的敏感度 来施加的。在人口普查重构攻击中，引入的统计噪声彻底破坏了这样一种攻击逻辑：攻击者通过对不同发布结果进行差分，获取数据集中某个特定个体的信息。

对于从网络构建的统计量，统计量的敏感度往往可能覆盖整个参数空间，这会限制私有化统计量在维持特定隐私保证的同时对其真实值的保真度。事实上，我们证明，本文所考虑的这类网络统计量确实存在这种情况，因此需要采用另一种方法。

本文提出了一种方法，可在提供形式化隐私保证的同时，发布一类特定的网络统计量；即使对于规模相对较小的现实网络，该方法也能保持较高的准确度。我们考虑的统计量是网络连通性指数，用于描述具有某种特征的节点与具有相同或不同特征的节点之间的连接比例。例如，白人个体的朋友中有多大比例也是白人？美国公司的供应商中有多大比例位于中国？有多大比例的银行贷款流向了对冲基金？

本文聚焦的应用问题，也是本文的研究动机，涉及跨越收入分布的友谊关系所占的比例。在一对广受引用的研究中，Chetty 等人（2022a）和 Chetty 等人（2022b）公开发布了利用 Facebook 和 Instagram 数据构建的社会资本指数，涵盖美国的县、邮政编码、大学和学校。具体而言，Chetty 等人（2022a）发布了一项指标，用于衡量某地区低社会经济地位个体中拥有高社会经济地位（SES）朋友的比例，并指出该指标是美国经济流动最有力的预测指标。（他们使用 Chetty 等人（2026）的数据，并将某地区的经济流动定义为：父母处于全国收入分布第 25 百分位数的孩子成年后的平均收入排名。）他们公开发布的统计量已被广泛下载，并用于后续研究。尽管这些研究的作者采用了差分隐私领域的工具来降低隐私损失风险，但并未为所发布的连通性指数提供形式化隐私保证。Harris 等人（2025）的后续研究将这些指标扩展到英国；他们在公开数据之前也加入了噪声，但同样没有提供形式化隐私保证。

本文提出了一种简单易用的方法，可在提供正式隐私保证的前提下公开发布网络连通性指数，例如经济连通性。在实践中，我们的方法可以从至少包含 200 个个体的现实网络中构建连通性指数并予以发布，同时将统计噪声控制在较低水平。

我们方法的核心思路是，先向数据中的节点特征注入随机噪声，使下游统计量基于已满足隐私保证的节点特征数据库计算。这一点至关重要，它使我们能够避免考察在“最坏情况”网络（例如现实中并不存在的星形网络）中，节点特征变化对我们关注的统计量所造成的敏感度问题。此外，这也使我们能够避开这样一个问题：节点的属性发生变化，可能会影响大量单元中的连通性，从而导致组合效果不佳。相反，通过对节点特征进行隐私化处理，并假设所有下游统计量均由这些经过净化的节点级数据计算得出，我们只需考察边集变化对统计量的敏感度；我们证明，这种敏感度要低得多。此外，某条边（或无边）只能改变至多两个单元中的连通性值，从而实现更理想的组合效果。

向节点特征中注入随机噪声通常会使下游统计量产生偏差，因此我们提供了解析校正公式，这些公式也可以在下游应用，以消除这些统计量的偏差，同时根据后处理定理（Dwork 和 Roth，2014年，命题 2.1）维持其隐私保证。

我们尤其指出，个体特征数据库通常并非关注对象，因此不必（而且通常不应）发布。尽管不发布这一中间数据集并不会降低数据发布所造成的数学隐私损失，因为我们不对外部人员的背景知识作任何假设，但在实践中，这提供了额外一层隐私保护，因为现实中的外部人员并不具备通常构成研究者试图提供隐私保证时最坏情形的那种极端信息水平（Bassily 等人，2013年）。

本文结构如下。第2节概述有关差分隐私以及我们所考察的网络统计量的相关文献。第3节介绍我们采用的隐私定义和网络连通性指数。第4节提出适用于二元标签的方法。第5节将我们的框架扩展至连续标签。第6节给出用于评估我们方法准确性的模拟。第7节展示我们的方法在现实世界带标签网络上的实证表现，第8节作结。

2 文献综述

我们的论文建立在差分隐私领域丰富的文献基础之上，该领域始于 Dwork 等人（2006）。标准差分隐私机制往往利用记录之间的独立性，但对于网络数据而言，记录显然并非相互独立（Kifer 和 Machanavajjhala，2011）。这种依赖性并非只是理论上的好奇现象，而是一种在实践中可被利用的漏洞。Liu 等人（2016）展示了这一点：他们根据在独立性假设下具有差分隐私保证的统计量，推断出用户的敏感位置数据。上述研究凸显了隐私应用有必要明确考虑数据中的统计依赖性，而我们的方法正是针对网络连通性指数做到这一点。

我们的主要贡献在于经济数据发布实践，因此本研究建立在 Chetty 等人（2022a）和 Chetty 等人（ 2022b ）的工作之上。他们提出了经济连通性的概念，并发布了美国地理区域、学校和大学的细粒度统计量。这些论文的核心创新在于表明：通过结合谁与谁相连的信息以及相连个体的特征信息构建的网络连通性度量，在解释代际流动性等经济结果的变异方面，显著优于仅基于网络边结构的度量（如聚类系数），或仅基于节点特征、忽略节点间连接的度量（如与某个邻域收入构成相关的变量）。这一洞见——连接结构和相连个体的特征都至关重要——直接启发了我们采用 Blocki 等人（2013）提出的边邻接隐私概念，并使我们能够保证对边的存在（或缺失）和节点特征都予以保护。

我们的论文还涉及有关依赖数据差分隐私的研究。依赖差分隐私及其推广形式处理成对相关性和任意相关性，但它们通常依赖于难以在缺乏显式概率模型的情况下进行估计的依赖参数（Liu 等人，2016年；Zhao 等人，2017年）。一种更灵活的替代方案是 Pufferfish 框架（Kifer 和 Machanavajjhala，2014年），该框架通过对秘密、判别对和数据生成分布进行显式建模，支持定制化的隐私定义，例如 Blowfish 和贝叶斯差分隐私。在图数据背景下，Wasserstein 和 Markov quilt 机制等 Pufferfish 实现通过利用贝叶斯网络对依赖关系进行建模，在群体差分隐私方面展现出更高的效用（Song 等人，2017年）。互补性方法还包括推断隐私，它利用马尔可夫链量化相关性（Ghosh 和 R. Kleinberg，2016年）；以及零知识隐私，它能够针对连通性等特定图性质提供更强的保证（Gehrke 等人，2011年）。

尤其是在图设置中，一种常见方法是将数据投影到最大度较低的图上，以降低统计量的敏感度（Kasiviswanathan 等人，2013年）。在许多设置中，这一技术效果很好；但对于我们所关注的应用而言，度异质性是数据的一个重要特征，而将数据投影到最大度较低的图上可能会丢失这一特征。（例如，Harris 等人（2025）表明，在考察连通性指数时，英国网络中按社会经济地位划分的度变异是一个重要特征——在这一

特定设置中施加较低的度上界可能会大幅改变连通性估计值。）有关网络差分隐私的各种概念及其解释，请参见 Hehir 等人（2025）的综述。

3 隐私与连通性指数

我们遵循 Dwork 等人（2006）及差分隐私文献中的做法，采用一种隐私定义：要求从两个相近数据集中生成任意统计输出的概率相似。

定义 1（ε-差分隐私）。若对于所有相邻数据集 D 和 D'，以及机制所有可能的输出集合 $\mathcal{S}$，均满足：则称随机化机制 $\mathcal{M}$ 是 ε-差分隐私的。

$$\left|\ln\left(\frac{\mathbb{P}(\mathcal{M}(D)\in\mathcal{S})}{\mathbb{P}(\mathcal{M}(D')\in\mathcal{S})}\right)\right|\le\varepsilon$$

对这个定义的一种直观理解是：观察到 M 的输出后，即使对手掌握了有关数据集其余部分的无限制信息，也无法大幅提高其确信真实数据集是 D 而非 D'（反之亦然）的程度。

本文的目标是构造一种机制，输出满足 ε-差分隐私的连通性指数估计值。为统一符号，我们先定义带标签网络。

定义 2（带标签网络）。带标签网络是一个三元组 $(\mathcal{V}、\mathcal{E}、\mathbf{L})$，其中 $\mathcal{V}$ 是顶点集合，$\mathcal{E}$ 是边集合，而

$$\mathbf{L}=(l_i)_{i\in\mathcal{V}},\qquad l_i\in\{a,b\},$$

是节点标签向量，它为每个节点分配一个来自标签集合 $\{a、\ b\}$ 的标签。

请注意，该定义类似于 Blocki 等人（2013）对社交网络的定义。节点可能具有的一对标签例如“高社会经济地位”或“低社会经济地位”。为便于记号，我们定义 $e_{ij}=1$ 当且仅当 $(i,j)\ \in$E 时（即 i和 j之间有一条边），否则为 $e_{ij}=0$。我们将节点 i的度记为：

$$d_i=\sum_{j\in\mathcal{V}}e_{ij}$$

并将节点 i 的邻域定义为与其相连的节点 i 的集合：

$$N(i)=\{j\in\mathcal{V}:e_{ij}=1\}$$

定义 3（跨类型连通性指数）。取一个带标签网络 (V、E、L)，其中

$$\mathbf{L}=(l_i)_{i\in\mathcal{V}},\qquad l_i\in\{a,b\}.$$

将节点的诱导划分定义为

$$\mathcal{A}:=\{i\in\mathcal{V}:l_i=a\},\qquad \mathcal{B}:=\{i\in\mathcal{V}:l_i=b\}.$$

跨类型连通性指数从A到B定义为

$$C^{\mathcal{A}\to\mathcal{B}}\ :=\ \frac{1}{\#(\mathcal{A})}\sum_{i\in\mathcal{A}}\frac{\sum_{j\in\mathcal{B}}e_{ij}}{d_i}.$$

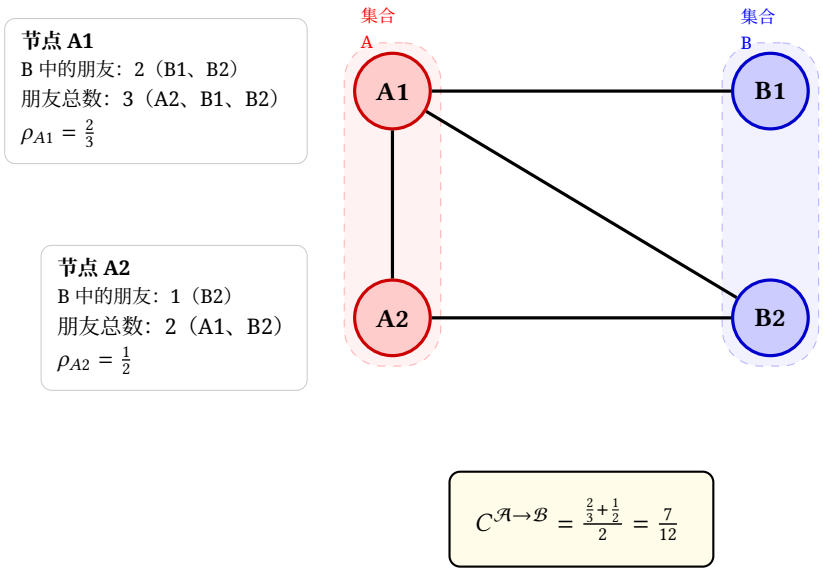


图 1. 跨类型连通性指数的计算示例。

也就是说，$C^{\mathcal{A}\to\mathcal{B}}$ 是群体$\mathcal{A}$ 中的个体与群体$\mathcal{B}$中的个体之间连接所占的平均比例。

Chetty 等人（2022a）和 Chetty 等人（2022b）针对美国的县、邮政编码、学校和大学，编制了三种不同类型的跨类型连通性指数：

（1）经济连通性（低 SES 个体的高社会经济地位朋友所占的平均比例）。（2）语言连通性（主要语言不是英语的个体中，朋友的主要语言为英语所占的平均比例）。（3）年龄连通性（25至34岁个体中，35至44岁朋友所占的平均比例）。

Bailey 等人（2025）构建了一个跨类型连通性指数，用于衡量男性个体中女性朋友所占平均比例。

关于符号表示，我们定义：

$$\rho_i := \frac{\sum_{j\in\mathcal{B}} e_{ij}}{d_i}$$

下文均采用这种表示方式，因此跨类型连通性指数可写为：

$$C^{\mathcal{A}\to\mathcal{B}} = \frac{1}{\#(\mathcal{A})}\sum_{i\in\mathcal{A}} \rho_i$$

我们在图1中提供了一个说明跨类型连通性指数的简单示例。在表1中，我们总结了所使用的符号。（更完整的符号列表见表B.2。）

我们还可以定义同类型连通性指数，即从一组节点连接到其自身的连接所占的比例。

表1。符号概述

符号	描述
$\mathcal{V}$	网络中顶点（节点）的集合
$\mathcal{E}$	节点之间的边（友谊关系/连接）集合
$\mathbf{L} = (l_i)_{i\in\mathcal{V}}$	每个节点的节点标签向量为 $l_i \in \{a, b\}$
e_{ij}	二元指示变量；$e_{ij} = 1$ 若 i 与 j 之间存在一条边，则为 1，否则为 0
d_i	节点 i 的度，$\#(j \in \mathrm{V} : e_{ij} = 1)$。
$N(i)$	节点 i 的邻域，$\{ j \in \mathrm{V} : e_{ij} = 1\}$。
$\mathcal{A}, \mathcal{B}$	基于标签划分的 V（a，b）
$\#(\cdot)$	基数算子（集合中的元素数量）
ρ_i	个体 i 的朋友中属于目标群体的比例
$C_{\mathrm{A}\to\mathrm{B}}$	跨类型连通性指数（A组中个体的平均值 ρ_i）

定义 ion 4（同类型连通性指数）。设有一个带标签网络(V，E，L)，其中

$$\mathbf{L} = (l_i)_{i\in\mathcal{V}}, \qquad l_i \in \{a, b\}.$$

定义节点的诱导划分为

$$\mathcal{A} := \{i \in \mathcal{V} : l_i = a\}, \qquad \mathcal{B} := \{i \in \mathcal{V} : l_i = b\}.$$

同类型连通性指数对于群体A 定义为

$$C^{\mathcal{A}\to\mathcal{A}} := \frac{1}{\#(\mathcal{A})} \sum_{i\in\mathcal{A}} \frac{\sum_{j\in\mathcal{A}} e_{ij}}{d_i}.$$

也就是说，$C^{\mathcal{A}\to\mathcal{A}}$ 是群体$\mathcal{A}$ 中的个体与同一群体$\mathcal{A}$内其他个体之间连接比例的平均值。

请注意，$C^{\mathcal{A}\to\mathcal{A}} = 1 - C_{\mathcal{A}\to\mathcal{B}}$。

我们通常需要计算特定单元的跨类型连通性指数，例如，单元可以是县、邮政编码区域或学校。在这种情况下，如果我们令 s 表示该单元（由一组用户构成，例如一所学校），则可以考虑以下统计量:

$$C_s^{\mathcal{A}\to\mathcal{B}} = \frac{1}{\#(\mathcal{A} \cap s)} \sum_{i\in\mathcal{A}\cap s} \frac{\sum_{j\in\mathcal{B}} e_{ij}}{d_i}$$

or:

$$C_s^{\mathcal{A}\to\mathcal{B}} = \frac{1}{\#(\mathcal{A} \cap s)} \sum_{i\in\mathcal{A}\cap s} \frac{\sum_{j\in\mathcal{B}\cap s} e_{ij}}{\sum_{j\in\mathcal{V}\cap s} e_{ij}}$$

区别在于，我们是只考虑单元内个体与单元内其他个体之间的友谊关系，还是考虑单元内个体与任何其他个体之间的连接。

对于我们的隐私保证，我们将遵循 Blocki 等人（2013）提出的边相邻差分隐私 概念，该概念能够保护网络中个体节点的特征（例如个人的收入或种族），以及网络中特定边的存在与否。将满足以下属性的带标签网络视为相邻时，这就是ε-差分隐私:

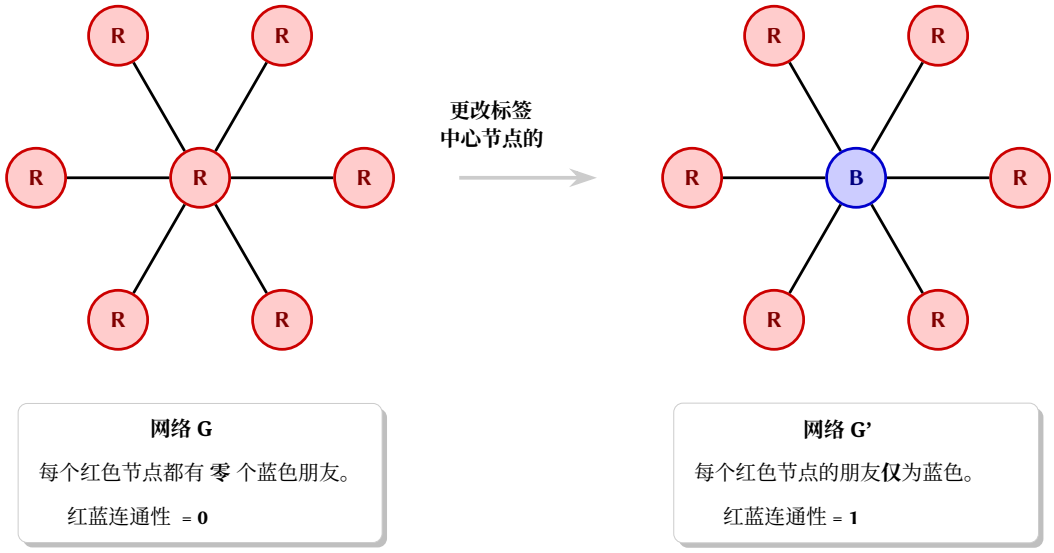


图 2. 星形网络示意图，说明节点级隐私概念对于连通性统计量可能过于严格：修改中心节点（或在节点差分隐私下保护其存在及关联边）可以使红蓝连通性从 0 变为 1，这意味着敏感度涵盖 $[0, 1]$，且不会随着网络规模增大而减小。

定义5（边相邻的带标签网络）。两个带标签网络 (V，E，L) 和(V，E′，L′)称为边相邻，当且仅当：

- 边集$\mathcal{E}$ 与$\mathcal{E}'$ 至多相差一条边，并且
- 标签向量$\mathbf{L} = (l_i)_{i\in\mathcal{V}}$ 与$\mathbf{L}' = (l'_i)_{i\in\mathcal{V}}$ 至多有一个分量不同，即至多存在一个节点 $u \in \mathcal{V}$ 使得 $l_u \neq l'_u$。

也就是说，如果两个网络至多只在一条边是否存在以及至多一个节点的特征上不同，那么它们就是边相邻的。边相邻这个名称不太恰当，因为它确实允许节点特征发生变化；但为与先前文献保持一致，我们沿用这一术语。Jorgensen 等人（2016）和 Chen 等人（2020）均采用了边相邻差分隐私的概念。需要指出的是，我们的边邻接概念比 Blocki 等人（2013）所述的略强，因为我们允许一条边和一个节点的特征同时 发生变化，而不是只允许一条边或 一个节点的特征发生变化。

请注意，我们的相邻关系并未将数据集中节点的存在 视为隐私。若要求保护节点存在隐私，则会使我们转向节点级差分隐私（节点差分隐私）；在这种隐私定义下，如果一个图可以通过添加或删除一个顶点以及（可能的）所有关联边而得到另一个图，那么这两个图就是相邻的。在社交网络场景中，对于连通性统计量而言，这种更强的隐私定义通常不可行：删除一个高阶顶点就可能大幅改变跨类型连通性，迫使任何差分隐私机制注入规模大到淹没信号的噪声。图 2 给出了一个简单示例：改变中心节点（等价地，在节点差分隐私下保护其存在及关联边）可能使红蓝连通性从 0 变为 1，因此敏感度覆盖整个参数空间，且不会随网络规模增大而减小。

因此，我们遵循 Chen 等人（2020）倡导的建模视角：在许多平台上，用户资料（以及顶点集 V）的存在实际上是公开的，而

资料中记录的个人特征则是敏感信息。因此，我们的目标是保护节点属性 和个体边 （通过边相邻差分隐私），而非节点存在性；否则，相关的连通性指数将无法实际应用。

3.1 A 实现边相邻差分隐私的组合定理

本节将介绍一种方法，使我们能够在提供形式化隐私保证的同时发布连通性指数。

我们的方法核心逻辑如下。连通性指数可能会泄露网络中某个给定节点的两项特征信息——节点标签和其友谊关系——这正是我们所采用的边相邻差分隐私概念所体现的。Chetty 等人（2022a）和 Chetty 等人（2022b）采用的方法是，通过计算经济连通性指标，再根据聚合统计量对节点特征和友谊关系变化的局部敏感度添加噪声，从而保护经济连通性。[1]

我们表明，先对节点特征添加噪声，再校正由此造成的经济连通性后续估计中的偏差，随后根据节点友谊网络变化的敏感度添加噪声，是一种更好的方法；它只需注入少量噪声，就能提供形式化隐私保证。[2]

我们的方法基于以下定理（证明见 A.1 节），该定理形式化了这一构造所依据的组合原则。定理指出，先对节点属性应用隐私机制，再对以隐私化属性为输入的聚合统计量应用一个单独的边差分隐私机制，二者组合后满足边相邻差分隐私。

定理1（边邻接下的（ε, δ）组合）。设 (V, E, L) 和 (V, E', L') 是边-$\widehat{\ }$相邻的带标签网络。设 $M_1 : (V, E, L) \mapsto (V, E, L)$ 是关于改变单个节点属性的 $(\varepsilon_\ell, \delta_\ell)$-差分隐私机制；对于每个固定的 L，$M_2(\cdot, L) : (V, E) \mapsto \mathbb{R}$ 是关于改变单条边的 $(\varepsilon_e, \delta_e)$-差分隐私机制。定义组合机制 $M := M_2 \circ M_1$。那么，M 满足 $(\varepsilon_\ell + \varepsilon_e, \delta_\ell + e^{\varepsilon_\ell}\delta_e)$ 边相邻差分隐私。

4 二元标签差分隐私：高、低状态连通性

现在，我们利用定理1中确立的通用组合框架。具体而言，我们首先构造一个针对节点属性的 ε_ℓ-差分隐私机制 M_1，随后将 ε_e 边差分隐私机制应用于根据隐私化标签计算出的聚合统计量。

因此，我们的第一步是使用随机响应，在 ε_l-差分隐私下保护节点属性。具体来说，对于每个节点 i，类别标签 $l_i \in \{a,\ b\}$ 会以概率独立地翻转其取值

$$p = \frac{1}{1 + e^{\varepsilon_l}},$$

$\hat{}$ 从而得到隐私标签 l_i。

[1]局部敏感度的概念源自 Nissim 等人（2007）。然而，尽管 Nissim 等人（2007）提出的平滑敏感度方法在形式上满足隐私要求，Chetty 等人（2022a）和 Chetty 等人（2022b）使用的基于局部敏感度的方法（该方法采用了 Chetty 和 Friedman（2019）的最大观测敏感度方法）却不具备形式上的隐私保障，因为最大观测敏感度本身是隐私数据的函数，因此可能泄露有关离群值的信息。[2]它还避开了这个问题中原本会遇到的一项关键难题：给定节点的标签可能是数据集中每个单个单元的连通性指数的输入（因为一个节点可能与其他每个单元中的某个节点相连）。对于基于节点标签变化的全局敏感度的方法，这会导致组合问题，因为隐私损失的组合规模等于单元数量与所发布统计量数量的乘积。我们的方法只需按所发布统计量的数量进行组合。这是因为节点标签一经去标识化，根据后处理定理，就可以从这些隐私标签中推导出任意数量的后续单元格级指数，而不会进一步消耗为节点特征预留的隐私预算。

事实 1. 令每个节点的属性为 $l_v \in \{a, b\}$，并令 $\mathbf{L} = (l_v)_{v \in \mathcal{V}}$ 表示标签向量。定义随机化机制 $\mathcal{M}_1$，该机制根据以下规则独立替换每个节点的属性

$$\mathcal{M}_1(l_v) = \begin{cases} l_v, & \text{with probability } 1 - p = \dfrac{e^{\varepsilon_l}}{1 + e^{\varepsilon_l}} \\ \text{the opposite value in } \{a, b\}, & \text{with probability } p = \dfrac{1}{1 + e^{\varepsilon_l}} \end{cases}$$

那么 $\mathcal{M}_1$ 满足关于单个节点属性变化的 ε_l-差分隐私。

证明。随机响应的这一性质在差分隐私文献中广为人知。例如，Wang 等人（2016）的结果 1 证明了这一性质。□

这种方法的一个潜在问题是，随机响应会导致连通性指标出现衰减偏差。采用高于中位数与低于中位数的划分时，随机翻转节点标签会机械地使观测到的群体构成趋近于 1/2，即使真实的同质性很强也是如此。例如，如果 $p = \frac{1}{2}$，那么无论社会中的真实同质性水平如何，根据隐私化标签，平均而言，每个节点的邻居中都有一半看起来属于群体 b。

因此，我们需要一种校正机制，仅使用隐私化标签 l_i，而不访问非隐私标签 l_i，即可恢复跨类型连通性的无偏估计。命题 1（证明见 A.2 节）构造了一个最小方差无偏估计量（MVUE），仅利用随机响应的输出即可校正这种衰减偏差。

命题 1（个体连通性的最小方差无偏估计量）。对于每个节点 i，定义真实的个体连通性

$$\rho_i := \sum_{j \in \mathcal{V}} a_{ij}\, \mathbf{1}\{l_j = b\}, \qquad a_{ij} := \frac{e_{ij}}{d_i}.$$

ˆ设隐私化标签 l_j 由翻转概率为.....的随机响应生成

$$p = \frac{1}{1 + e^{\varepsilon_l}},$$

并定义观测代理量

$$\hat{\rho}_i := \sum_{j \in \mathcal{V}} a_{ij}\, \mathbf{1}\{\hat{l}_j = b\}.$$

那么，仅基于隐私化ˆ标签$(l_j)^{j \in V}$得到的 ρ_i 的唯一最小方差无偏估计量（*MVUE*）为

$$\tilde{\rho}_i := \frac{\hat{\rho}_i - p}{1 - 2p}. \tag{1}$$

4.1 网络连通性的去偏私有估计的一致性

命题 1 证明，给定隐私化标签后，$\tilde{\rho}_i$ 是个体连通性的无偏估计量（实际上是最小方差无偏估计量 ρ_i）。然而，我们的目标估计量是单元级连通性指数，它对潜在 节点子集（例如 $A \cap s$）上的 ρ_i 进行汇总，并通过比率（Hájek 型标准化）实现。个体分量的无偏性并不自动意味着所得比率估计量无偏或在有限样本中表现良好；尤其是，它无法控制随机分母接近零的概率。此外，最终发布的统计量还包含根据随总体规模增大而减小的边敏感度校准的边差分隐私噪声。

定义属于 A 的去偏权重：

$$w_i := \frac{\mathbf{1}\{\widehat{l}_i = a\} - p}{1 - 2p}. \tag{2}$$

Let

$$S_{0,n} := \sum_{i \in V} w_i, \qquad S_{1,n} := \sum_{i \in V} w_i \, \tilde{\rho}_i,$$

其中， $\tilde{\rho}_i$ 是式（1）中 ρ_i 的去偏代理。Hájek 型估计量为

$$\widetilde{C}^{\mathcal{A} \to \mathcal{B}} := \frac{S_{1,n}}{S_{0,n}}, \tag{3}$$

当$S_{0,\ n} \neq 0$ 时（我们证明，在温和条件下，$S_{0,\ n}$ 以趋于 1 的概率有正的下界）。

接下来，我们给出一个大样本保证：在温和的正则条件下，去偏 Hájek 估计量是一致的，且额外的边差分隐私噪声渐近可忽略，因此整体隐私发布依概率收敛于真实的连通性指数。为陈述该结果，我们先引入两个正则条件。第一个条件确保目标单元渐近地不趋于可忽略，从而 Hájek 归一化中的随机分母不会趋于零。第二个条件是有界度条件，用于控制局部依赖，并确保估计量及其边敏感度均一致有界。

假设 1（类型为 A 的节点占非零比例）。存在 $\pi_{\mathcal{A}} \in (0, 1]$ 使得#(A)/$n \to \pi_{\mathcal{A}}$ 当 $n \to \infty$。

假设 2（有界度）。存在 $\Delta < \infty$，使得对于所有 n 和所有 $i \in V$，

$$d_i \leq \Delta. \tag{4}$$

为了实现第 3.1 节中构造的第二阶段，我们必须校准添加到聚合统计量中的边差分隐私噪声。由于拉普拉斯机制要求释放量的边敏感度具有一个上界，因此我们计算$S_{1,\ n}$在添加或移除单条边时的敏感度。

重要的是，在此阶段，隐私化标签 $\mathbf{L}$ 已通过 $\mathbf{M}_1$生成。因此，$(w_i)_{i \in V}$、 S_0以及去偏因子都已固定，我们只需控制边集变化时 $S_{1,\ n}$如何变化。以下引理给出了该边敏感度的一个统一上界（证明见第 A.3 节，其中还给出了其他辅助引理，以证明隐私网络连通性估计量的一致性）。

引理*1*（$S_{1,n}$ 的边敏感度）。固定 $(l_i)_{i \in V}$，使得 $(w_i)_{i \in V}$ 固定。设 E 和 E' 仅相差一条边 (u, v) 的添加或删除。假设 $\hat{\rho}_i(E)$ 按行计算为 $\hat{\rho}_i(E) = \sum_j a_{ij}(E)\, \mathbf{1}\{l_j = b\}$，其中 $a_{ij}(E) = e_{ij}/d_i$（因此，添加或删除 (u, v) 时，只有 u 和 v 对应的归一化权重行会发生变化）。那么，

$$\left| S_{1,n}(E) - S_{1,n}(E') \right| \ \leq \ \frac{2(1-p)}{(1-2p)^2}.$$

接下来，我们研究所提出的估计量的渐近行为。尽管 $\widetilde{C}^{\mathcal{A} \to \mathcal{B}}$ 由无偏分量构造而成，但它是一个比率估计量，且最终发布的结果还包含边差分隐私噪声。因此，随着 n 增大，隐私估计量能否继续保持良好的性质并非显而易见。

下面的定理证明了一致性，并表明添加的边差分隐私噪声在渐近意义下可以忽略（证明见附录 A.3）。

定理 2（去偏私有估计量的一致性）。假设假设 *1* 和 *2* 成立，且 $p \in (0, 1/2)$。令 $\widetilde{C}^{\mathcal{A}\to\mathcal{B}} = S_{1,n}/S_{0,n}$ 为式 (3) 中的去偏 ***Hájek*** 型估计量。定义边差分隐私发布

$$\widehat{C}^{\mathcal{A}\to\mathcal{B}}_{\mathrm{DP}} := \widetilde{C}^{\mathcal{A}\to\mathcal{B}} + Z_n, \qquad Z_n \sim \mathrm{Lap}\left(0, \frac{2(1-p)}{(1-2p)^2\,\varepsilon_e\,S_{0,n}}\right),$$

其中 Z_n 与随机响应扰动相互独立。因此，对于任意固定的 $\varepsilon_e > 0$，

$$\widehat{C}^{\mathcal{A}\to\mathcal{B}}_{\mathrm{DP}} \xrightarrow{p} C^{\mathcal{A}\to\mathcal{B}}, \ as\ n \to \infty.$$

我们在算法 1 中概述了我们的方法。

算法 1 差分隐私跨类型连通性指数

要求： 网络 $(\mathcal{V},\mathcal{E})$(带权重 $e_{ij} \geq 0$)，节点标签 $l_i \in \{a, b\}$，隐私参数 $\varepsilon_l, \varepsilon_e$。**确保：** 差分隐私连通性指数 $\widehat{C}^{\mathcal{A}\to\mathcal{B}}$ DP

11: $p \leftarrow \frac{1}{1+e^{\varepsilon_l}}$ 2: **对于** 每个节点 $i \in \mathrm{V}$ **执行**

3: $\hat{l}_i \leftarrow \begin{cases} \text{取相反值 } \{a, b\}, \text{ 概率为 } p, \\ l_i \text{ ，概率为 } 1-p \end{cases}$ ▷ 对标签进行随机响应

4: **结束循环**

5: **对于** 每个节点 $i \in \mathrm{V}$ **执行** ▷ 计算去偏权重和个体连通性

6: $\hat{\rho}_i \leftarrow \begin{cases} \frac{1}{d_i} \sum_{j\in N(i)} \mathbf{1}\{\hat{l}_j = b\}, & d_i > 0, \\ 0, & d_i = 0 \end{cases}$

7: $w_i \leftarrow \frac{\mathbf{1}\{\hat{l}_i = a\} - p}{1-2p}$

8: $\tilde{\rho}_i \leftarrow \frac{\hat{\rho}_i - p}{1-2p}$

9: </style>结束循环

10: $S_{0,n} \leftarrow \sum_{i\in\mathcal{V}} w_i$

11: $S_{1,n} \leftarrow \sum_{i\in\mathcal{V}} w_i\,\tilde{\rho}_i$

12: $\widehat{C}^{\mathcal{A}\to\mathcal{B}}_{\mathrm{DP}} \leftarrow \frac{S_{1,n}}{S_{0,n}} + Z_n, \quad Z_n \sim \mathrm{Lap}\left(0, \frac{2(1-p)}{(1-2p)^2\,\varepsilon_e\,S_{0,n}}\right)$ ▷ 私有化 Hájek 估计量

13: **return** $\widehat{C}^{\mathcal{A}\to\mathcal{B}}_{\mathrm{DP}}$

基于算法 1，我们现陈述所提出程序的整体隐私保证。该结果直接由定理1以及随机响应的 ε_ℓ-差分隐私性质和拉普拉斯机制的 ε_e 边差分隐私校准（在 A.4 节中证明）得出。

定理 3（算法 1 的隐私保证）。算法 *1* 针对输入 $(\varepsilon_\ell + \varepsilon_e)$ $(\mathcal{V}, \mathcal{E}, \mathrm{L})$满足边相邻差分隐私。

如定理 3 所述，算法 1 清晰地分解了整体隐私预算：标签私有化步骤占用 ε_ℓ，而最终的边隐私发布占用ε_e。通过顺序组合，整个过程因此满足 $(\varepsilon_\ell + \varepsilon_e)$ 边相邻差分隐私。

定理 4（渐近正态性与扰动阶）。在算法 *1* 的设定下，假设假设 *1–2* 成立、最大度一致有界，并且

$$\frac{\#(\mathcal{A})}{n} \to \pi_{\mathcal{A}} \in (0,1).$$

Let

$$\hat{\theta}_n := \frac{S_{1,n}}{S_{0,n}}, \qquad \theta_n := \frac{\mathbf{E}[S_{1,n}]}{\mathbf{E}[S_{0,n}]},$$

并定义

$$T_n := \begin{pmatrix} S_{0,n} \\ S_{1,n} \end{pmatrix}.$$

令最终的隐私发布为

$$\hat{\theta}_n^{DP} := \hat{\theta}_n + Z_n,$$

其中

$$Z_n \mid S_{0,n} \sim \mathrm{Lap}\left(0, \frac{2(1-p)}{(1-2p)^2 \varepsilon_e S_{0,n}}\right).$$

在假设 *1–2* 下，可以推得

$$\frac{1}{n}\,\mathrm{Cov}(T_n) \to \Sigma$$

对于某个有限矩阵 Σ，且渐近方差

$$\sigma^2 = \nabla g(\mu_0,\mu_1)^\top \Sigma\, \nabla g(\mu_0,\mu_1)$$

有明确定义，其中 $g(x,y) = y/x$和

$$(\mu_0,\mu_1) = \lim_{n\to\infty} \mathbf{E}\left[\frac{T_n}{n}\right], \qquad \mu_0 > 0.$$

Then

$$\sqrt{n}(\hat{\theta}_n - \theta_n) \Rightarrow \mathcal{N}(0,\sigma^2).$$

此外，

$$\mathrm{Var}(Z_n) = \Theta(n^{-2}).$$

因此，相对于隐私化比率估计量的根-n 波动尺度，额外的边隐私拉普拉斯扰动在渐近意义上可以忽略不计。

正如定理 4 所确立的，不确定性的主要来源是 θ_n的标签随机化，而非额外的边隐私扰动。由于拉普拉斯噪声的方差为 n^{-2}阶，因此不会影响发布的估计量的一阶渐近分布。

5 差异 连续标签的差分隐私：以自身排名回归朋友排名 k

到目前为止，我们的研究结果主要聚焦于采用二元标签的网络连通性指数，但我们预计，直接对敏感属性添加噪声并根据私有输入向量计算统计量的方法，将有助于处理与带标签网络相关的各种统计量。我们通过将这一方法应用于即将在后续论文中发布的统计量来说明这一点。这些统计量使用网络节点的连续有界标签，而非我们此前研究的离散标签。具体而言，这些论文构建的统计量涉及具有特定排名的节点，或属于特定排名集合的节点的朋友标签均值。更广泛地说，我们的方法有助于针对以下问题提供差分隐私答案：流动性较差的企业是否也倾向于依赖流动性较差的供应商，

绿色企业是否倾向于从其他绿色企业采购原材料，或者学生在考试成绩方面是否表现出同质性。

首先，为数据集中的所有个体 j赋予一个特征 x_j，其取值属于 $\mathbb{R}$中的一个连通有界域。一般而言，不失一般性，我们可以将该域设为 $[0,1]$。例如，x_j可以是如上文论文中所述的连续型社会经济地位指标，但更一般地，它通常是某种数值特征的百分位排名。

对于用户集合 A，我们可以考察集合—单元组合中个体的平均好友排名 。令 y_i 表示个体 i的平均好友排名：

$$y_i = \frac{\sum_{j \in N(i)} x_j}{d_i}.$$

˜考虑对 x_i进行 y_i的回归，并令 $\tilde{\alpha}$和 β分别表示该回归截距和斜率的估计值。

定义 6（平均好友排名均值）。集合$\mathcal{A}$ 中个体的平均好友排名均值为：

$$\mathrm{MAFR}_s^{\mathcal{A}} = \frac{\tilde{\alpha} + \tilde{\beta} \int_{\mathcal{A}} x \, \mathrm{d}x}{\int_{\mathcal{A}} 1 \, \mathrm{d}x}. \tag{5}$$

也就是说，我们首先求出每个个体 $i \in$ A的朋友平均排名，然后再对 A 中所有个体的这些排名取均值。“平均好友排名均值”这一术语体现了公式 (5) 中涉及的两重平均：先对特定个体的朋友取平均，再对相关集合中的个体取平均。

本节所考虑的所有统计量都可以从回归直线中推导出来，其中x 变量是节点标签， y 变量是节点连接对象的平均标签。根据后处理定理，问题由此转化为构建一条差分隐私回归直线。

首先请注意，只有 x变量中的测量误差（此处由差分隐私程序引入）会使回归直线产生偏差，而对 y变量施加均值为零的噪声会增大回归输出的标准误，但不会影响其期望值。

为校正这种偏差，假设

$$y_i = \alpha + \beta x_i + \nu_i, \qquad \mathbb{E}[\nu_i \mid x_i] = 0,$$

并令

$$\hat{y}_i = y_i + \eta_i, \qquad \hat{x}_i = x_i + u_i,$$

其中

$$\mathbb{E}[\eta_i \mid y_i] = 0, \qquad \mathbb{E}[u_i \mid x_i] = 0, \qquad \mathbb{E}[u_i^2 \mid x_i] = \sigma^2.$$

如果 β^* 表示以ˆx为自变量、ˆy为因变量进行回归所得的斜率系数，那么标准误差变量修正为

$$\tilde{\beta} = \beta^* \frac{\left(\frac{1}{n-1}\right) \sum_i (\hat{x}_i - \bar{\hat{x}})^2}{\left(\frac{1}{n-1}\right) \sum_i (\hat{x}_i - \bar{\hat{x}})^2 - \sigma^2} \tag{6}$$

该估计量对真实回归系数 β 是一致的。相应的截距估计量为

$$\tilde{\alpha} = \bar{y} - \tilde{\beta} \bar{x},$$

该方法对 α是一致的。我们在下文将此校正应用于隐私化回归输出。为保证完备性，我们在命题 2 中正式陈述这一结果，并在附录 A.6 中给出证明；相关讨论另见 Greene（2012）。

现在我们来描述这一隐私机制，它由两个部分组成。首先，由于 $x_j \in [0,1]$，我们通过添加截断拉普拉斯噪声对每个个体特征进行隐私化，从而得到受隐私保护的排名 $\hat{x}_j$。按照 Geng 等人（2020）的做法，该机制经过校准，以达到所需的隐私级别，并使扰动后的特征保持有界；这与支持集无界的标准拉普拉斯机制不同。第一步保护的是敏感的节点标签本身。

其次，我们使用隐私化协变量和相应的隐私化好友排名结果计算回归，然后对得到的有界回归问题应用算法 3。第二步在回归阶段增加隐私保护；结合第一步对节点标签的扰动，最终发布的回归系数满足边相邻差分隐私。

在第一步中加入有界噪声在此尤其有用，因为有界线性回归的隐私保证要求协变量和结果保持在有界范围内（Holohan 等人，2019），且算法 3 所添加的后续噪声的尺度取决于这些界限。

由此，我们可以为每个人 i 计算平均好友排名，而不会泄露有关任何 x_j 的信息：

$$\widehat{\mathrm{AFR}}_i = \frac{\sum_{j \in N(i)} \hat{x}_j}{d_i}$$

请注意，这些隐私化好友排名可能也不在 $[0,1]$ 之间，因为 $\hat{x}_j$ 可能不在 $[0,1]$ 之间。不过，由于我们使用有界域中的噪声，AFR_i 仍然有界，因此回归斜率和截距的敏感度也有界（Alabi 等人，2022年）。

我们在算法 2 中总结了我们的方法。

算法 2 差分隐私平均平均好友排名

要求： 节点属性 $\{x_i\}_{i \in \mathcal{V}}$ 及其 $x_i \in [0,1]$、网络 (V, E)、隐私参数 $(\varepsilon_\ell, \delta_\ell)$ 和 ε_e

确保： 差分隐私平均平均好友排名 $\mathrm{MAFR}^{\mathcal{A}}_{\mathrm{DP}}$ 1: 设置 $\Delta = 1$, $\lambda \leftarrow \frac{\Delta}{\varepsilon_\ell}$, $A \leftarrow \frac{\Delta}{\varepsilon_\ell} \log\left(1 + \frac{e^{\varepsilon_\ell} - 1}{2\delta_\ell}\right)$, $B \leftarrow \frac{1}{2\lambda\left(1 - e^{-A/\lambda}\right)}$. 2: 对 $i \in \mathrm{V}$中的每个节点执行 3: 从密度为 $f(x) = Be^{-|x|/\lambda}$ 且定义在 $[-A, A]$上的截断拉普拉斯分布中采样 z_i，并计算 $\hat{x}_i \leftarrow x_i + z_i$。 4: 结束循环 5: 对 $i \in \mathrm{V}$中的每个节点执行 6: 计算 $\hat{y}_i \leftarrow \frac{1}{d_i} \sum_{N(i)} \hat{x}_j$ (若 $d_i = 0$，则设置 $\hat{y}_i = 0$)。 7: 结束循环 ˆ 8: 将算法 3 应用于 $\{(\hat{x}_i, \hat{y}_i)\}_{i \in \mathcal{V}}$，使用隐私参数 ε_e，以获得 $(\hat{\alpha}, \beta)$。 ˆ ˜9: 将公式 (6) 中所述的去偏方法应用于 $(\hat{\alpha}, \beta)$，以获得 $(\tilde{\alpha}, \beta)$。 10: 根据 $(\tilde{\alpha}, \tilde{\beta})$ 构造 $\mathrm{MAFR}^{\mathcal{A}}$ DP 并返回。

我们对简单回归线两个系数的估计量是一致的。

命题 2. 算法 *2* 得出的回归斜率估计值是真实回归系数的一致估计量。

该命题表明，隐私机制不会影响斜率估计量在大样本下的有效性。具体而言，在考虑截断拉普拉斯机制引入的扰动以及 DPSuffStats 注入的额外噪声后，最终的去偏估计量仍是真实回归系数的一致估计量。

算法 3 广义 DPSuffStats：$(\varepsilon,\ 0)$-差分隐私有界输入算法

要求： 数据：$\{(\hat{x}_i,\ \hat{y}_i)\}_{i=1}^{n} \in [a,b]\times[a',b']$

要求： 隐私参数：ε

确保： 差分隐私回归系数 $\tilde{\alpha},\ \tilde{\beta}$

计算 $\bar{x} = \frac{1}{n}\sum_i \hat{x}_i,\quad \bar{y} = \frac{1}{n}\sum_i \hat{y}_i$

计算

$$\mathrm{nvar}(x) = \sum_i (\hat{x}_i - \bar{x})^2, \qquad \mathrm{ncov}(x,y) = \sum_i (\hat{x}_i - \bar{x})(\hat{y}_i - \bar{y})$$

设置敏感度界：

$$\Delta_1 = \left(1 - \frac{1}{n}\right)(b-a)^2, \qquad \Delta_2 = 2\left(1 - \frac{1}{n}\right)(b-a)(b'-a')$$

采样 $L_1 \sim \mathrm{Lap}(0,\ 3\Delta_1/\varepsilon)$

采样 $L_2 \sim \mathrm{Lap}(0,\ 3\Delta_2/\varepsilon)$

如果 $\mathrm{nvar}(x) + L_2 > 0$ **则**

$$\hat{\beta} = \frac{\mathrm{ncov}(x,y) + L_1}{\mathrm{n方差}(x) + L_2}$$

设置 $\Delta_3 = \frac{b'-a'}{n} + |\hat{\beta}|\frac{b-a}{n}$

采样 $L_3 \sim \mathrm{Lap}(0,\ 3\Delta_3/\varepsilon)$

$\hat{\alpha} = (\bar{y} - \hat{\beta}\bar{x}) + L_3$

返回 $\hat{\alpha},\ \hat{\beta}$

否则

返回 $\perp$

结束如果

前述结果确立了所提出的估计量的统计有效性，表明去偏私有估计量仍然是一致的。现在我们转向对机制的隐私分析。具体而言，我们在边相邻邻接关系下刻画其正式的差分隐私保证（证明见第 A.7 节）。

定理 5（算法 2 的隐私保证）。*算法 2 对输入* $(\mathcal{V}、\mathcal{E}、\mathbf{L})$*满足* $(\varepsilon_\ell+\varepsilon_e,\ \delta_\ell)$ *边相邻差分隐私。*

证明可通过分别分析算法的两个阶段得出。截断拉普拉斯扰动确保节点属性满足（$\varepsilon_\ell,\ \delta_\ell$）-差分隐私，而依赖边的估计步骤在全局敏感度界下满足 ε_e 边相邻差分隐私。结合定理1中确立的组合性质，即可得到该结果。

定理6（私有去偏回归估计量的渐近正态性）。*在算法 2 中的连续标签私有回归程序下，定义*

$$\hat{x}_i = x_i + z_i, \qquad \hat{y}_i = \frac{1}{d_i}\sum_{j\in N(i)} \hat{x}_j,$$

并令

$$M_n := \sum_{i=1}^{n} Z_{i,n}, \qquad Z_{i,n} := \begin{pmatrix} \hat{x}_i \\ \hat{y}_i \\ \hat{x}_i^2 \\ \hat{x}_i\hat{y}_i \end{pmatrix}, \qquad \bar{M}_n := \frac{1}{n} M_n = \begin{pmatrix} \bar{M}_{1,n} \\ \bar{M}_{2,n} \\ \bar{M}_{3,n} \\ \bar{M}_{4,n} \end{pmatrix}.$$

定义

$$\beta_n^* := \frac{\bar{M}_{4,n} - \bar{M}_{1,n}\bar{M}_{2,n}}{\bar{M}_{3,n} - \bar{M}_{1,n}^2}, \qquad \tilde{\beta}_n := \frac{\bar{M}_{4,n} - \bar{M}_{1,n}\bar{M}_{2,n}}{\bar{M}_{3,n} - \bar{M}_{1,n}^2 - \sigma_z^2},$$

其中 $\sigma_z^2 = \text{方差}(z_i)$，并令

$$h(m_1, m_2, m_3, m_4) := \frac{m_4 - m_1 m_2}{m_3 - m_1^2 - \sigma_z^2},$$

使得 $\tilde{\beta}_n = h(\bar{M}_n)$。令 $\hat{}\,\beta_n$ 表示经过算法 *3* 中的 DPSuffStats扰动后最终发布的估计量。

假设 $\mu_3 - \mu_1^2 > 0$ 和 $\mu_3 - \mu_1^2 - \sigma_z^2 > 0$，则有

$$\mu_n := \mathbf{E}[\bar{M}_n] \to \mu, \qquad \frac{1}{n}\,\mathrm{Cov}(M_n) \to \Sigma,$$

对于某个确定性向量 $\mu = [\mu_1, \mu_2, \mu_3, \mu_4]^T \in \mathbb{R}^4$ 和确定性矩阵 $\Sigma \in \mathbb{R}^{4\times 4}$，并且还有 $h(\mu) = \beta$，其中 β 是真实回归系数。于是

$$\sqrt{n}\,(\tilde{\beta}_n - \beta) \;\Rightarrow\; N(0, \tau^2), \qquad \tau^2 = \nabla h(\mu)^\top \Sigma \nabla h(\mu).$$

此外，额外的DPSuffStats扰动在渐近意义下可以忽略：

$$\sqrt{n}\,(\hat{\beta}_n - \tilde{\beta}_n) \xrightarrow{p} 0.$$

因此，

$$\sqrt{n}\,(\hat{\beta}_n - \beta) \;\Rightarrow\; N(0, \tau^2).$$

定理6表明，与离散标签情形一样，DPSuffStats引入的额外扰动在 theroot-n 尺度下渐近可忽略，且估计量渐近正态。

我们将该方法应用于第7.2节中来自 Twitch 直播平台的数据，以及第B.1节中经常被一起购买的亚马逊商品，并表明我们的隐私估计量在这两种情况下都能很好地恢复真实的回归斜率。

6 模拟网络上的准确性

6.1 Erdős-Rényi 和随机块模型上的二元标签

为了检验我们的方法，并了解其有效性如何随网络特征和隐私参数而变化，我们将该方法应用于根据 Erdős-Rényi 模型和随机块模型模拟生成的网络。

首先，我们考察在基于随机块模型模拟的网络中改变同质性程度时，隐私连通性估计值的准确性（相对于模拟网络中的真实连通性指数）如何变化。图 3a 显示，与无同配性的基准情形相比，较低程度的同质性不会显著影响隐私经济连通性统计量的准确性。然而，极高程度的同质性会降低隐私指数的准确性。这是因为，同质性程度较高的随机块模型具有更高的聚类程度，因此节点状态翻转所导致的方差更大，因为这可能会影响该组中更多节点的 ρ_i 值。不过，

对于所有程度的同质性，随着我们增加 ε，隐私估计量都会收敛到网络连通性的真实值，如图 3b 所示。

在图 B.2 中，我们展示了对于给定的单元大小，随着单元组合发生变化，隐私指数的准确性如何变化。直观而言，对于给定的单元大小，当单元中低社会经济地位个体较少时（沿用 Chetty 等人（2022a）对两组节点的称呼），我们的隐私指数准确性较低：这是因为我们要对更少个体所受到的噪声取平均，这些噪声来自交换纳入 $\tilde{\rho}_i$ 中关系对象的社会经济地位；同时，为保护边而注入的噪声规模也会随低社会经济地位个体数量减少而降低。最后，在图 B.3 中保持平均度不变时，准确性会随网络规模迅速提高：随着 n 增大，在所有隐私预算下，隐私连通性估计量的均方误差都会下降，这与边敏感度降低以及标签噪声的平均效应相一致。

我们还展示了如何在 ε_e 和 ε_l 之间分配隐私预算会如何影响均方误差。对于给定的 $\varepsilon_\ell + \varepsilon_e$，从图 4 可以看出，随着网络规模扩大，为了使均方误差最小化，增加 ε_ℓ 相对于 ε_e 的比例会变得更优。这与定理 4 一致；该定理表明，从渐近角度看，施加于标签的噪声所导致的方差将占主导地位。

在图 B.3 中，我们绘制了网络的均方误差随节点数的变化。可以看到，在固定$\varepsilon_\ell + \varepsilon_e$且网络平均度也保持不变时，均方误差会随着网络节点数的增加而降低。这体现了我们的估计量具有一致性，详见定理 4。

6.2 图元上的连续标签

我们还在模拟网络上测试了针对连续标签的方法，以了解其有效性如何随网络特征而变化。

我们构建包含 N 个节点的网络，并为每个节点 i 从 $[0,1]$上的均匀分布中抽取 x_i。然后，我们以以下概率填充边：[3]

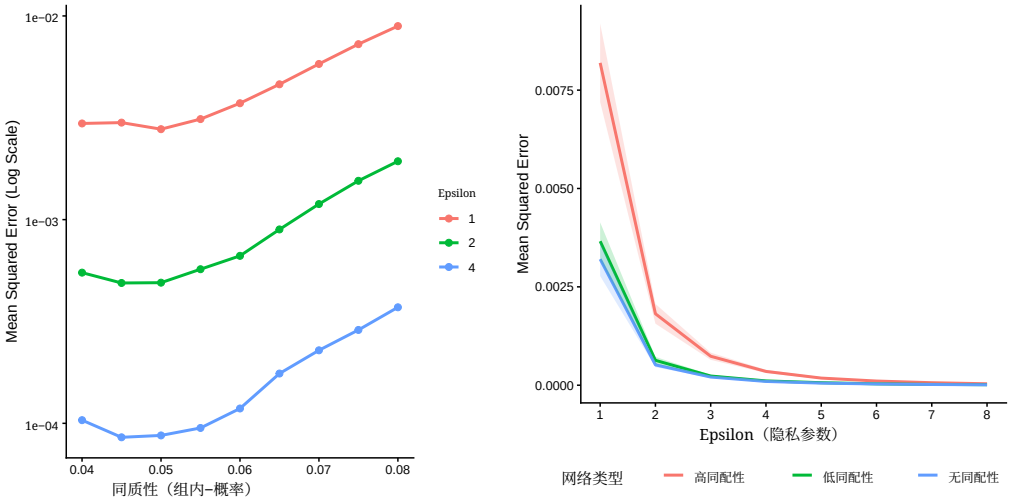
$$\mathbb{P}\{e_{ij}=1 \mid x_i, x_j\} = \frac{\bar{d}}{(N-1)\left(\frac{2}{h} - \frac{2}{h^2}(1-e^{-h})\right)} e^{-h|x_i - x_j|}$$

其中，$\bar{d}$是网络的目标平均度。前面的分数是一个缩放参数，用于使网络的平均度在期望上保持为 $\bar{d}$。项 $e^{-h|x_i-x_j|}$（同质性参数为 $h \geq 0$）刻画了取值 x 彼此接近的节点之间形成边的可能性，比取值 x 相距较远的节点之间形成边的可能性高出多少。[4]

图 5总结了模拟环境的两个基本特征。面板a显示，同质性参数 h 越高，在以自身排名为自变量、平均好友排名为因变量的回归中，R^2 值也越高。面板b展示了在不同网络规模下，ε_e 与 ε_l 之间的分配如何影响差分隐私回归斜率估计的均方误差。与二元情形一样，随着网络规模增大，最优做法是让 ε_l 相对于 ε_e更高；这与定理6的结果一致，即用于保护边的噪声相较于注入标签中的噪声所产生的影响，在渐近意义下可以忽略不计。

图6显示，由回归斜率构建的底四分位节点的回归斜率和 MAFR 的均方误差均随网络规模增大而降低，但相对而言不受网络同质性程度的影响。面板a报告回归斜率的结果，面板b报告底四分位 MAFR 的结果。均方误差

[3]根据这一设定，这些模拟图可以视为从图子中抽样得到。[4]请注意，当 $h \to 0^+$ 时，该表达式的极限为 $\frac{\bar{d}}{N-1}$。在我们的模拟中，有时会设定 $h=0$；在这种情况下，我们将 i 与 j之间相应的边概率设为 $\frac{\bar{d}}{N-1}$。



(a) 均方误差与同质性。

（b）均方误差与隐私预算的关系。

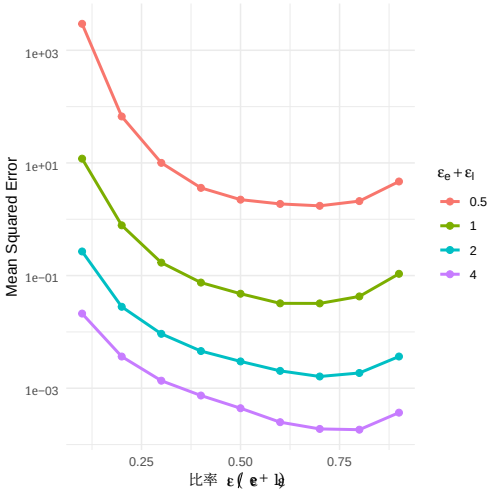
图 3。**均方误差与隐私预算下同质性的关系**。面板（a）：该图展示了在四种隐私预算（$\varepsilon \in \{0.5, 1, 2, 4\}$）下，同质性对差分隐私机制准确性的影响。在每种情况下，我们将隐私预算在 ε_e 和 ε_l之间平均分配。我们的模拟基于由随机块模型生成的网络，这些网络包含 5000 个节点和两个规模相等的群组。为隔离社群结构与网络密度的影响，将总平均度固定为 ≈ 80。具体而言，在将群组内连接概率从 0.04 提高到 0.08 的同时，将群组间概率从 0.04 降低到 0，使群组内连接概率与群组间连接概率之和保持为 0.08。每个数据点的结果取 1,125 个模拟样本的平均值（75 个固定图 $\times$ 15 耦合噪声种子）。纵轴采用对数刻度。面板（b）：该图绘制了三种模拟网络情景下，我们的差分隐私机制的均方误差随隐私预算（ε）变化的情况：无同配性、低同配性和高同配性。准确性基于每个 Epsilon 值的 500 个样本计算，使用 50 个预先生成的固定图，以及每个图对应的 10 个耦合隐私噪声过程，以隔离隐私导致的方差。阴影区域表示 95% 置信区间。隐私预算在 ε_e 和 ε_l之间平均分配。每个模拟网络由 $N = 2,000$ 个节点组成。我们使用包含两个规模相等群组的随机块模型模拟网络。对于无同配性情况，我们将群组内和群组间的连接概率均设为 0.04。对于低同配性情况，我们将群组内连接概率设为 0.06，将群组间连接概率设为 0.02。对于高同配性情况，我们将群组内连接概率设为 0.08，将群组间连接概率设为 0。

随网络规模增大而降低，这再次与我们的估计量会随着网络规模扩大而收敛于真实值这一事实一致，详见定理6。

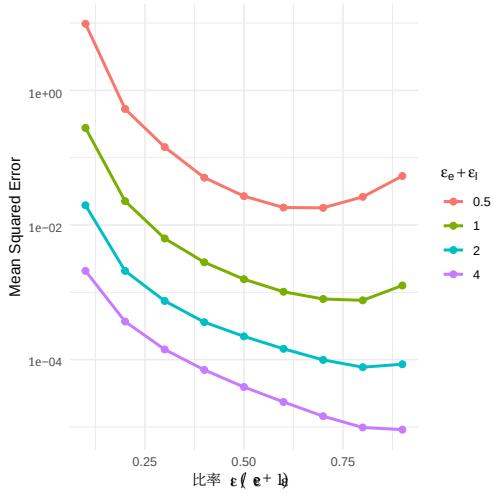
7 实证应用

7.1 印度农村的扩散网络

我们将该方法应用于 Banerjee 等人（2013）的数据，以展示其经验表现。该数据集被广泛使用，包含来自印度南部卡纳塔克邦农村村庄的社交网络。这些数据非常适合用于应用和检验我们的方法，因为其中包含一组规模较小、稀疏但类型多样的社交网络。如果我们的方法在这些数据上表现良好，那么我们预计它在大多数社会科学数据集上也能表现良好。

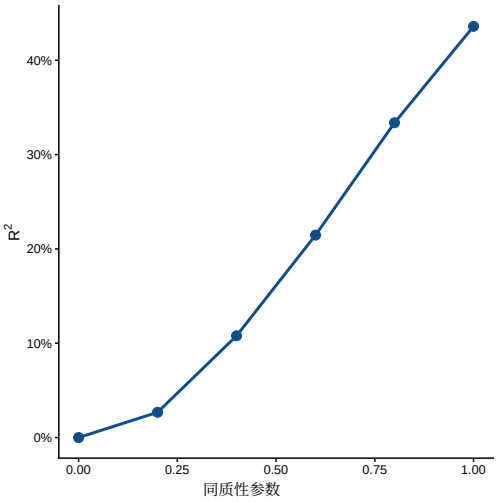


(a) 2,000 个节点。

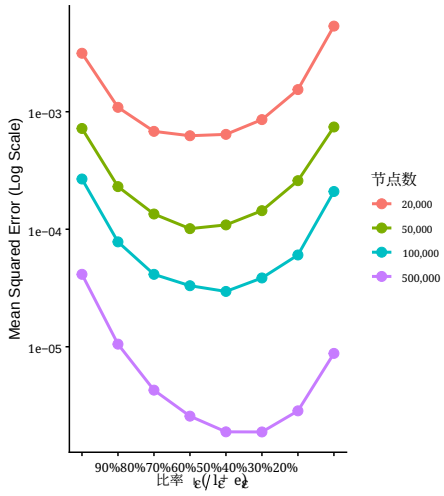


(b) 20,000 个节点。

图 4. 固定 $\varepsilon_e + \varepsilon_l$ 时的均方误差。

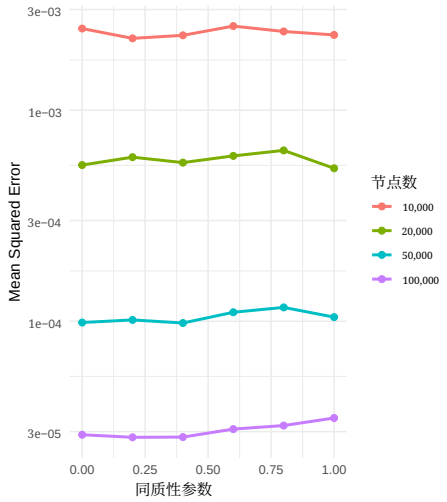


(a) R^2 由平均好友排名对自身排名的回归得出，随同质性变化。

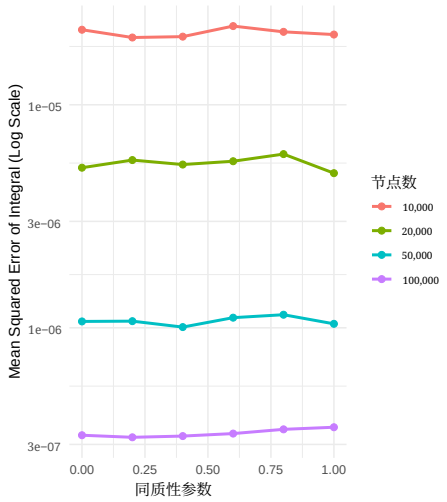


(b)不同隐私预算分配下差分隐私回归斜率估计的均方误差。

图 5. **连续标签的模拟。**面板 (a) 绘制了在网络中将节点标签从 $[0,1]$ 均匀抽样，并以概率$20\,99,\ 999\left(\frac{2}{h}-\frac{2}{h^2}(1-e^{-h})\right)^{e^{-h|x_i-x_j|}}$添加边所生成的网络中，平均好友排名对自身排名回归所得的 R^2，其中同质性水平各不相同。每个模拟网络包含 100,000 个节点，平均度为 20。面板 (b) 展示了四种网络规模下的结果，其同质性为 $h = 0.8$，期望度为 $\bar{d} = 20$；每个数据点是 3,000 次模拟的平均值。



(a) 不同网络规模和同质性下回归斜率的均方误差。



(b) 不同网络规模和同质性下底四分位 MAFR 的均方误差。

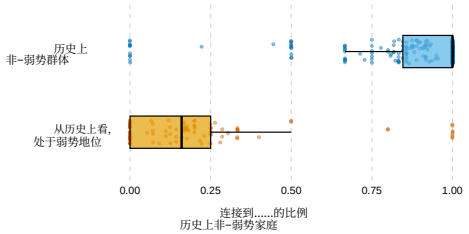
图 6。**连续标签模拟中的估计准确度**。 在两个面板中，我们通过从 $[0,1]$ 均匀抽取节点标签，并以概率 $e^{-h|x_i-x_j|}$ 填充边来生成网络，f 对于变化的 levels of h。面板l a sh显示了 h均方 d 误差 o f 我们的隐私保护回归斜l率估计 f 由平均好友排名对自身排名的回归得出的估计。面板b展示了根据回归斜率构造的底四分位节点的 MAFR 均方误差。在两个面板中，我们设定 $\varepsilon_e = \varepsilon_l = 4$。

数据集包含家庭种姓的信息。为了创建二元标签集，我们将家庭分为两组——如果家庭自认属于表列种姓、表列部落或少数群体，则归入历史上处于弱势的家庭；如果家庭报告属于高种姓，或属于落后种姓但不属于表列种姓，则归入另一组（历史上非弱势家庭）。如果两个家庭报告在所询问的任一网络层中存在联系，我们便认为它们彼此相连。

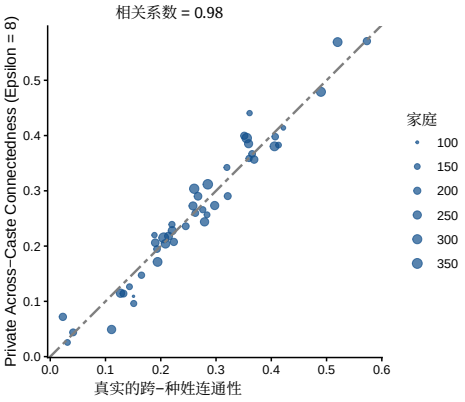
数据中有50个村庄包含至少部分家庭的种姓信息。我们剔除了4个村庄（村庄37、41、43和56），因为这些村庄中历史上处于弱势地位的家庭不超过6户。这样，我们最终纳入46个村庄进行分析。我们在表 B.1 中列出样本的汇总统计量。表 B.1显示，总体而言，种姓之间存在很强的同质性。图7通过三个互为补充的示例，展示了这一模式以及隐私保护程序的实际表现。面板a展示了村庄60的一个示例；该村庄规模最大。

根据这些数据，我们计算每个村庄历史上处于弱势地位的家庭中，与历史上非弱势家庭建立连接的平均比例，从而构建跨种姓连通性指数。我们采用的隐私参数 ε 为8，与 Chetty 等人（2022a）针对 Facebook 数据以及 Chetty 等人（2026）针对美国税务数据所使用的参数一致。我们为每个村庄计算500次带噪隐私跨种姓连通性估计值，并在最右侧一列报告这些估计值的标准差。结果发现，相对于跨种姓连通性的估计值，该标准差通常较低。在各村庄中，信号标准差为0.12，平均噪声标准差为0.04，因此信号方差与噪声方差之比为10.8。因此，即使纳入我们的形式化隐私程序，变异的大部分仍来自

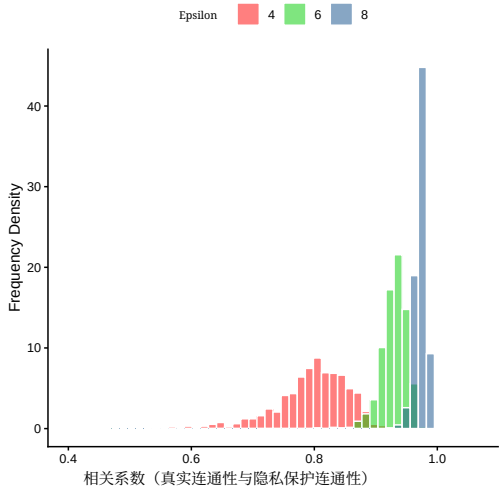
数据集中各村庄之间跨种姓连通性的真实潜在变异，而我们的隐私保护程序引入的变异仅占数据总变异中相对较小的比例。图7的面板b展示了真实跨种姓连通性与隐私保护后的跨种姓连通性示例，表明真实数据与隐私保护后的数据高度相关；规模较大的村庄往往更接近45度线。面板c展示了各村庄真实跨种姓连通性与隐私保护后的跨种姓连通性之间相关系数的分布，说明对于 $\varepsilon = 8$，这些相关性通常非常高，但随着我们逐渐转向 $\varepsilon = 4$，相关性开始减弱。



(a) 村庄60中，按家庭状态划分的历史上非弱势家庭连接占比。



（b）在 $\varepsilon = 8$的一次模拟中，真实跨种姓连通性与隐私跨种姓连通性的示例。



(c) 在不同的 ε设置下，各次模拟中真实跨种姓连通性与隐私跨种姓连通性之间相关系数的分布。

图7. **卡纳塔克邦村庄网络中的跨种姓连通性，以及差分隐私估计量的表现**。面板a展示了村庄60中，按家庭状态分别统计的每个家庭与历史上非弱势家庭连接所占比例的分布。面板b展示了在 $\varepsilon = 8$下某次特定模拟中，村庄层面真实跨种姓连通性与差分隐私跨种姓连通性的散点图；虚线是经过原点的45度线。面板c展示了在 $\varepsilon \in \{4, 6, 8\}$下，各次模拟迭代中真实跨种姓连通性与差分隐私跨种姓连通性之间皮尔逊相关系数的分布。我们将隐私预算平均分配给 ε_e和 ε_l。

7.2 Twitch 上的相互关注

我们还将针对连续标签的方法应用于 Rozemberczki 和 Sarkar（2021年）收集的直播服务 Twitch 数据。该数据集包含约 168,000 名 Twitch 用户。如果一对节点在该平台上相互关注，则数据中存在一条边；数据集约有 680 万条边。我们根据语言将用户划分为 21 个不同群组。数据集中的每位用户都关联有观看次数，我们根据观看次数为每位用户分配其在数据集中的百分位排名。随后，我们根据这些排名为每位用户分配平均好友排名，并通过对平均好友排名关于自身排名的回归来估计平均好友排名均值。

总体而言，在 Twitch 上，观看次数较高的用户往往会相互关注观看次数较低的用户，如图 8 面板a所示，该图展示了英语用户子网络。

我们针对二元标签的设定进行了类似的实验：对每种语言计算回归斜率的隐私估计值，并将这些隐私估计值与真实回归斜率进行比较。图 8 面板b展示了该实验的一次模拟，其中 $\varepsilon = 8$。面板c展示了在不同 ε取值下，各次模拟中隐私系数与真实斜率系数之间相关性的分布。我们得出结论，对于由 Rozemberczki 和 Sarkar（2021年）收集的、规模和结构与 Twitch 数据相当的网络，当 ε 介于 4 到 8 之间时，我们的方法效果良好。

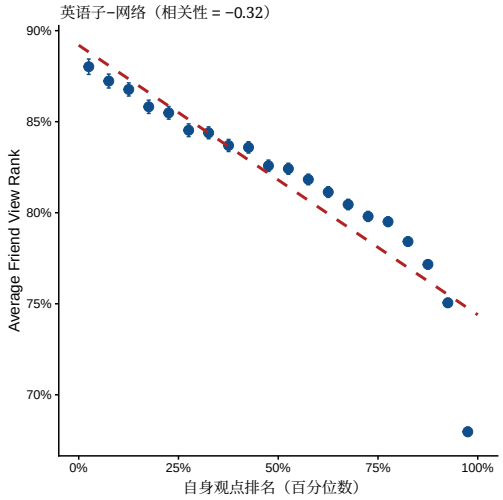
在附录 B.1 中，我们将针对连续标签的方法进一步应用于亚马逊商品的共同购买数据。

8 讨论

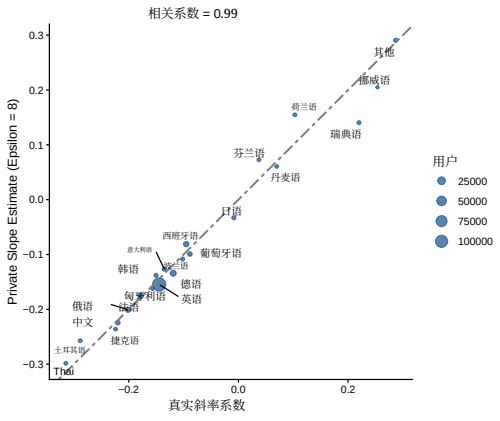
本文提出了一种实用方法，可在边相邻差分隐私下发布网络连通性指数。我们研究了标签为二元标签和连续标签的情形。我们的方法分两步进行：首先直接对节点属性进行隐私处理，然后对得到的统计量进行解析去偏，并添加第二层噪声，以保护单条边是否存在。我们为这两种方法都提供了形式化隐私保证，证明了所得估计量的一致性和渐近正态性，并通过模拟和实证应用表明，该方法在提供隐私保护的同时，即使在规模相对较小的网络中，也能保留实用性。

网络连通性指数能让政策制定者深入了解其所在地区及其管理机构中连接结构的细致、局部特征；我们的方法则使网络连通性指数能够在发布的同时，确保数据中个体的隐私不受侵犯。更广泛地说，我们的研究有助于更轻松地发布各种社会和经济网络的统计量，或具有类图结构的数据集，从而解答社会科学中的各类研究问题。

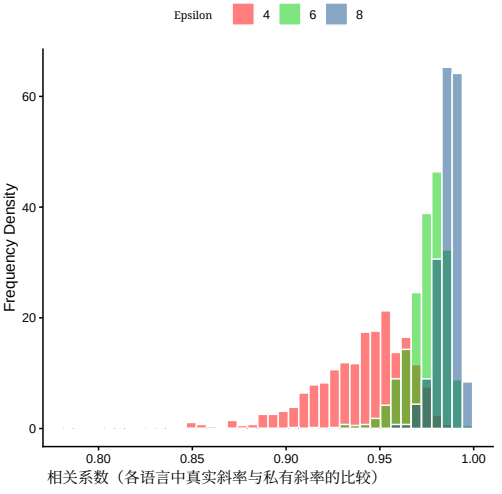
我们的标签优先方法有两个重要优点。首先，它避开了这样一个问题：在最坏情况下的图上，网络连通性统计量具有很高的全局敏感度。其次，它也避开了组合问题；这些问题源于一个节点的属性可能作为输入，影响大量单元。由于这两个问题是许多网络统计量的共同特点，我们认为，标签优先方法很有前景，可用于在隐私保证下发布由带标签网络导出的各种统计量。通过多项随机响应，将标签优先方法扩展到多项属性，是顺理成章的下一步。（例如，个人的就业状态不一定只有就业或失业，也可能是全职工作、兼职工作或未工作。又如，个人的婚姻状况不一定只有已婚或单身，也可能是已婚、从未结婚或离异；个人的受教育程度也可能是高中学历、大学学历或研究生学历。）



（a）英语子网络中平均好友排名与自身排名的关系。



（b）在 $\varepsilon = 8$时，一次模拟中各语言的私有斜率系数与真实斜率系数的比较。



(c) 按 ε划分的模拟中隐私斜率系数与真实斜率系数之间相关性的分布。

图 8. **Twitch 网络中按观看次数排名划分的连通性，以及差分隐私估计量的表现**。面板a绘制了英语子网络中平均好友排名与自身排名的关系。面板b展示了在 $\varepsilon = 8$下，一次模拟中不同语言的真实回归斜率估计值和差分隐私估计值；虚线为45度线。面板c展示了$\varepsilon \in \{4、6、8\}$对应的模拟迭代中真实斜率系数与差分隐私斜率系数之间相关性的分布。

然而，从更广泛的角度来看，可能有许多网络统计量顽固地难以通过隐私保护方法加以保护。例如，Chandrasekhar 等（2024）指出，与扩散过程的范围（例如病毒的传播）相关的结果，往往会对初始节点特征或边集的任何微小变化高度敏感。未来的研究可以探讨动态网络过程中的网络扩散指标的隐私估计（例如，测量感染邻居的比例，以评估不同地区的风险暴露），并将其扩展至高阶网络邻域（例如，两跳范围内或通过超边产生的暴露）。

参考文献

Alabi, Daniel、Audra McMillan、Jayshree Sarathy、Adam Smith 和 Salil Vadhan（2022）。《差分隐私简单线性回归》。载于：隐私增强技术会议录（*PoPETs*）（2），第184–204页。Bailey, Michael、Drew Johnston、Theresa Kuchler、Ayush Kumar 和 Johannes Stroebel（2025）。《世界各地的跨性别社会联系》。载于：美国经济学会论文与会议录。第115卷，第132–138页。Banerjee, Abhijit、Arun G. Chandrasekhar、Esther Duflo 和 Matthew O. Jackson（2013年）。《小额信贷的扩散》。载于：《科学》341.1236498。Bassily, Raef、Adam Groce、Jonathan Katz 和 Adam Smith（2013年10月）。《耦合世界隐私：利用统计数据隐私中的对抗性不确定性》。载于：*2013 IEEE*第*54*届计算机科学基础年度研讨会，第439–448页。doi：10.1109/FOCS.2013.54。Blocki, Jeremiah、Avrim Blum、Anupam Datta 和 Or Sheffet（2013年）。《通过受限敏感度进行社交网络的差分隐私数据分析》。载于：第*4*届理论计算机科学创新会议论文集，第87–96页。Chandrasekhar, Arun G、Paul Goldsmith-Pinkham、Tyler McCormick、Samuel A Thau 和 Jerry Wei（2024）。测量误差下网络扩散估计值的非稳健性。技术报告。美国国家经济研究局。Chen, Xihui、Sjouke Mauw 和 Yunior Ramírez-Cruz（2020）。《以差分隐私保证发布保持社区结构的属性社交图》。载于：隐私增强技术会议录4，第131–152页。url：https://api.semanticscholar.org/CorpusID:202540124。Chetty, Raj 和 John N Friedman（2019）。《在披露基于小样本的统计量时降低隐私损失的实用方法》。载于：《隐私与保密性杂志》9.2。Chetty, Raj、John N Friedman、Nathaniel Hendren、Maggie R Jones 和 Sonya R Porter（2026）。《机会地图集：绘制社会流动的童年根源》。载于：《美国经济评论》116.1，第1–51页。Chetty, Raj、Matthew O. Jackson、Theresa Kuchler、Johannes Stroebl、Nathan Hendren、Robert Fluegge、Sara Gong、Federico Gonzalez、Armelle Grondin、Matthew Jacob、Drew Johnston、Martin Koenen、Eduardo Laguna-Muggenburg、Florian Mudekereza、Tom Rutter、Nicolaj Thor、Wilbur Townsend、Ruby Zhang、Mike Bailey、Pablo Barbera、Monica Bhole 和 Nils Wernerfelt（2022a年）。《美国社会资本 I：衡量及其与经济流动的关联》。载于：《自然》608.7921。doi：doi10.1038/s41586-022-04996-4。–（2022b年）。《美国社会资本 II：经济连通性的决定因素》。载于：《自然》608.7921。doi：10.1038/s41586-022-04997-3。Dick, Travis、Cynthia Dwork、Michael Kearns、Terrance Liu、Aaron Roth、Giuseppe Vietri 和 Zhiwei Steven Wu（2023）。《从已发布统计数据中按置信度排名重构人口普查微观数据》。载于：《美国国家科学院院刊》120.8，e2218605120。Dwork, Cynthia、Frank McSherry、Kobbi Nissim 和 Adam Smith（2006）。《根据隐私数据分析中的敏感度校准噪声》。载于：密码学理论：第三届密码学理论会议，*TCC 2006*，美国纽约州纽约市，*2006*年*3*月*4*日至*7*日。会议录*3*。Springer，第265–284页。

Dwork, Cynthia 和 Aaron Roth（2014年）。《差分隐私的算法基础》。载于：理论计算机科学基础与趋势9.3–4，第211–407页。Easley, David 和 Jon Kleinberg（2010）。网络、群体与市场：对高度连通世界的推理。剑桥大学出版社。Gehrke, Johannes、Edward Lui 和 Rafael Pass（2011年）。《迈向社交网络隐私：一种基于零知识的隐私定义》。载于：密码学理论会议。Springer，第432–449页。Geng, Quan、Wei Ding、Ruiqi Guo 和 Sanjiv Kumar （ 2018 ）。《用于近似差分隐私的截断拉普拉斯机制》。载于：*arXiv* 预印本 *arXiv:1810.00877*。–（2020）。《近似差分隐私中隐私与效用权衡的紧致分析》。载于：人工智能与统计国际会议。PMLR，第89–99页。Gentzkow, Matthew（2007）。《在具有互补性的模型中评估新产品：在线报纸》。载于：《美国经济评论》97.3，第713–744页。Ghosh, Arpita 和 Robert Kleinberg（2016）。《差分隐私机制的推断隐私保证》。载于：*arXiv*预印本 *arXiv:1603.01508*。Goyal, Sanjeev（2023）。网络：一种经济学方法。麻省理工学院出版社。Greene,WilliamH.（2012）。《计量经济分析》。第7版。波士顿：培生教育。isbn：9780131395381。Harris, Tom、Shankar Iyer、Tom Rutter、Guanghua Chi、Drew Johnston、Patrick Lam、Lucy Makinson、Antonio S Silva、Martin Wessel、Mei-Chen Liou、Yingcan Wang、Qamar Zaman 和 Michael Bailey（2025）。英国的社会资本：来自六十亿份友谊关系的证据。技术报告。开放科学中心。Hehir, Jonathan、Xiaoyue Niu 和 Aleksandra Slavkovic（2025）。《解读网络差分隐私》。载于：*arXiv*预印本 *arXiv:2504.12520*。Holohan, Naoise、Spiros Antonatos、Stefano Braghin 和 Pól Mac Aonghusa（2019年12月）。《差分隐私中的有界拉普拉斯机制》。载于：《隐私与保密性杂志》10.1。doi：10.29012/jpc.715。url：https://journalprivacyconfidentiality.org/index.php/jpc/article/ view/715。Jackson, Matthew O.（2019）。《人类网络：你的社会位置如何决定你的权力、信念与行为》。万神殿出版社：纽约。Jorgensen, Zach、Ting Yu 和 Graham Cormode（2016）。《以形式化隐私保证发布带属性的社交图》。载于：*2016*年国际数据管理会议论文集，第107–122页。Kasiviswanathan, Shiva Prasad、Kobbi Nissim、Sofya Raskhodnikova 和 Adam Smith（2013年）。《使用节点差分隐私分析图》。载于：密码学理论会议。Springer，第457–476页。Kifer, Daniel 和 Ashwin Machanavajjhala（2011年6月）。《数据隐私中没有免费午餐》。载于：*2011*年*ACM SIGMOD*数据管理国际会议论文集。SIGMOD '11。计算机协会，第193–204页。isbn：978-1-4503-0661-4。doi：10.1145/1989323.1989345。url：https://doi.org/10.1145/1989323.1989345（访问于2022年02月03日）。–（2014年）。《Pufferfish：数学化隐私定义的框架》。载于：《*ACM*数据库系统汇刊》（*TODS*）39.1，第1–36页。Leskovec, Jure、Lada A Adamic 和 Bernardo A Huberman（2007）。《病毒式营销的动态》。载于：《*ACM*网络汇刊》（*TWEB*）1.1，5–es。Liu, Changchang、Supriyo Chakraborty 和 Prateek Mittal（2016）。《依赖性使你易受攻击：相关元组下的差分隐私》。载于：*NDSS*。第16卷，第21–24页。Narayanan, Arvind 和 Vitaly Shmatikov（2008）。《大型稀疏数据集的稳健去匿名化》。载于：*2008 IEEE*安全与隐私研讨会（*sp 2008*）。IEEE，第111–125页。

Nissim、Kobbi、Sofya Raskhodnikova 和 AdamSmith（2007年6月）。“隐私数据分析中的平滑敏感度与抽样”。载于：第三十九届 *ACM* 计算理论研讨会年会论文集。STOC ’07。美国纽约州纽约：计算机协会，第 75–84 页。

Rozemberczki、Benedek 和 Rik Sarkar（2021年）。*Twitch* 玩家：用于评估邻近性保持和基于结构角色的节点嵌入的数据集。arXiv：2101。03091 [cs.SI]。

Song、Shuang、Yizhen Wang 和 Kamalika Chaudhuri（2017年）。“针对相关数据的 Pufferfish 隐私机制”。载于：*2017*年 *ACM* 国际数据管理会议论文集，第 1291–1306 页。

Wang、Yue、XintaoWu 和 Donghui Hu（2016年）。“使用随机响应实现差分隐私数据收集。”载于：*EDBT/ICDT*研讨会。第 1558 卷，第 0090–6778 页。

Yang、Jaewon 和 Jure Leskovec（2012）。“基于真实值定义和评估网络社区”。载于：*ACM SIGKDD*数据语义挖掘研讨会论文集，第1–8页。Zhao、Jun、Junshan Zhang 和 H Vincent Poor（2017）。“针对相关数据的依赖差分隐私”。载于：*2017*年 *IEEE Globecom*研讨会（*GC Wkshps*）。IEEE，第1–7页。

A 证明

A.1 定理1的证明：(ε,δ) 边邻接下的组合

证明。设 $(\mathrm{V},\mathrm{E},\mathbf{L})$ 和 $(\mathrm{V},\mathrm{E}',\mathbf{L}')$ 是边相邻带标签网络。任取可测集 $\widehat{\mathrm{S}} \subseteq \mathbb{R}$。通过对 M_1，的输出$\mathbf{L}$ 进行条件化

$$\Pr[\mathcal{M}(\mathcal{V},\mathcal{E},\mathbf{L}) \in \mathcal{S}] = \sum_{\widehat{\mathbf{L}}} \Pr\left[\mathcal{M}_1(\mathcal{V},\mathcal{E},\mathbf{L}) = \widehat{\mathbf{L}}\right] \Pr\left[\mathcal{M}_2(\mathcal{V},\mathcal{E},\widehat{\mathbf{L}}) \in \mathcal{S}\right].$$

由于$\mathbf{L}$和$\mathbf{L}'$至多在一个节点属性上不同，且 M_1关于节点属性满足 $(\varepsilon_\ell,\delta_\ell)$-差分隐私，因此我们有

$$\Pr\left[\mathcal{M}_1(\mathcal{V},\mathcal{E},\mathbf{L}) = \widehat{\mathbf{L}}\right] \le e^{\varepsilon_\ell} \Pr\left[\mathcal{M}_1(\mathcal{V},\mathcal{E}',\mathbf{L}') = \widehat{\mathbf{L}}\right] + \delta_\ell(\widehat{\mathbf{L}}),$$

因此，

$$\Pr[\mathcal{M}(\mathcal{V},\mathcal{E},\mathbf{L}) \in \mathcal{S}] \le e^{\varepsilon_\ell} \sum_{\widehat{\mathbf{L}}} \Pr\left[\mathcal{M}_1(\mathcal{V},\mathcal{E}',\mathbf{L}') = \widehat{\mathbf{L}}\right] \Pr\left[\mathcal{M}_2(\mathcal{V},\mathcal{E},\widehat{\mathbf{L}}) \in \mathcal{S}\right] + \delta_\ell,$$

由于 E 和 $\widehat{\ }$d E′至多相差一条边，且 M_2满足（ε_e， δ_e）边相邻差分隐私（对任意固定的 $\mathbf{L}$），

$$\Pr\left[\mathcal{M}_2(\mathcal{V},\mathcal{E},\widehat{\mathbf{L}}) \in \mathcal{S}\right] \le e^{\varepsilon_e} \Pr\left[\mathcal{M}_2(\mathcal{V},\mathcal{E}',\widehat{\mathbf{L}}) \in \mathcal{S}\right] + \delta_e.$$

$\widehat{\ }$入此界到前一个不等式中，并使用 $\sum_{\sim\mathbf{L}}\Pr[\mathrm{M}_1(\cdot) = \mathbf{L}] = 1$ gi

代入 ve

$$\begin{aligned}\Pr[\mathcal{M}(\mathcal{V},\mathcal{E},\mathbf{L}) \in \mathcal{S}] &\le e^{\varepsilon_\ell+\varepsilon_e} \sum_{\widehat{\mathbf{L}}} \Pr\left[\mathcal{M}_1(\mathcal{V},\mathcal{E}',\mathbf{L}') = \widehat{\mathbf{L}}\right] \Pr\left[\mathcal{M}_2(\mathcal{V},\mathcal{E}',\widehat{\mathbf{L}}) \in \mathcal{S}\right] + \delta_\ell + \delta_e \\ &= e^{\varepsilon_\ell+\varepsilon_e} \Pr[\mathcal{M}(\mathcal{V},\mathcal{E}',\mathbf{L}') \in \mathcal{S}] + \delta_\ell + e^{\varepsilon_\ell}\delta_e,\end{aligned}$$

由此可证明（$\varepsilon_\ell + \varepsilon_e$， $\delta_\ell + e^{\varepsilon_\ell}\delta_e$）边相邻差分隐私。 □

A.2 命题 1 的证明：个体连通性的最小方差无偏估计量

证明。我们首先证明无偏性。在随机响应机制下，

$$\mathbb{E}[\mathbf{1}\{\hat{l}_j = b\} \mid l_j] = (1-p)\mathbf{1}\{l_j = b\} + p\mathbf{1}\{l_j = a\} = (1-2p)\mathbf{1}\{l_j = b\} + p.$$

因此，

$$\begin{aligned}\mathbb{E}[\hat{\rho}_i \mid \mathbf{L}] &= \sum_{j\in\mathcal{V}} a_{ij}\,\mathbb{E}[\mathbf{1}\{\hat{l}_j = b\} \mid l_j] \\ &= \sum_{j\in\mathcal{V}} a_{ij}\Big((1-2p)\mathbf{1}\{l_j = b\} + p\Big) \\ &= (1-2p)\sum_{j\in\mathcal{V}} a_{ij}\mathbf{1}\{l_j = b\} + p\sum_{j\in\mathcal{V}} a_{ij} \\ &= (1-2p)\rho_i + p,\end{aligned}$$

因为 $\sum_j a_{ij} = 1$。重新整理可得

$$\rho_i = \frac{\mathbb{E}[\hat{\rho}_i \mid \mathbf{L}] - p}{1-2p},$$

这意味着 $\tilde{\rho}_i = (\hat{\rho}_i - p)/(1-2p)$ 是无偏的。

^ 接下来，我们通过 Lehmann–Scheffé 定理证明其唯一性。隐私化标签 $\{l_j\}$ 是独立生成的，其似然为

$$f(\hat{l}_j \mid l_j) = \big((1-p)\mathbf{1}\{l_j = \hat{l}_j\} + p\mathbf{1}\{l_j \neq \hat{l}_j\}\big),$$

因此，联合似然可分解为

$$f(\hat{\mathbf{L}} \mid \mathbf{L}) = \prod_{j\in\mathcal{V}} f(\hat{l}_j \mid l_j),$$

^这意味着$\mathbf{L}$ 是$\mathbf{L}$ 的充分统计量，这是由 Neyman因子分解定理得出的 .

^ 完备性源于似然严格为正：在任意$\mathbf{L}$下，每种配置$\mathbf{L}$ 的概率都严格大于零，因此对于任意可测函数 h,

$$\mathbb{E}[h(\hat{\mathbf{L}})] = 0 \text{ for all } \mathbf{L} \quad \Rightarrow \quad h(\hat{\mathbf{L}}) = 0 \text{ a.s.}$$

^因此，$\mathbf{L}$ 是$\mathbf{L}$的完备充分统计量；由于 $\tilde{\rho}_i$ 是一个无偏估计量，且仅通过数据中的$\mathbf{L}$而依赖于数据，Lehmann–Scheffé定理表明， $\tilde{\rho}_i$ 是ρ_i的唯一最小方差无偏估计量。□

A.3 定理 2 的证明：去偏私有估计量的一致性

在证明定理 2 中隐私Hájek比率估计量的一致性之前，我们先给出引理1的证明，该引理讨论分子中加权和的边敏感度，随后再给出另外三个引理。

引理1的证明。切换单条边（u, v）只会改变归一化权重行 $a_{u\cdot}$和 $a_{v\cdot}$，因此只有^ρ_u, ^ρ_v （进而只有 $\widetilde{\rho}_u, \widetilde{\rho}_v$）可能发生变化；其他所有 $\widetilde{\rho}_i$ 都保持不变。^此外， w_i 仅取决于 l_i，因此不受影响。

Thus

$$|S_{1,n}(E) - S_{1,n}(E')| = |w_u(\widetilde{\rho}_u(E) - \widetilde{\rho}_u(E')) + w_v(\widetilde{\rho}_v(E) - \widetilde{\rho}_v(E'))| \le |w_u|\Delta\widetilde{\rho}_u + |w_v|\Delta\widetilde{\rho}_v.$$

由于 ^$\rho_i \in [0,1]$，我们有

$$\widetilde{\rho}_i = \frac{\hat{\rho}_i - p}{1-2p} \in \left[-\frac{p}{1-2p}, \frac{1-p}{1-2p}\right],$$

因此 $|\widetilde{\rho}_i(E) - \widetilde{\rho}_i(E')| \le \frac{1}{1-2p}$。利用 $|w_u|, |w_v| \le W$可得

$$|S_{1,n}(E) - S_{1,n}(E')| \le 2W \cdot \frac{1}{1-2p} = \frac{2W}{1-2p}.$$

□

引理 2（去偏恒等式）。给定网络 E 和真实标签向量$\mathbf{L} = (l_i)_{i\in V}$的条件下：

$$\mathbf{E}[w_i \mid E, \mathbf{L}] = \mathbf{1}\{l_i = a\}, \tag{7}$$

$$\mathbf{E}[\widetilde{\rho}_i \mid E, \mathbf{L}] = \rho_i. \tag{8}$$

ˆ ˆ此外，对于每个固定的 i， w_i 仅取决于 l_i，而 $\widetilde{\rho}_i$ 取决于$(l_j)_{j\in V}$；在独立的ˆ扰动下，给定$(E, \mathbf{L})$后， w_i与$(l_j)_{j\neq i}$相互独立。

ˆ 证明。对于 (7)，当 $l_i = a, \mathbb{P}(l_i = a) = 1 - p$ 时，因此

$$\mathbf{E}[w_i \mid E, l_i = a] = \frac{(1-p) - p}{1-2p} = 1.$$

ˆ 当 $l_i = b, \mathbb{P}(l_i = a) = p$ 时，因此

$$\mathbf{E}[w_i \mid E, l_i = b] = \frac{p - p}{1-2p} = 0.$$

对于 (8)，注意，对于任意 j,

$$\mathbf{E}[\mathbf{1}\{\hat{l}_j = b\} \mid E, l_j] = \begin{cases} 1-p, & l_j = b, \\ p, & l_j = a, \end{cases} = p + (1-2p)\mathbf{1}\{l_j = b\}.$$

取条件期望并利用线性性,

$$\mathbf{E}[\hat{\rho}_i \mid E, \mathbf{L}] = \sum_j a_{ij}\Big(p + (1-2p)\mathbf{1}\{l_j = b\}\Big) = p + (1-2p)\sum_j a_{ij}\mathbf{1}\{l_j = b\} = p + (1-2p)\rho_i.$$

Thus

$$\mathbf{E}[\widetilde{\rho}_i \mid E, \mathbf{L}] = \frac{\mathbf{E}[\hat{\rho}_i \mid E, \mathbf{L}] - p}{1-2p} = \rho_i.$$

□

引理 3（$S_{0,n}$的集中性）。令 $W :=$最大值$\{\frac{p}{1-2p}, \frac{1-p}{1-2p}\}$。若 $p < \frac{1}{2}$，则 $|w_i| \le W$几乎必然成立，且

$$\mathbb{P}\left(\left|\frac{S_{0,n}}{n} - \frac{\mathbf{E}[S_{0,n}]}{n}\right| > t\right) \le 2\exp\left(-\frac{2nt^2}{(2W)^2}\right), \qquad t > 0.$$

特别地，$S_{0,n} - \mathbf{E}[S_{0,n}] \xrightarrow{p} 0$。

ˆ证明。每个 w_i都只是 L_i的函数，因此在假设 1 下，$(w_i)_{i\in V}$相互独立。它们有界于 W。对和式 $S_{0,n} = \sum_i w_i$应用霍夫丁不等式。□

引理 4（有界差分：关于 $S_{1,n}$）。设 ˆ$\mathbf{L}$ 和 ˆ$\mathbf{L}'$ 仅在节点 k 处不同（即，对于所有$j \neq k$, ˆ $l_j = \hat{l}'_j$）。在假设 2（有界度）下，

$$\left|S_{1,n}(\hat{\mathbf{L}}) - S_{1,n}(\hat{\mathbf{L}}')\right| \le c_k,$$

其中可以取

$$c_k \;\le\; \frac{2W}{1-2p}\Big(1+\sum_{i\in V} a_{ik}\Big) \;\le\; \frac{2W}{1-2p}(1+\Gamma) \;=:\; C. \tag{9}$$

ˆ 证明。改变 l_k 会影响：(i) $S_{1,n}$ 中第 k 个加项的权重 w_k；以及 (ii) 每个 ˆρ_i 内的指示变量 ˆ$\mathbf{1}\{l_k = b\}$。

首先，由于 $|w_k| \le W$ 且 $|\widetilde{\rho}_k| \le \frac{1}{1-2p}$（因为 ˆ$\rho_k \in [0,1]$），第 k 个加项的变化至多为

$$\frac{2W}{1-2p}.$$

其次，对于 $i \ne k$，w_i 不变，只有 ˆρ_i 会通过索引为 k 处的单个指示变量发生变化：

$$\big|\hat{\rho}_i(\hat{\mathbf{L}}) - \hat{\rho}_i(\hat{\mathbf{L}}')\big| = \left|\sum_j a_{ij}\big(\mathbf{1}\{\hat{l}_j = b\} - \mathbf{1}\{\hat{l}'_j = b\}\big)\right| \le a_{ik}.$$

因此，

$$\big|\widetilde{\rho}_i(\hat{\mathbf{L}}) - \widetilde{\rho}_i(\hat{\mathbf{L}}')\big| \le \frac{a_{ik}}{1-2p}, \quad \Rightarrow \quad \big|w_i\widetilde{\rho}_i(\hat{\mathbf{L}}) - w_i\widetilde{\rho}_i(\hat{\mathbf{L}}')\big| \le \frac{W a_{ik}}{1-2p}.$$

对 $i \ne k$ 求和可得，额外变化的上界为

$$\frac{W}{1-2p}\sum_{i\in V} a_{ik}.$$

加上第 k项的界，得到 (9)。 □

定理 2 的证明。由公式 (7)，

$$\mathbf{E}[S_{0,n} \mid E, \mathbf{L}] = \sum_i \mathbf{1}\{l_i = a\} = \#(\mathcal{A}),$$

因此 $\mathbf{E}[S_{0,n}] = \#(\mathbf{A})$（将 E,$\mathbf{L}$ 视为固定，若其为随机变量，则取外期望）。此外，

$$\mathbf{E}[S_{1,n} \mid E, \mathbf{L}] = \sum_i \mathbf{E}[w_i\widetilde{\rho}_i \mid E, \mathbf{L}].$$

同理，对每个 i,

$$\mathbf{E}[w_i\widetilde{\rho}_i \mid E, \mathbf{L}] = \mathbf{1}\{l_i = a\}\rho_i,$$

因此

$$\mathbf{E}[S_{1,n} \mid E, \mathbf{L}] = \sum_{i\in A}\rho_i, \qquad \Rightarrow \qquad \mathbf{E}[S_{1,n}] = \sum_{i\in A}\rho_i.$$

由引理 3,

$$\frac{S_{0,n}}{n} - \frac{\mathbf{E}[S_{0,n}]}{n} \xrightarrow{p} 0.$$

此外，引理 4 给出了映射 ˆ$\mathbf{L} \mapsto S_{1,n}(\hat{\mathbf{L}})$ 的有界差分常数（关于 k 一致），因此由 McDiarmid 不等式可得

$$\frac{S_{1,n}}{n} - \frac{\mathbf{E}[S_{1,n}]}{n} \xrightarrow{p} 0.$$

根据假设 1，$\mathbf{E}[S_{0,n}]/n = \#(\mathbf{A})/n \to \pi_{\mathcal{A}} > 0$。结合$S_{0,\,n}/n \xrightarrow{p} \mathbf{E}[S_{0,n}]/n$，我们有$S_{0,\,n}/n \xrightarrow{p} \pi_{\mathcal{A}}$，因此

$$\mathbb{P}\big(S_{0,n}/n > \pi_{\mathcal{A}}/2\big) \to 1.$$

特别地，$S_{0,\ n} > 0$ 以趋于 1 的概率成立，因此 $\widetilde{C}^{A\to B} = S_{1,\ n}/S_{0,\ n}$ 以趋于 1 的概率有定义。

考虑向量

$$\left(\frac{S_{1,n}}{n}, \frac{S_{0,n}}{n}\right) \xrightarrow{p} \left(\frac{\mathbf{E}[S_{1,n}]}{n}, \frac{\mathbf{E}[S_{0,n}]}{n}\right) = \left(\frac{1}{n}\sum_{i\in A}\rho_i,\ \frac{\#(\mathcal{A})}{n}\right).$$

映射 $f(x,y) = x/y$ 在 $\{(x,y):\ y \neq 0\}$ 上连续。由于第二个分量依概率收敛于 $\pi_{\mathcal{A}} > 0$，由连续映射定理可得

$$\widetilde{C}^{\mathcal{A}\to\mathcal{B}} = \frac{S_{1,n}/n}{S_{0,n}/n} \xrightarrow{p} \frac{\frac{1}{n}\sum_{i\in\mathcal{A}}\rho_i}{\#(\mathcal{A})/n} = \frac{1}{\#(\mathcal{A})}\sum_{i\in A}\rho_i = C^{\mathcal{A}\to\mathcal{B}}.$$

令 $Z_n \mid S_{0,n} \sim$ 拉普拉斯$\left(0,\ \sqrt{\frac{2}{(1-p)\,(1-2p)^2\varepsilon_e S_{0,n}}}\right)$。任取 $t > 0$ 并定义

$$\mathcal{G}_n := \left\{S_{0,n}/n > \pi_{\mathcal{A}}/2\right\}.$$

则 $\mathbb{P}(\mathcal{G}_n) \to 1$。在 $\mathcal{G}_n$ 上，我们有

$$\frac{2(1-p)}{(1-2p)^2\varepsilon_e S_{0,n}} \le \frac{4(1-p)}{(1-2p)^2\varepsilon_e \pi_{\mathcal{A}} n}.$$

利用拉普拉斯尾界 $\mathbb{P}(|\mathrm{Lap}(0,b)| > t) = \exp(-t/b)$，可得

$$\begin{aligned}
\mathbb{P}(|Z_n| > t) &\le \mathbb{P}(\mathcal{G}_n^c) + \mathbb{P}(|Z_n| > t, \mathcal{G}_n) \\
&\le \mathbb{P}(\mathcal{G}_n^c) + \exp\left(-\frac{t}{2(1-p)/((1-2p)^2\varepsilon_e S_{0,n})}\right)\Big|_{\mathcal{G}_n} \\
&\le \mathbb{P}(\mathcal{G}_n^c) + \exp\left(-\frac{(1-2p)^2\varepsilon_e\pi_{\mathcal{A}} n}{4(1-p)}\,t\right) \longrightarrow 0,
\end{aligned}$$

因此 $Z_n \xrightarrow{p} 0$。

最后，$\widetilde{C}^{\mathcal{A}\to\mathcal{B}} \xrightarrow{p} C^{\mathcal{A}\to\mathcal{B}}$，由斯卢茨基定理可得

$$\widehat{C}^{\mathcal{A}\to\mathcal{B}}_{\mathrm{DP}} = \widetilde{C}^{\mathcal{A}\to\mathcal{B}} + Z_n \xrightarrow{p} C^{\mathcal{A}\to\mathcal{B}}.$$

□

A.4 定理 3 的证明：二元标签的边相邻差分隐私保证

证明。考虑输入的带标签网络 $(\mathbf{V}, \mathbf{E}, \mathbf{L})$。我们通过将算法 1 分解为两个阶段来分析其隐私性。

在第一阶段，算法对 $\mathbf{L}$ 中的每个节点标签独立应用随机响应，翻转概率为 $p = \frac{1}{1+e^{\varepsilon_\ell}}$，从而生成扰动标签 $\hat{\mathbf{L}}$。根据事实 1，该机制相对于单个节点标签的变化满足 ε_ℓ-差分隐私。因此，映射

$$(\mathcal{V}, \mathcal{E}, \mathbf{L}) \longmapsto (\mathcal{V}, \mathcal{E}, \hat{\mathbf{L}})$$

满足 ε_ℓ 节点级差分隐私。

Fix $\hat{\mathbf{L}}$，使得权重 $(w_i)_{i\in\mathcal{V}}$ 是确定的。设 $\mathbf{E}$ 和 $\mathbf{E}'$ 为两个边相邻边集，它们相差一条无向边的添加或删除。根据引理 2，统计量

$$S_{1,n}(\mathcal{E}) = \sum_{i\in\mathcal{V}} w_i\,\tilde{\rho}_i(\mathcal{E})$$

具有有界的边敏感度，即

$$\left|S_{1,n}(\mathcal{E}) - S_{1,n}(\mathcal{E}')\right| \;\le\; \frac{2(1-p)}{(1-2p)^2}.$$

因此，根据拉普拉斯机制，发布 Hájek估计量

$$\frac{\widehat{S}_{1,n}}{S_{0,n}} = \frac{S_{1,n}(\mathcal{E})}{S_{0,n}} + Z_n, \qquad Z_n \sim \mathrm{Lap}\!\left(0, \frac{2(1-p)}{(1-2p)^2\,\varepsilon_e\,S_{0,n}}\right),$$

ˆ 在给定L的条件下满足 ε_e 边相邻差分隐私。用于计算 w_i、 $\tilde{\rho}_i$、 $S_{0,\ n}$和$S_{1,\ n}$的任何确定性后处理都会保留这一保证。

最后，算法 1 是 ε_ℓ-差分隐私标签随机化机制与ε_e 边隐私发布的组合。根据引理1（组合定理），整体机制对于输入 (V、E、L)满足$(\varepsilon_\ell + \varepsilon_e)$边相邻差分隐私。□

A.5 定理 4 的证明：渐近正态性与扰动阶

证明。令

$$I_i := \mathbf{1}\{\hat{I}_i = a\}, \qquad w_i = \frac{I_i - p}{1-2p}, \qquad \tilde{\rho}_i = \frac{\hat{\rho}_i - p}{1-2p} = \frac{1}{1-2p}\left(\sum_{j\in N(i)} a_{ij}(1-I_j) - p\right),$$

其中

$$a_{ij} := \frac{e_{ij}}{d_i}, \qquad d_i = \sum_{j\in V_n} e_{ij}.$$

定义节点级向量

$$Z_i := \begin{pmatrix} w_i \\ y_i \end{pmatrix}, \qquad y_i := w_i\tilde{\rho}_i, \qquad T_n := \begin{pmatrix} S_{0,n} \\ S_{1,n} \end{pmatrix} = \sum_{i=1}^{n} Z_i.$$

由于随机响应是在各节点间独立应用的，因此变量 $\{I_i\}_{i=1}^n$ 相互独立。注意， w_i 仅依赖于 I_i，而 $\tilde{\rho}_i$ 仅依赖于 $\{I_j\colon\ j \in N(i)\}$。因此， Z_i 关于以下内容可测

$$\sigma\big(I_i, \{I_j : j \in N(i)\}\big).$$

因此，若 $\mathrm{dist}(i,\ j)\ \ge 3$，则 Z_i 和Z_j 所使用的随机响应变量集合互不相交，因而相互独立，并且

$$\mathrm{Cov}(Z_i, Z_j) = 0, \qquad \mathrm{dist}(i,j) \ge 3.$$

因此，我们得到精确的两跳协方差分解

$$\frac{1}{n}\,\mathrm{Cov}(T_n) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(Z_i) + \frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{j\neq i:\\ \mathrm{dist}(i,j)\le 2}}\mathrm{Cov}(Z_i, Z_j). \tag{10}$$

由于存在一个常数 $\Delta < \infty$ （与 n无关），使得

$$\max_{i\in V_n} d_i \le \Delta,$$

每个 i 在图距离不超过 2 的范围内都只有有限个节点。因此，(10) 是局部协方差贡献之和。此外，下面推导出的显式公式表明

$$\frac{1}{n}\,\mathrm{Cov}(T_n) \to \Sigma,$$

因此，协方差(T_n) 关于 n是线性阶。

回顾 $Z_i := (w_i, y_i)^\top$，其中 $y_i := w_i \tilde{\rho}_i$且

$$w_i = \frac{I_i - p}{1-2p}, \qquad \tilde{\rho}_i = \frac{1}{1-2p}\left(\sum_{k \in N(i)} a_{ik}(1 - I_k) - p \right), \qquad a_{ik} = \frac{e_{ik}}{d_i}, \ \sum_{k \in N(i)} a_{ik} = 1, \ a_{ii} = 0.$$

假设不存在自环。由于随机响应是在各节点之间独立应用的， I_i 与$\{I_k : k \in N(i)\}$相互独立，因此

$$w_i \perp\!\!\!\perp \tilde{\rho}_i.$$

记

$$\mathrm{Var}(Z_i) = \begin{pmatrix} \mathrm{Var}(w_i) & \mathrm{Cov}(w_i, y_i) \\ \mathrm{Cov}(y_i, w_i) & \mathrm{Var}(y_i) \end{pmatrix}.$$

(i)$\mathrm{Var}(w_i)$。 由于 $w_i = \frac{I_i - p}{1-2p}$和 $\mathrm{Var}(I_i) = p(1-p)$,

$$\mathrm{Var}(w_i) = \frac{\mathrm{Var}(I_i)}{(1-2p)^2} = \frac{p(1-p)}{(1-2p)^2}. \tag{11}$$

(ii)$\mathrm{Cov}(w_i, \ y_i)$。 利用 $y_i = w_i \tilde{\rho}_i$和独立性 $w_i \perp\!\!\!\perp \tilde{\rho}_i$,
$\mathrm{Cov}(w_i, \ y_i) = \mathrm{Cov}(w_i, \ w_i \tilde{\rho}_i) = \mathbf{E}[w_i^2 \tilde{\rho}_i] - \mathbf{E}[w_i]\mathbf{E}[w_i \tilde{\rho}_i] = \mathbf{E}[\tilde{\rho}_i]\left(\mathbf{E}[w_i^2] - \mathbf{E}[w_i]^2 \right) = \mathrm{Var}(w_i)\ \mathbf{E}[\tilde{\rho}_i]$.接下来，根据随机响应去偏，对每个 k,

$$\mathbf{E}\left[\frac{(1 - I_k) - p}{1-2p} \right] = \mathbf{1}\{l_k = b\},$$

因此

$$\mathbf{E}[\tilde{\rho}_i] = \sum_{k \in N(i)} a_{ik}\, \mathbf{1}\{l_k = b\} =: \rho_i. \tag{12}$$

结合（11）和（12），

$$\mathrm{Cov}(w_i, y_i) = \frac{p(1-p)}{(1-2p)^2}\, \rho_i. \tag{13}$$

（*iii*）方差(y_i) =方差$(w_i \tilde{\rho}_i)$。再次使用 $w_i \perp\!\!\!\perp \tilde{\rho}_i$,

$$\mathrm{Var}(y_i) = \mathrm{Var}(w_i \tilde{\rho}_i) = \mathbf{E}[w_i^2]\, \mathrm{Var}(\tilde{\rho}_i) + \mathrm{Var}(w_i)\big(\mathbf{E}[\tilde{\rho}_i] \big)^2. \tag{14}$$

我们首先计算方差$(\tilde{\rho}_i)$。由于

$$\tilde{\rho}_i = \frac{1}{1-2p}\left(\sum_{k \in N(i)} a_{ik}(1 - I_k) - p \right),$$

常数不会影响方差，且 $\{I_k : k \in N(i)\}$ 的独立性可得

$$\mathrm{Var}(\tilde{\rho}_i) = \frac{1}{(1-2p)^2}\, \mathrm{Var}\Big(\sum_{k \in N(i)} a_{ik}(1 - I_k) \Big) = \frac{1}{(1-2p)^2} \sum_{k \in N(i)} a_{ik}^2\, \mathrm{Var}(I_k) = \frac{p(1-p)}{(1-2p)^2} \sum_{k \in N(i)} a_{ik}^2. \tag{15}$$

接下来，计算 $\mathbf{E}[w_i^2]$。根据 $\mathbf{E}[w_i^2] = \mathrm{Var}(w_i) + (\mathbf{E}[w_i])^2$和随机响应无偏性,

$$\mathbf{E}[w_i] = \mathbf{1}\{l_i = a\} =: A_i,$$

so

$$\mathbf{E}[w_i^2] = \frac{p(1-p)}{(1-2p)^2} + A_i. \tag{16}$$

将 (15)、(16) 和 (12) 代入 (14)，得

$$\mathrm{Var}(y_i) = \left(\frac{p(1-p)}{(1-2p)^2} + A_i\right) \cdot \frac{p(1-p)}{(1-2p)^2} \sum_{k \in N(i)} a_{ik}^2 + \frac{p(1-p)}{(1-2p)^2}\rho_i^2. \tag{17}$$

由 (11)、(13) 和 (17),

$$\mathrm{Var}(Z_i) = \begin{pmatrix} \dfrac{p(1-p)}{(1-2p)^2} & \dfrac{p(1-p)}{(1-2p)^2}\,\rho_i \\ \dfrac{p(1-p)}{(1-2p)^2}\,\rho_i & \left(\dfrac{p(1-p)}{(1-2p)^2} + A_i\right) \cdot \dfrac{p(1-p)}{(1-2p)^2} \displaystyle\sum_{k \in N(i)} a_{ik}^2 + \dfrac{p(1-p)}{(1-2p)^2}\rho_i^2 \end{pmatrix}.$$

特别地，$\mathrm{Var}(Z_i)$ 的所有元素均由局部量 p、A_i、ρ_i 和 $\sum_{k \in N(i)} a_{ik}^2$ 决定，且每个局部量都有界。因此，$\mathrm{Var}(Z_i)$ 关于 i 和 n 一致有界。

对于 $i \neq j$,

$$\mathrm{Cov}(Z_i, Z_j) = \begin{pmatrix} \mathrm{Cov}(w_i, w_j) & \mathrm{Cov}(w_i, y_j) \\ \mathrm{Cov}(y_i, w_j) & \mathrm{Cov}(y_i, y_j) \end{pmatrix}.$$

(a)$\mathrm{Cov}(w_i,\ w_j) = 0$ 对于 $i \neq j$。由于 w_i 和 w_j 相互独立,

$$\mathrm{Cov}(w_i, w_j) = 0, \qquad (i \neq j). \tag{18}$$

(b)$\mathrm{Cov}(w_i,\ y_j)$。固定 $i \neq j$。由于 $y_j = w_j\tilde{\rho}_j$，且对于 $i \neq j$，随机变量 w_j 与 $(w_i,\ \tilde{\rho}_j)$ 独立，我们有

$$\mathrm{Cov}(w_i, y_j) = \mathrm{Cov}(w_i, w_j\tilde{\rho}_j) = \mathbf{E}[w_j]\,\mathrm{Cov}(w_i, \tilde{\rho}_j).$$

此外,

$$\tilde{\rho}_j = \frac{1}{1-2p}\left(\sum_{k \in N(j)} a_{jk}(1 - I_k) - p\right),$$

因此，协方差$(w_i,\ \tilde{\rho}_j) \neq 0$仅当 $i \in N(j)$ 时成立，而在这种情况下

$$\mathrm{Cov}(w_i, \tilde{\rho}_j) = \frac{a_{ji}}{1-2p}\,\mathrm{Cov}(w_i, 1 - I_i) = -\frac{a_{ji}}{1-2p}\,\mathrm{Cov}(w_i, I_i).$$

由于

$$w_i = \frac{I_i - p}{1-2p}, \qquad \mathrm{Cov}(w_i, I_i) = \frac{\mathrm{Var}(I_i)}{1-2p} = \frac{p(1-p)}{1-2p},$$

由此可得

$$\mathrm{Cov}(w_i, \tilde{\rho}_j) = -\frac{a_{ji}\,p(1-p)}{(1-2p)^2}.$$

由于 $\mathbf{E}[w_j] = A_j := \mathbf{1}\{l_j = a\}$，我们得到

$$\mathrm{Cov}(w_i, y_j) = -A_j\,\frac{a_{ji}\,p(1-p)}{(1-2p)^2}\,\mathbf{1}\{i \in N(j)\}. \tag{19}$$

根据对称性,

$$\mathrm{Cov}(y_i, w_j) = -A_i\,\frac{a_{ij}\,p(1-p)}{(1-2p)^2}\,\mathbf{1}\{j \in N(i)\}. \tag{20}$$

*(c)*协方差$(y_i,\ y_j)$。为计算协方差$(y_i,\ y_j)$，我们对独立伯努利变量 $\{I_u\}$ 进行精确的多项式展开。令

$$w_i = \frac{I_i - p}{1-2p}, \qquad \tilde{\rho}_i = \frac{1}{1-2p}\left(\sum_{k\in N(i)} a_{ik}(1-I_k) - p\right) = \frac{1 - p - \sum_{k\in N(i)} a_{ik} I_k}{1-2p},$$

并定义常数

$$c_0 := \frac{1}{(1-2p)^2}, \qquad \beta := (1-p)c_0, \qquad \gamma_{ik} := p\, c_0\, a_{ik}, \qquad \delta_{ik} := -c_0\, a_{ik}.$$

由于 $a_{ii} = 0$，我们有 $i \notin N(i)$，展开 $y_i = w_i\tilde{\rho}_i$ 可得恒等式

$$y_i = -p(1-p)c_0 + \beta\, I_i + \sum_{k\in N(i)} (\gamma_{ik} + \delta_{ik} I_i)\, I_k. \tag{21}$$

由于 $\{I_u\}$ 相互独立，$\mathrm{Cov}(y_i,y_j)$ 只能来自至少共享一个伯努利因子的单项式。在无向图中，当 $\mathrm{dist}(i,j) \le 2$ 时，有两种可能。情况 A：$\mathrm{dist}(i,j) = 2$。假设 $i \notin N(j)$ 且 $j \notin N(i)$，但是

$$C_{ij} := N(i) \cap N(j) \neq \emptyset.$$

固定 $k \in C_{ij}$。y_i 中涉及 I_k 的项可以分组如下

$$(\gamma_{ik} + \delta_{ik} I_i) I_k,$$

类似地，y_j 中涉及 I_k 的项等于

$$(\gamma_{jk} + \delta_{jk} I_j) I_k.$$

y_i和y_j中的所有其余项都依赖于互不相交的伯努利变量集合，因此彼此独立。

由于 I_i, I_j, I_k彼此独立，且$(\gamma_{ik} + \delta_{ik} I_i)$、$(\gamma_{jk} + \delta_{jk} I_j)$与 I_k独立，我们得到

$$\mathrm{Cov}\big((\gamma_{ik} + \delta_{ik} I_i) I_k, (\gamma_{jk} + \delta_{jk} I_j) I_k\big) = \mathbf{E}[\gamma_{ik} + \delta_{ik} I_i]\ \mathbf{E}[\gamma_{jk} + \delta_{jk} I_j]\ \mathrm{Var}(I_k).$$

对所有共同邻居求和可得

$$\mathrm{Cov}(y_i, y_j) = \sum_{k\in C_{ij}} (\gamma_{ik} + \delta_{ik} q_i)\, (\gamma_{jk} + \delta_{jk} q_j)\ \mathrm{Var}(I_k),$$

其中 $q_i = \mathbf{E}[I_i]$ 且 $q_j = \mathbf{E}[I_j]$。

在随机响应下，

$$\mathrm{Var}(I_k) = p(1-p),$$

因此我们得到

$$\mathrm{Cov}(y_i, y_j) = p(1-p) \sum_{k\in N(i)\cap N(j)} (\gamma_{ik} + \delta_{ik} q_i)\, (\gamma_{jk} + \delta_{jk} q_j). \tag{22}$$

Case B：$\mathrm{dist}(i,\ j) = 1$。假设 $j \in N(i)$（等价地，$i \in N(j)$）。回顾精确展开式

$$y_i = -p(1-p)c_0 + \beta I_i + \sum_{k\in N(i)} \gamma_{ik} I_k + \sum_{k\in N(i)} \delta_{ik} I_i I_k, \qquad y_j = -p(1-p)c_0 + \beta I_j + \sum_{\ell\in N(j)} \gamma_{j\ell} I_\ell + \sum_{\ell\in N(j)} \delta_{j\ell} I_j I_\ell.$$

由于 I 在各节点之间相互独立，$\mathrm{Cov}(y_i,\ y_j)$ 完全由涉及共享伯努利变量的项决定。当 $j \in N(i)$ 时，共享变量为：

$$I_i, \quad I_j, \quad \{I_k : k \in N(i) \cap N(j)\}, \quad \text{and the product } I_i I_j.$$

所有其他项都依赖于互不相交的 $\{I_u\}$ 集合，因此协方差为零。

令 $q_u := \mathbf{E}[I_u]$，并注意，在随机响应下，对于所有 u，$\mathrm{Var}(I_u) = q_u(1-q_u) = p(1-p)$。现在我们来计算产生贡献的协方差。

(1) 线性–线性贡献。只有相同的指标才有贡献：

$$\mathrm{Cov}(\beta I_i, \gamma_{ji} I_i) = \beta\gamma_{ji}\,\mathrm{Var}(I_i), \qquad \mathrm{Cov}(\gamma_{ij} I_j, \beta I_j) = \gamma_{ij}\beta\,\mathrm{Var}(I_j),$$

对于每个共同邻居 $k \in N(i) \cap N(j)$，

$$\mathrm{Cov}(\gamma_{ik} I_k, \gamma_{jk} I_k) = \gamma_{ik}\gamma_{jk}\,\mathrm{Var}(I_k).$$

(2) 线性–二次贡献。与距离为 2 的情形不同，此处 I_i 和 I_j 都出现在两个展开式中，因此会与共享二次单项式 I_iI_j 产生非零协方差：

$$\mathrm{Cov}(I_i, I_iI_j) = \mathbf{E}[I_j]\,\mathrm{Var}(I_i) = q_j\ \mathrm{Var}(I_i), \qquad \mathrm{Cov}(I_j, I_iI_j) = \mathbf{E}[I_i]\,\mathrm{Var}(I_j) = q_i\ \mathrm{Var}(I_j),$$

同理，对于共同邻居 $k \in N(i) \cap N(j)$，

$$\mathrm{Cov}(I_k, I_iI_k) = \mathbf{E}[I_i]\,\mathrm{Var}(I_k) = q_i\ \mathrm{Var}(I_k), \qquad \mathrm{Cov}(I_k, I_jI_k) = \mathbf{E}[I_j]\,\mathrm{Var}(I_k) = q_j\ \mathrm{Var}(I_k).$$

因此，非零的线性–二次贡献为

$$\mathrm{Cov}(\beta I_i, \delta_{ji} I_iI_j) = \beta\delta_{ji}\,q_j\ \mathrm{Var}(I_i), \qquad \mathrm{Cov}(\delta_{ij} I_iI_j, \beta I_j) = \delta_{ij}\beta\,q_i\ \mathrm{Var}(I_j),$$

并且对于每个 $k \in N(i) \cap N(j)$，

$$\mathrm{Cov}(\gamma_{ik} I_k, \delta_{jk} I_jI_k) = \gamma_{ik}\delta_{jk}\,q_j\ \mathrm{Var}(I_k), \qquad \mathrm{Cov}(\delta_{ik} I_iI_k, \gamma_{jk} I_k) = \delta_{ik}\gamma_{jk}\,q_i\ \mathrm{Var}(I_k).$$

(3) 二次–二次贡献。共享二次单项式 I_iI_j 贡献

$$\mathrm{Cov}(\delta_{ij} I_iI_j, \delta_{ji} I_iI_j) = \delta_{ij}\delta_{ji}\,\mathrm{Var}(I_iI_j), \qquad \mathrm{Var}(I_iI_j) = q_iq_j\big(1 - q_iq_j\big).$$

此外，对于每个共同邻居 $k \in N(i) \cap N(j)$，二次单项式 I_iI_k 和 I_jI_k 共享因子 I_k，因此其协方差非零：

$$\mathrm{Cov}(I_iI_k,\ I_jI_k) = \mathbf{E}[I_i]\,\mathbf{E}[I_j]\,\mathrm{Var}(I_k) = q_iq_j\ \mathrm{Var}(I_k),$$

so

$$\mathrm{Cov}(\delta_{ik} I_iI_k, \delta_{jk} I_jI_k) = \delta_{ik}\delta_{jk}\,q_iq_j\ \mathrm{Var}(I_k).$$

综合 (1)–(3)，并利用对所有 u 均有 $\mathrm{Var}(I_u) = p(1-p)$，可得

$$\begin{aligned}\mathrm{Cov}(y_i, y_j) = {} & \beta\gamma_{ji}\,\mathrm{Var}(I_i) + \gamma_{ij}\beta\,\mathrm{Var}(I_j) + \sum_{k\in N(i)\cap N(j)} \gamma_{ik}\gamma_{jk}\,\mathrm{Var}(I_k) \\ & + \beta\delta_{ji}\,q_j\ \mathrm{Var}(I_i) + \delta_{ij}\beta\,q_i\ \mathrm{Var}(I_j) + \sum_{k\in N(i)\cap N(j)} \Big(\gamma_{ik}\delta_{jk}\,q_j + \delta_{ik}\gamma_{jk}\,q_i\Big)\,\mathrm{Var}(I_k) \\ & + \delta_{ij}\delta_{ji}\,\mathrm{Var}(I_iI_j) + \sum_{k\in N(i)\cap N(j)} \delta_{ik}\delta_{jk}\,q_iq_j\ \mathrm{Var}(I_k), \end{aligned} \tag{23}$$

对于 $j \in N(i)$。

特别地，对于每个固定的局部构型，$\mathrm{Cov}(Z_i,\ Z_j)$ 是局部参数的显式函数，因此是有限的局部常数。由于最大度一致有界，对于所有满足 $\mathrm{dist}(i,\ j) \le 2$ 的点对 $(i,\ j)$，这些局部常数一致有界。

定义滤过

$$\mathcal{F}_k := \sigma(I_1, \ldots, I_k), \qquad k = 0, 1, \ldots, n,$$

其中 $\mathcal{F}_0$ 是平凡的。定义取值于 $\mathbb{R}^2$的 Doob鞅

$$M_{n,k} := \mathbf{E}[T_n \mid \mathcal{F}_k], \qquad k = 0, 1, \ldots, n,$$

以及鞅差

$$D_{n,k} := M_{n,k} - M_{n,k-1}, \qquad k = 1, \dots, n.$$

那么，$\{(M_{n,\,k},\ \mathrm{F}_k)\}$ 是一个二元鞅，$M_{n,\,0} = \mathbf{E}[T_n]$, $M_{n,\,n} = T_n$，并且

$$T_n - \mathbf{E}[T_n] = \sum_{k=1}^{n} D_{n,k}, \qquad \mathbf{E}[D_{n,k} \mid \mathcal{F}_{k-1}] = 0. \tag{24}$$

我们证明，在 $\max_i d_i \le \Delta$ 给定的情况下，鞅差$D_{n,\,k}$一致有界。改变 I_k 的值，只会影响依赖于 I_k 的随机变量。由于 w_i 仅依赖于 I_i，因此只有w_k 会直接受到影响。此外， $\tilde{\rho}_i$ 依赖于 $\{I_j :\ j \in N(i)\}$，因此只有当 $k \in N(i)$，即 $i \in N(k)$ 时， I_k 才会影响 $\tilde{\rho}_i$。

因此，改变 I_k 只会影响以下索引处的 $y_i = w_i\tilde{\rho}_i$

$$i \in \{k\} \cup N(k),$$

根据有界度假设，其基数至多为 $1 + \Delta$。

接下来注意到

$$|w_k| = \left|\frac{I_k - p}{1 - 2p}\right| \le \frac{1}{1 - 2p}, \qquad |\tilde{\rho}_i| = \left|\frac{\hat{\rho}_i - p}{1 - 2p}\right| \le \frac{1}{1 - 2p},$$

因为 $\hat{\rho}_i \in [0, 1]$。

因此，$(S_{0,\,n},\ S_{1,\,n})$ 中每个受影响的加项变化至多为一个仅依赖于p的常数，且至多有 $1 + \Delta$ 个加项会发生变化。因此，存在一个常数 $C = C(p,\ \Delta) < \infty$，使得

$$\|T_n(I_k = 1) - T_n(I_k = 0)\|_\infty \le C.$$

由于

$$D_{n,k} = \mathbf{E}[T_n \mid \mathcal{F}_k] - \mathbf{E}[T_n \mid \mathcal{F}_{k-1}]$$

是通过揭示 I_k 得到的条件期望差，因此

$$\|D_{n,k}\|_\infty \le C \qquad \text{a.s.}$$

对所有 n 及所有 k均成立。

定义可预测二次变差矩阵

$$V_n := \frac{1}{n} \sum_{k=1}^{n} \mathbf{E}\big[D_{n,k} D_{n,k}^{\top} \mid \mathcal{F}_{k-1}\big].$$

由于鞅差$D_{n,\,k}$一致有界，对于任意 $\varepsilon > 0$,

$$\frac{1}{n} \sum_{k=1}^{n} \mathbf{E}\Big[\|D_{n,k}\|_2^2 \mathbf{1}\{\|D_{n,k}\|_2 > \varepsilon\sqrt{n}\} \,\Big|\, \mathcal{F}_{k-1}\Big] = 0 \quad \text{a.s. for all sufficiently large } n,$$

因此Lindeberg条件成立。

回顾

$$\Sigma_n := \frac{1}{n} \operatorname{Cov}(T_n) = \frac{1}{n} \operatorname{Var}(T_n - \mathbf{E}[T_n]).$$

对于杜布分解（24），鞅差的正交性意味着 s

$$\operatorname{Var}(T_n - \mathbf{E}[T_n]) = \sum_{k=1}^{n} \mathbf{E}\big[D_{n,k} D_{n,k}^{\top}\big] = \mathbf{E}\left[\sum_{k=1}^{n} \mathbf{E}\big[D_{n,k} D_{n,k}^{\top} \mid \mathcal{F}_{k-1}\big]\right]. \tag{25}$$

因此

$$\mathbf{E}[V_n] = \Sigma_n \qquad \text{for every } n.$$

为验证鞅中心极限定理所要求的条件方差收敛，ite

$$\Psi_{n,k} := \mathbf{E}\big[D_{n,k}D_{n,k}^{\top} \mid \mathcal{F}_{k-1}\big], \qquad V_n = \frac{1}{n}\sum_{k=1}^{n}\Psi_{n,k}.$$

固定任意矩阵元素（$a,\ b$）$\in\{1、2\}^2$,，并将相应的元素记为（V_n）$_{ab}$和（$\Psi_{n,\ k}$）$_{ab}$。由于$\|D_{n,\ k}\|_\infty \le C$，$\Psi_{n,\ k}$的每个元素都由仅依赖于 C 的常数一致有界。

此外，根据上文所使用的相同局部性论证，改变一个隐私化比特 I_r 只会影响那些 $D_{n,\ k}$，因而也只会影响那些 $\Psi_{n,\ k}$，其中 k位于 r 的有界半径邻域内。由于最大度一致有界，这类指标的数量 k 在 r和 n 上一致有界。因此，存在一个常数 $C' < \infty$，使得对于每个 r,

$$\big|(V_n)_{ab}(I) - (V_n)_{ab}(I^{(r)})\big| \le \frac{C'}{n},$$

其中 $I_{(r)}$表示从 $I=(I_1,\ \ldots,\ I_n)$ 中仅改变第 r 个坐标所得到的向量。根据 McDiarmid 不等式，对于每个固定的（$a,\ b$）和每个 $t>0$,

$$\Pr\big(\big|(V_n)_{ab} - \mathbf{E}[(V_n)_{ab}]\big| > t\big) \le 2\exp(-cnt^2)$$

对于某个常数 $c>0$。因此

$$(V_n)_{ab} - \mathbf{E}[(V_n)_{ab}] \xrightarrow{p} 0.$$

由于 V_n 为 2×2，可得

$$V_n - \mathbf{E}[V_n] \xrightarrow{p} 0.$$

此外，若对于某个有限矩阵 Σ，$\Sigma_n \to \Sigma$，则

$$V_n \xrightarrow{p} \Sigma.$$

因此，根据鞅中心极限定理,

$$\frac{T_n - \mathbf{E}[T_n]}{\sqrt{n}} = \sqrt{n}\Big(\frac{T_n}{n} - \mathbf{E}\Big[\frac{T_n}{n}\Big]\Big) \Rightarrow \mathcal{N}(0,\Sigma).$$

令 $\bar{T}_n := T_n/n = (S_{0,\ n}/n,\ S_{1,\ n}/n)^{\top}$，并定义 $g(x,\ y) = y/x$。假设

$$\mathbf{E}[\bar{T}_n] \to (\mu_0,\mu_1), \qquad \mu_0 > 0.$$

则$\hat{}\ \theta_n = g(\bar{T}_n)$，并且根据多元Delta方法,

$$\sqrt{n}\big(\hat{\theta}_n - \theta_n\big) \Rightarrow \mathcal{N}(0,\sigma^2), \qquad \theta_n := \frac{\mathbf{E}[S_{1,n}]}{\mathbf{E}[S_{0,n}]} \longrightarrow \theta := \frac{\mu_1}{\mu_0},$$

with

$$\sigma^2 = \nabla g(\mu_0,\mu_1)^{\top}\Sigma\,\nabla g(\mu_0,\mu_1), \qquad \nabla g(\mu_0,\mu_1) = \Big(-\frac{\mu_1}{\mu_0^2},\ \frac{1}{\mu_0}\Big)^{\top}.$$

由于尺度参数为 b 的拉普拉斯随机变量的方差为 $2b^2$，因此我们有

$$\mathrm{Var}(Z_n \mid S_{0,n}) = 2\frac{c^2}{S_{0,n}^2} \qquad \text{on } \{S_{0,n}>0\}.$$

因此

$$\mathrm{Var}(Z_n) = \mathbf{E}[\mathrm{Var}(Z_n \mid S_{0,n})] + \mathrm{Var}(\mathbf{E}[Z_n \mid S_{0,n}]).$$

由于 $\mathbf{E}[Z_n \mid S_{0,n}] = 0$，因此可得

$$\mathrm{Var}(Z_n) = 2c^2\ \mathbf{E}\left[\frac{1}{S_{0,n}^2}\mathbf{1}\{S_{0,n}>0\}\right].$$

现在定义

$$G_n := \left\{ \left| \frac{S_{0,n}}{n} - \pi_{\mathcal{A}} \right| \le \frac{\pi_{\mathcal{A}}}{2} \right\}.$$

由于 $S_{0,\ n} = \sum_i w_i$，其中 w_i 相互独立且一致有界，霍夫丁不等式推出

$$\Pr(G_n^c) \le 2\exp(-c_0 n)$$

对于某个常数 $c_0 > 0$ 以及所有充分大的 n,

在事件 G_n发生时,

$$\frac{\pi_{\mathcal{A}}}{2} n \le S_{0,n} \le \frac{3\pi_{\mathcal{A}}}{2} n,$$

因此

$$\frac{4}{9\pi_{\mathcal{A}}^2 n^2} \le \frac{1}{S_{0,n}^2} \le \frac{4}{\pi_{\mathcal{A}}^2 n^2}.$$

由此,

$$\mathbf{E}\left[\frac{1}{S_{0,n}^2} \mathbf{1}(G_n) \right] \le \frac{4}{\pi_{\mathcal{A}}^2 n^2},$$

and

$$\mathbf{E}\left[\frac{1}{S_{0,n}^2} \mathbf{1}(G_n) \right] \ge \frac{4}{9\pi_{\mathcal{A}}^2 n^2} \Pr(G_n).$$

由于 $\Pr(G_n) \to 1$，因此得到

$$\mathbf{E}\left[\frac{1}{S_{0,n}^2} \mathbf{1}(G_n) \right] = \Theta(n^{-2}).$$

□

A.6 命题 2 的证明：私有回归估计量的一致性

引理 5（误差变量校正）。设

$$y_i = \alpha + \beta x_i + v_i, \qquad \mathbf{E}[v_i \mid x_i] = 0,$$

并令

$$\hat{y}_i = y_i + \eta_i, \qquad \hat{x}_i = x_i + u_i,$$

其中

$$\mathbf{E}[\eta_i \mid y_i] = 0, \qquad \mathbf{E}[u_i \mid x_i] = 0, \qquad \mathbf{E}[u_i^2 \mid x_i] = \sigma^2.$$

如果 β^* 表示以$\hat{}x$ 对$\hat{}y$ 进行回归所得的斜率系数，那么标准误差变量修正为

$$\tilde{\beta} = \beta^* \frac{\left(\frac{1}{n-1}\right) \sum_i (\hat{x}_i - \bar{\hat{x}})^2}{\left(\frac{1}{n-1}\right) \sum_i (\hat{x}_i - \bar{\hat{x}})^2 - \sigma^2},$$

其　　是对真实回归系数 β 的一致估计量。相应的截距估计量为

$$\tilde{\alpha} = \bar{y} - \tilde{\beta}\, \bar{x},$$

这对于 α是一致的。

˜证明。我们首先证明 β 的一致性。

令 β^* 表示将ˆy 对ˆx进行回归所得的系数，即

$$\beta^* = \frac{\sum_i (\hat{x}_i - \bar{\hat{x}})(\hat{y}_i - \bar{\hat{y}})}{\sum_i (\hat{x}_i - \bar{\hat{x}})^2}.$$

我们先来看分子。由于

$$y_i = \alpha + \beta x_i + \nu_i, \qquad \hat{y}_i = y_i + \eta_i, \qquad \hat{x}_i = x_i + u_i,$$

我们有

$$\hat{y}_i = \alpha + \beta x_i + \nu_i + \eta_i.$$

因此，

$$\sum_i (\hat{x}_i - \bar{\hat{x}})(\hat{y}_i - \bar{\hat{y}}) = \sum_i (\hat{x}_i - \bar{\hat{x}})(\alpha + \beta x_i + \nu_i + \eta_i - \bar{\hat{y}}).$$

¯由于 $\sum_i (\hat{x}_i - \hat{x}) = 0$，截距项消失，因此

$$\sum_i (\hat{x}_i - \bar{\hat{x}})(\hat{y}_i - \bar{\hat{y}}) = \beta \sum_i (\hat{x}_i - \bar{\hat{x}}) x_i + \sum_i (\hat{x}_i - \bar{\hat{x}}) \nu_i + \sum_i (\hat{x}_i - \bar{\hat{x}}) \eta_i.$$

现在注意到

$$\hat{x}_i - \bar{\hat{x}} = (x_i + u_i) - (\bar{x} + \bar{u}) = (x_i - \bar{x}) + (u_i - \bar{u}).$$

因此，

$$\begin{aligned} \sum_i (\hat{x}_i - \bar{\hat{x}}) x_i &= \sum_i (x_i - \bar{x}) x_i + \sum_i (u_i - \bar{u}) x_i \\ &= \sum_i (x_i - \bar{x})^2 + \sum_i (u_i - \bar{u}) x_i. \end{aligned}$$

根据大数定律和假设 $\mathbf{E}[u_i \mid x_i] = 0$，

$$\frac{1}{n} \sum_i (u_i - \bar{u}) x_i \xrightarrow{p} 0.$$

类似地，由于 $\mathbf{E}[\nu_i \mid x_i] = 0$，我们有

$$\frac{1}{n} \sum_i (\hat{x}_i - \bar{\hat{x}}) \nu_i \xrightarrow{p} 0.$$

此外，由于 nce $\mathbf{E}[\eta_i \mid y_i] = 0$ 且 ˆx_i 在模型假设下具有有界的二阶矩，

$$\frac{1}{n} \sum_i (\hat{x}_i - \bar{\hat{x}}) \eta_i \xrightarrow{p} 0.$$

因此，

$$\frac{1}{n} \sum_i (\hat{x}_i - \bar{\hat{x}})(\hat{y}_i - \bar{\hat{y}}) \xrightarrow{p} \beta \operatorname{Var}(x).$$

现在考虑分母：

$$\begin{aligned} \frac{1}{n} \sum_i (\hat{x}_i - \bar{\hat{x}})^2 &= \frac{1}{n} \sum_i \big[(x_i - \bar{x}) + (u_i - \bar{u}) \big]^2 \\ &= \frac{1}{n} \sum_i (x_i - \bar{x})^2 + \frac{2}{n} \sum_i (x_i - \bar{x})(u_i - \bar{u}) + \frac{1}{n} \sum_i (u_i - \bar{u})^2. \end{aligned}$$

根据大数定律，

$$\frac{1}{n}\sum_i (x_i - \bar{x})^2 \xrightarrow{p} \mathrm{Var}(x),$$
$$\frac{1}{n}\sum_i (x_i - \bar{x})(u_i - \bar{u}) \xrightarrow{p} 0,$$
$$\frac{1}{n}\sum_i (u_i - \bar{u})^2 \xrightarrow{p} \mathbf{E}[u_i^2] = \sigma^2.$$

因此，

$$\frac{1}{n}\sum_i (\hat{x}_i - \bar{\hat{x}})^2 \xrightarrow{p} \mathrm{Var}(x) + \sigma^2.$$

结合分子与分母，并应用连续映射定理，我们得到

$$\beta^* \xrightarrow{p} \frac{\beta\ \mathrm{Var}(x)}{\mathrm{Var}(x) + \sigma^2}.$$

此外，

$$\frac{1}{n-1}\sum_i (\hat{x}_i - \bar{\hat{x}})^2 \xrightarrow{p} \mathrm{Var}(x) + \sigma^2.$$

因此，再次应用连续映射定理，

$$\tilde{\beta} = \beta^* \frac{\left(\frac{1}{n-1}\right)\sum_i (\hat{x}_i - \bar{\hat{x}})^2}{\left(\frac{1}{n-1}\right)\sum_i (\hat{x}_i - \bar{\hat{x}})^2 - \sigma^2}$$
$$\xrightarrow{p} \frac{\beta\ \mathrm{Var}(x)}{\mathrm{Var}(x) + \sigma^2} \cdot \frac{\mathrm{Var}(x) + \sigma^2}{\mathrm{Var}(x)} = \beta.$$

˜这证明了 β的一致性。

接下来，我们证明 $\tilde{\alpha}$ 的一致性。由于

$$\hat{y}_i = y_i + \eta_i = \alpha + \beta x_i + \nu_i + \eta_i,$$

我们有

$$\bar{\hat{y}} = \alpha + \beta\bar{x} + \bar{\nu} + \bar{\eta}.$$

此外，

$$\bar{\hat{x}} = \bar{x} + \bar{u}.$$

因此，

$$\begin{aligned}\tilde{\alpha} &= \bar{\hat{y}} - \tilde{\beta}\bar{\hat{x}} \\ &= \alpha + \beta\bar{x} + \bar{\nu} + \bar{\eta} - \tilde{\beta}(\bar{x} + \bar{u}) \\ &= \alpha + (\beta - \tilde{\beta})\bar{x} - \tilde{\beta}\bar{u} + \bar{\nu} + \bar{\eta}.\end{aligned}$$

根据大数定律，

$$\bar{u} \xrightarrow{p} 0, \qquad \bar{\nu} \xrightarrow{p} 0, \qquad \bar{\eta} \xrightarrow{p} 0.$$

由于 $\tilde{\beta} \xrightarrow{p} \beta$，因此

$$\tilde{\alpha} \xrightarrow{p} \alpha.$$

因此， $\tilde{\alpha}$ 对 α是一致的。 □

基于上述引理，我们给出命题 2 的形式化证明。

证明。我们证明算法 2 得出的斜率估计是一致的。首先，令

$$\beta^* := \frac{\mathrm{ncov}(x, y)}{\mathrm{nvar}(x)}$$

表示根据有界私有化样本计算出的回归系数

$$\{(\hat{x}_i, \hat{y}_i)\}_{i=1}^n$$

由算法 2 生成。根据算法 3，发布的斜率估计量为

$$\hat{\beta} = \frac{\mathrm{ncov}(x, y) + L_1}{\mathrm{nvar}(x) + L_2},$$

当 $\mathrm{nvar}(x) + L_2 > 0$ 时，其中

$$L_1 \sim \mathrm{Lap}\left(0, \frac{3\Delta_1}{\varepsilon}\right), \qquad L_2 \sim \mathrm{Lap}\left(0, \frac{3\Delta_2}{\varepsilon}\right).$$

由于输入对 $(\hat{x}_i, \hat{y}_i)$ 有界，敏感度界 Δ_1，Δ_2 关于 n 一致有界，因此

$$L_1 = O_p(1), \qquad L_2 = O_p(1).$$

Now,

$$\hat{\beta} - \beta^* = \frac{\mathrm{ncov}(x, y) + L_1}{\mathrm{nvar}(x) + L_2} - \frac{\mathrm{ncov}(x, y)}{\mathrm{nvar}(x)} = \frac{L_1 \,\mathrm{nvar}(x) - \mathrm{ncov}(x, y) L_2}{\mathrm{nvar}(x)\big(\mathrm{nvar}(x) + L_2\big)}.$$

由于 $(\hat{x}_i, \hat{y}_i)$ 有界，根据大数定律可得

$$\frac{1}{n}\,\mathrm{nvar}(x) \xrightarrow{p} \mathrm{Var}(x), \qquad \frac{1}{n}\,\mathrm{ncov}(x, y) \xrightarrow{p} \mathrm{Cov}(x, y).$$

假设 $\mathrm{Var}(x) > 0$。则

$$\mathrm{nvar}(x) = \Theta_p(n), \qquad \mathrm{ncov}(x, y) = O_p(n).$$

因此，

$$L_1 \,\mathrm{nvar}(x) = O_p(n), \qquad \mathrm{ncov}(x, y) L_2 = O_p(n),$$

而

$$\mathrm{nvar}(x)\big(\mathrm{nvar}(x) + L_2\big) = \Theta_p(n^2).$$

由此可得

$$\hat{\beta} - \gamma_n = o_p(1).$$

此外，由于 $\mathrm{nvar}(x) = \Theta_p(n)$ 且 $L_2 = O_p(1)$，

$$\Pr\big(\mathrm{nvar}(x) + L_2 > 0\big) \to 1,$$

因此，算法 3 返回 $\perp$ 的事件渐近上可忽略不计。因此，DPSuffStats 生成的斜率估计量对于 β^* 是一致的；β^* 是基于经截断拉普拉斯扰动的样本计算得到的回归系数。

根据引理 5，经过去偏步骤后，基于经截断拉普拉斯扰动的样本计算得到的回归系数可为真实回归系数 β 提供一致估计量。由于步骤 1 表明，DPSuffStats 估计值 β 与 β^* 相差一个 $o_p(1)$ 项，对 β 应用相同的去偏映射并利用连续映射定理，可知所得去偏估计量对于 β 也具有一致性。

因此，算法 2、算法 3 与去偏步骤的组合，能够得到真实回归系数的一致估计量。□

A.7 定理 5 的证明：回归估计的边相邻差分隐私保证

我们首先给出一个辅助引理，用以确定回归系数中方差项和协方差项（$\mathrm{ncov}(x,\ y)$ 和 $\mathrm{nvar}(x)$）的全局敏感度。

引理 6（一般区间上方差和协方差的全局敏感度）。设$n \geq 2$。假设 $x,\ \hat{}x \in [a,b]^n$和 $y,\ \hat{}y \in [a',b']^n$是最多在两个索引处不同的相邻数据库（即存在 $k \in \{1,\ldots,n\}$，使得对于所有 $i \neq k$，$(x_i,\ y_i) = (\hat{}x_i,\ \hat{}y_i)$），其中 $a < b$且$a' < b'$。定义

$$\mathrm{nvar}(x) = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)^2, \qquad \mathrm{ncov}(x,y) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)\Big(\sum_{i=1}^{n} y_i\Big).$$

Then

$$\big|\mathrm{nvar}(x) - \mathrm{nvar}(\hat{x})\big| \leq \Big(1 - \frac{1}{n}\Big)(b-a)^2,$$

and

$$\big|\mathrm{ncov}(x,y) - \mathrm{ncov}(\hat{x},\hat{y})\big| \leq 2\Big(1 - \frac{1}{n}\Big)(b-a)(b'-a').$$

因此，

$$\mathrm{GS}_{\mathrm{nvar}} = \Big(1 - \frac{1}{n}\Big)(b-a)^2, \qquad \mathrm{GS}_{\mathrm{ncov}} = 2\Big(1 - \frac{1}{n}\Big)(b-a)(b'-a').$$

证明。对于 $x \in [a,b]^n$ 和 $y \in [a',b']^n$，定义仿射缩放

$$\tilde{x} = \frac{x - a\mathbf{1}}{b-a} \in [0,1]^n, \qquad \tilde{y} = \frac{y - a'\mathbf{1}}{b'-a'} \in [0,1]^n,$$

$\tilde{}\tilde{}$并类似地根据$\hat{}x, \hat{}y$定义$\hat{}x, \hat{}y$。相邻关系在这些逐坐标仿射映射下保持不变。

令 $\bar{x}$（分别为 $\tilde{x}$）表示 x（分别为 $\tilde{x}$）的均值，y 也同理。由于

$$x - \bar{x}\,\mathbf{1} = (b-a)\big(\tilde{x} - \bar{\tilde{x}}\,\mathbf{1}\big), \qquad y - \bar{y}\,\mathbf{1} = (b'-a')\big(\tilde{y} - \bar{\tilde{y}}\,\mathbf{1}\big),$$

我们得到缩放恒等式

$$\mathrm{nvar}(x) = (b-a)^2\,\mathrm{nvar}(\tilde{x}), \qquad \mathrm{ncov}(x,y) = (b-a)(b'-a')\,\mathrm{ncov}(\tilde{x},\tilde{y}).$$

因此，

$$\big|\mathrm{nvar}(x) - \mathrm{nvar}(\hat{x})\big| = (b-a)^2\big|\mathrm{nvar}(\tilde{x}) - \mathrm{nvar}(\widetilde{\hat{x}})\big|,$$

and

$$\big|\mathrm{ncov}(x,y) - \mathrm{ncov}(\hat{x},\hat{y})\big| = 2(b-a)(b'-a')\big|\mathrm{ncov}(\tilde{x},\tilde{y}) - \mathrm{ncov}(\widetilde{\hat{x}},\widetilde{\hat{y}})\big|.$$

由于 x_i 的单次变化会导致 y 和 $\hat{}y$ 的两个坐标发生变化，因此得到的相邻数据集在两个索引上不同。

根据 Alabi 等人（2022年，引理 19），将其应用于 $[0,1]^n$ 中的向量，可得

$$\big|\mathrm{nvar}(\tilde{x}) - \mathrm{nvar}(\widetilde{\hat{x}})\big| \leq 1 - \frac{1}{n}, \qquad \big|\mathrm{ncov}(\tilde{x},\tilde{y}) - \mathrm{ncov}(\widetilde{\hat{x}},\widetilde{\hat{y}})\big| \leq 1 - \frac{1}{n}.$$

乘以相应的缩放因子即可得到所述界。 □

基于上述引理，我们给出定理 5 的正式证明。

证明。设 (V，E，L) 和 (V，E′，L′) 是两个边相邻带标签网络。我们验证算法 2 满足 $(\varepsilon_\ell + \varepsilon_e,\ \delta_\ell)$-边相邻差分隐私。

算法 2 的第一阶段使用截断拉普拉斯噪声扰动节点属性。根据 Geng 等人（2018），截断拉普拉斯机制可针对节点属性实现 $(\varepsilon_\ell, \delta_\ell)$-差分隐私。因此，发布的扰动标签 L 满足 $(\varepsilon_\ell, \delta_\ell)$-差分隐私。

在给定扰动标签 L的条件下，第二阶段执行边依赖估计。根据 Alabi等人（2022，引理4），结合引理 6 中确立的全局敏感度界，对于任意固定的输入标签，该估计量满足 $(\varepsilon_e,0)$ 边相邻差分隐私。

因此，整体机制是一个 $(\varepsilon_\ell, \delta_\ell)$-差分隐私机制（针对节点属性）与一个 $(\varepsilon_e,0)$ 边相邻差分隐私机制（针对边）的顺序组合。应用定理1可得，组合机制针对 (V, E,L) 满足 $(\varepsilon_\ell + \varepsilon_e, \delta_\ell)$ 边相邻差分隐私。□

A.8 定理6的证明：私有去偏回归估计量的渐近正态性

证明。定义

$$Z_{i,n} = \begin{pmatrix} \hat{x}_i \\ \hat{y}_i \\ \hat{x}_i^2 \\ \hat{x}_i\hat{y}_i \end{pmatrix}, \qquad M_n = \sum_{i=1}^{n} Z_{i,n}, \qquad \bar{M}_n = \frac{1}{n} M_n.$$

Thus

$$\bar{M}_n = (\bar{M}_{1,n}, \bar{M}_{2,n}, \bar{M}_{3,n}, \bar{M}_{4,n})^\top$$

它汇集了表示回归斜率及其去偏版本所需的四个经验矩。事实上，

$$\beta_n^* = \frac{\bar{M}_{4,n} - \bar{M}_{1,n}\bar{M}_{2,n}}{\bar{M}_{3,n} - \bar{M}_{1,n}^2}, \qquad \tilde{\beta}_n = \frac{\bar{M}_{4,n} - \bar{M}_{1,n}\bar{M}_{2,n}}{\bar{M}_{3,n} - \bar{M}_{1,n}^2 - \sigma_z^2} = h(\bar{M}_n).$$

令 I_1、……、I_n 表示私有化逐节点输入。定义滤过

$$\mathcal{F}_{n,k} := \sigma(I_1, \ldots, I_k), \qquad k = 0, 1, \ldots, n,$$

其中 $\mathbf{F}_{n,\ 0}$ 是平凡的。定义取值于 $\mathbb{R}^4$ 的 Doob鞅

$$U_{n,k} := \mathbf{E}[M_n \mid \mathcal{F}_{n,k}], \qquad k = 0, 1, \ldots, n,$$

以及鞅差

$$D_{n,k} := U_{n,k} - U_{n,k-1}, \qquad k = 1, \ldots, n.$$

Then

$$U_{n,0} = \mathbf{E}[M_n], \qquad U_{n,n} = M_n,$$

因此

$$M_n - \mathbf{E}[M_n] = \sum_{k=1}^{n} D_{n,k}, \qquad \mathbf{E}[D_{n,k} \mid \mathcal{F}_{n,k-1}] = 0.$$

根据构造，$\hat{x}_i$ 仅取决于节点 i处的隐私化输入，而$\hat{y}_i$ 仅取决于 i的一跳邻域中的隐私化输入。因此，每个 $Z_{i,n}$ 都是节点 i的有界半径邻域内隐私化输入的可测函数。由于最大度是一致有界的，改变单个输入 I_k 最多只会影响一致有界数量的求和项$Z_{i,n}$。此外，由于采用了截断拉普拉斯噪声，$\hat{x}_i$ 和$\hat{y}_i$ 是一致有界的，因此 $Z_{i,n}$ 的每个坐标也都是一致有界的。因此，存在一个常数 $C < \infty$ 使得

$$\|D_{n,k}\|_\infty \le C \qquad \text{a.s. for all } n, k.$$

然后定义

$$\Psi_{n,k} := \mathbf{E}[D_{n,k} D_{n,k}^\top \mid \mathcal{F}_{n,k-1}], \qquad V_n := \frac{1}{n}\sum_{k=1}^{n} \Psi_{n,k}.$$

我们声称

$$V_n - \mathbf{E}[V_n] \xrightarrow{p} 0.$$

固定任意一个矩阵元素（a, b）。由于$\|D_{n,\,k}\|_\infty \le C$，$\Psi_{n,\,k}$的每个元素都一致有界。根据上文相同的局部性论证，更改一个隐私化输入 I_r 只会影响和式中定义（V_n）$_{ab}$的数量一致有界的项。因此，存在 $C' > 0$，使得更改一个坐标 I_r 至多使（V_n）$_{ab}$ 变化 C'/n。应用 McDiarmid 不等式可得，对每个 $t > 0$,

$$\Pr\Big(\big|(V_n)_{ab} - \mathbf{E}[(V_n)_{ab}]\big| > t\Big) \le 2\exp(-cnt^2)$$

其中某个常数为 $c > 0$。因此

$$(V_n)_{ab} - \mathbf{E}[(V_n)_{ab}] \xrightarrow{p} 0.$$

由于维数固定，这意味着

$$V_n - \mathbf{E}[V_n] \xrightarrow{p} 0.$$

接下来，根据鞅差的正交性,

$$\mathbf{E}[V_n] = \frac{1}{n}\sum_{k=1}^{n} \mathbf{E}[D_{n,k} D_{n,k}^\top] = \frac{1}{n}\operatorname{Var}(M_n).$$

根据由有界局部依赖得出的方差极限结果,

$$\frac{1}{n}\operatorname{Cov}(M_n) \to \Sigma.$$

因此

$$V_n \xrightarrow{p} \Sigma.$$

考虑归一化鞅差数组

$$X_{n,k} := \frac{1}{\sqrt{n}} D_{n,k}.$$

由于$D_{n,\,k}$一致有界,

$$\|X_{n,k}\|_\infty \le \frac{C}{\sqrt{n}} \to 0$$

关于 k一致地。因此，Lindeberg条件自动成立。条件协方差过程满足

$$\sum_{k=1}^{n} \mathbf{E}[X_{n,k} X_{n,k}^\top \mid \mathcal{F}_{n,k-1}] = \frac{1}{n}\sum_{k=1}^{n} \Psi_{n,k} = V_n \xrightarrow{p} \Sigma.$$

因此，根据多元鞅中心极限定理,

$$\frac{1}{\sqrt{n}}\big(M_n - \mathbf{E}[M_n]\big) \Rightarrow N(0, \Sigma).$$

¯等价地，如果 μ_n :=$\mathbf{E}[M_n]$，那么

$$\sqrt{n}(\bar{M}_n - \mu_n) \Rightarrow N(0, \Sigma).$$

根据 U_n 的渐近正态性和 Delta方法，我们证明 β^* 渐近服从正态分布。定义

$$g(m_1, m_2, m_3, m_4) := \frac{m_4 - m_1 m_2}{m_3 - m_1^2}.$$

于是 $\beta_n^* = g(\bar{M}_n)$。由于 $\mu_n \to \mu$ 和 $\mu_3 - \mu_1^2 > 0$，映射 g 在 μ的某个邻域内连续可微。根据多元Delta方法，

$$\sqrt{n}\big(\beta_n^* - g(\mu_n)\big) \Rightarrow N\big(0, \nabla g(\mu)^\top \Sigma \nabla g(\mu)\big).$$

˜接下来，我们建立无偏估计量 β_n的渐近正态性。定义

$$h(m_1, m_2, m_3, m_4) := \frac{m_4 - m_1 m_2}{m_3 - m_1^2 - \sigma_z^2}.$$

于是 $\tilde{\beta}_n = h(\bar{M}_n)$。假设 $\mu_3 - \mu_1^2 - \sigma_z^2 > 0$，映射 h在 μ的某个邻域内连续可微。因此，多元Delta方法给出

$$\sqrt{n}\big(\tilde{\beta}_n - h(\mu_n)\big) \Rightarrow N\big(0, \nabla h(\mu)^\top \Sigma \nabla h(\mu)\big).$$

根据总体层面去偏校正的定义，

$$h(\mu) = \beta.$$

由于 $\mu_n \to \mu$ 和 h 连续，

$$h(\mu_n) \to h(\mu) = \beta.$$

因此

$$\sqrt{n}(\tilde{\beta}_n - \beta) \Rightarrow N(0, \tau^2), \qquad \tau^2 := \nabla h(\mu)^\top \Sigma \nabla h(\mu).$$

接下来，我们证明 DPSuffStats 扰动可以忽略。令

$$\mathrm{ncov}_n := \mathrm{ncov}(\hat{x}, \hat{y}), \qquad \mathrm{nvar}_n := \mathrm{nvar}(\hat{x}).$$

根据定义，

$$\beta_n^* = \frac{\mathrm{ncov}_n}{\mathrm{nvar}_n}.$$

ˆ最终发布的估计量 β_n 是通过向协方差项和方差项分别添加拉普拉斯噪声 L_1 和 L_2 得到的。根据拉普拉斯噪声的有界敏感度，

$$L_1 = O_p(1), \qquad L_2 = O_p(1).$$

另一方面，根据大数定律和非退化条件，

$$\mathrm{nvar}_n = \Theta_p(n), \qquad \mathrm{ncov}_n = O_p(n).$$

因此

$$\hat{\beta}_n - \beta_n^* = \frac{L_1\ \mathrm{nvar}_n - \mathrm{ncov}_n\, L_2}{\mathrm{nvar}_n(\mathrm{nvar}_n + L_2)} = O_p(n^{-1}).$$

˜由于 β_n 是通过一个在极限点邻域内导数有界的光滑校正映射从 β_n^* 得到的，因此去偏后仍保持相同的阶。于是

$$\hat{\beta}_n - \tilde{\beta}_n = O_p(n^{-1}),$$

因此

$$\sqrt{n}\,(\hat{\beta}_n - \tilde{\beta}_n) \xrightarrow{p} 0.$$

最后，

$$\sqrt{n}(\hat{\beta}_n - \beta) = \sqrt{n}(\tilde{\beta}_n - \beta) + \sqrt{n}(\hat{\beta}_n - \tilde{\beta}_n).$$

因此，根据斯卢茨基定理，

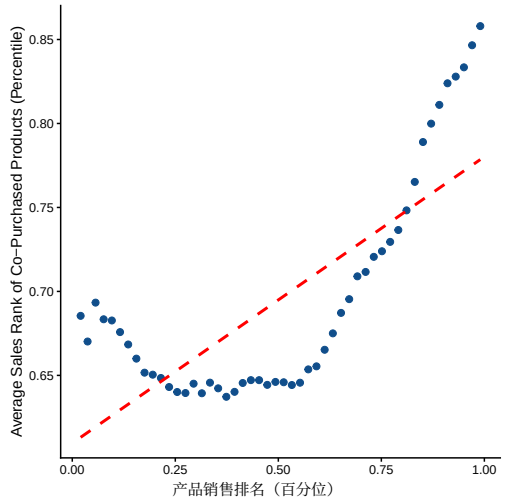
$$\sqrt{n}(\hat{\beta}_n - \beta) \Rightarrow N(0, \tau^2).$$

证明至此完成。 □

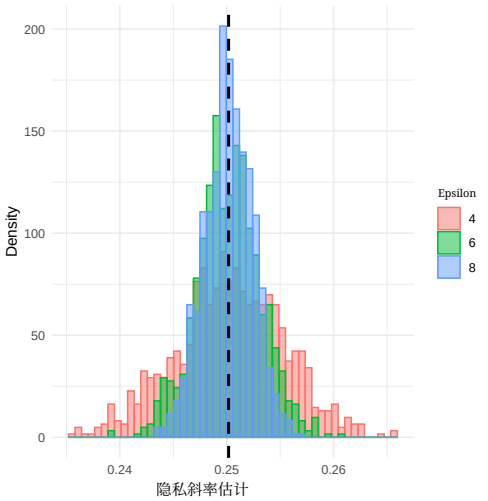
B 补充材料

B.1 亚马逊上经常共同购买的商品

本节展示了我们的方法在连续标签上的一项实证应用，所用数据由 Leskovec 等人（2007）以及 Yang 和 Leskovec（2012）收集，涉及亚马逊商品的特征以及这些商品是否经常被一起购买。不同商品之间互补性的模式是经济学中的一个经典问题，对商业定价、市场结构和反垄断监管都有影响（例如，参见 Gentzkow（2007））。具体而言，我们考虑一个标记图，其中每个节点的标签对应于该节点在销售分布中的百分位排名（缩放至 $[0, 1]$），如果亚马逊网站显示两种商品经常被一起购买，则这两种商品之间存在一条边。图 B.1a 显示，销量高的商品经常与其他高销量商品一起被购买。我们针对连续标签运行了 1,000 次方法模拟，并在图 B.1b 中绘制回归斜率的分布。



(a) 经常共同购买的产品的销售排名与其自身销售排名。



(b) 不同 ε取值下隐私回归斜率的分布。

图 B.1. **经常一起购买的亚马逊商品。** 面板 (a) 绘出了共同购买产品的平均销售排名（百分位数）与商品自身销售排名（百分位数）的关系。面板 (b) 显示了 1,000 次模拟中隐私斜率估计值的分布；红色虚线标示真实斜率。我们设定 $\varepsilon_e = \varepsilon_l = 4$。

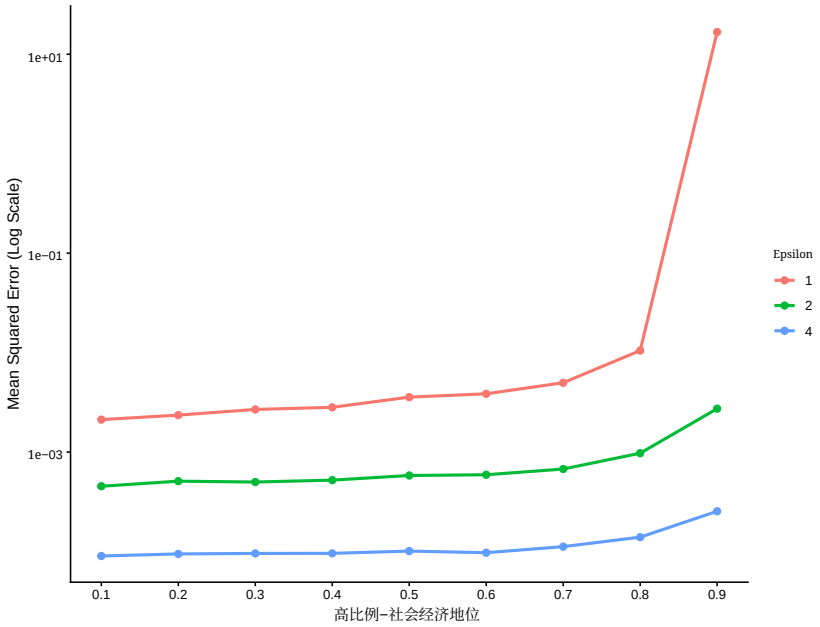


图 B.2。**模拟网络中均方误差随细胞组成的变化**。该图展示了网络中高社会经济地位个体的比例与隐私估计量所得均方误差之间的关系，涵盖三个隐私预算（$\varepsilon \in \{1、2、4\}$），隐私预算在 ε_e 和 ε_l 之间均等分配。每个数据点的结果均基于 1,125 次模拟（75 个图 $\times$15噪声随机种子）的平均值。网络生成方式为埃尔德什-雷尼图，连接概率为 0.04，包含 2,000 个节点。在 x轴上，我们改变“高社会经济地位”集合中节点的比例（对应于 Chetty 等人（2022a）中的应用）。 y轴采用对数刻度。

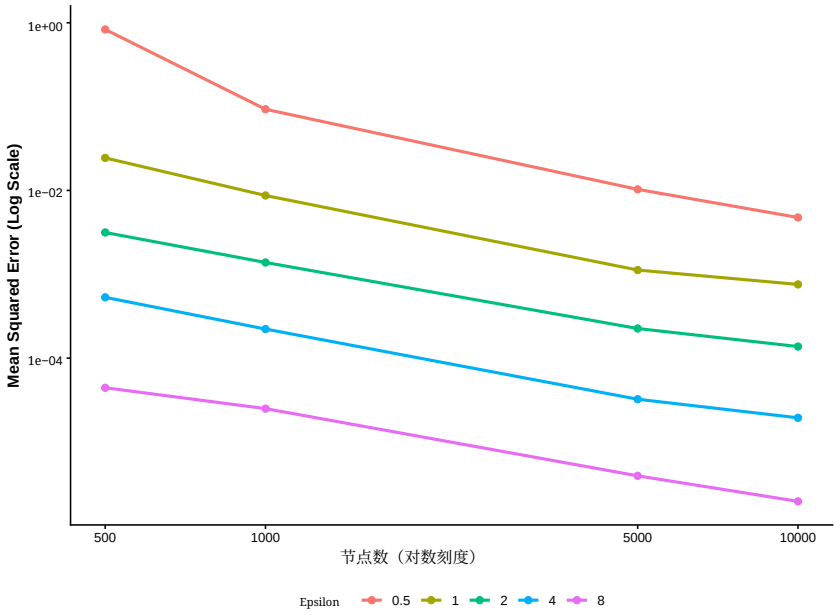


图 B.3. **模拟网络的均方误差与网络规模的关系**。该图展示了在不同隐私预算（ε）下，我们的差分隐私估计量的均方误差（MSE）随网络规模变化的情况。结果基于对模拟的 Erdős-Rényi 图进行的 10,000 次蒙特卡洛模拟；图中的节点随机分为两个规模相等的组，并随着图的增长调整连接概率，以保持平均度为 20。我们将隐私预算平均分配给 ε_e 和 ε_l。

表 B.1. Banerjee 等人（2013年）的村庄汇总统计量。

村庄编号	# HH	历史疾病数	历史非疾病数	平均度数	跨种姓连接	标准差（ε=8）
28	315	93	222	8.7	0.26	0.04
29	272	155	117	7.3	0.21	0.02
30	138	64	74	8.7	0.19	0.04
31	151	141	10	7.9	0.03	0.03
32	241	116	125	9.7	0.22	0.03
33	204	26	178	7.4	0.35	0.10
34	165	103	62	6.1	0.22	0.03
35	206	107	99	6.7	0.22	0.03
36	289	92	197	10.4	0.41	0.03
38	157	31	126	6.7	0.32	0.07
39	287	64	223	8.4	0.52	0.04
40	240	62	178	8.1	0.21	0.05
42	192	126	66	7.4	0.04	0.03
44	227	32	195	9.5	0.27	0.09
45	219	59	160	7.7	0.37	0.04
46	261	169	92	8.3	0.13	0.02
47	137	26	111	9.3	0.36	0.08
48	182	86	96	9.8	0.26	0.03
49	193	44	149	9.3	0.32	0.06
50	244	60	184	10.3	0.26	0.05
51	248	103	145	12.9	0.36	0.03
52	327	79	248	12.3	0.36	0.04
53	151	48	103	11.5	0.41	0.04
54	99	57	42	11.1	0.15	0.04
55	257	52	205	6.9	0.11	0.07
57	208	201	7	9.1	0.02	0.02
58	177	51	126	9.2	0.41	0.05
59	328	110	218	8.6	0.29	0.03
60	354	101	253	8.0	0.21	0.04
61	121	31	90	8.3	0.42	0.06
62	189	67	122	8.7	0.25	0.04
63	161	29	132	6.6	0.15	0.09
64	257	78	179	7.5	0.28	0.04
65	285	86	199	10.7	0.19	0.04
66	183	40	143	8.5	0.37	0.06
67	193	92	101	10.5	0.13	0.04
68	153	35	118	9.7	0.28	0.06
69	180	94	86	13.4	0.19	0.03
70	205	45	160	12.6	0.57	0.04
71	297	63	234	10.2	0.49	0.04
72	223	123	100	11.0	0.19	0.03
73	164	78	86	10.5	0.28	0.03
74	170	32	138	7.4	0.17	0.08
75	166	55	111	11.3	0.36	0.04
76	251	62	189	7.6	0.30	0.05
77	153	83	70	7.6	0.14	0.03

表 *B.1* 注释：本表汇总了 Banerjee 等人（2013）研究中符合以下条件的村庄子集的网络和人口统计特征：具有种姓覆盖，且我们划分的两个种姓集合中每个集合至少包含七个家庭。

B.2 符号列表

表B.2：符号列表

符号	描述
$\mathcal{V}$	网络中的顶点（节点）集合
$\mathcal{E}$	节点之间的边（友谊关系/连接）集合
$\mathbf{L} = (l_i)_{i\in\mathcal{V}}$	节点标签向量，每个节点 i 对应 $l_i \in \{a, b\}$
e_{ij}	二元指示变量；$e_{ij} = 1$ 如果 i 和 j 之间存在一条边，则为 1，否则为 0
d_i	节点 i 的度，$\#(j \in \mathrm{V} : e_{ij} = 1)$.
a_{ij}	e_{ij}/d_i。
$N(i)$	节点 i 的邻域，$\{ j \in \mathrm{V} : e_{ij} = 1\}$。
$\mathcal{A}, \mathcal{B}$	基于标签划分的 V（a, b）
π_A	集合 A 中的节点比例。
$\#(\cdot)$	基数算子（集合中的元素数量）
ρ_i	个人 i属于目标群体的朋友比例
$C_{\mathrm{A}\to\mathrm{B}}$	跨类型连通性指数（A组个体的平均值 ρ_i）
$C_{\mathrm{A}\to\mathrm{A}}$	同类型连通性指数
s	特定的“单元”或用户子集（例如，某个县或学校）
$C_{\mathrm{A}\to\mathrm{B}}^{s}$	专门针对单元 s计算的连通性指数
ε	隐私损失预算（差分隐私参数）
ε_ℓ	标签隐私损失预算
ε_e	边隐私损失预算
$\mathrm{M}_1, \mathrm{M}_2$	用于保护标签 M_1 和边 M_2的隐私保护机制
$D,\ D'$	边相邻带标签网络