预印本平台的普及是否缩小了学术论文间的“贫富差距”？

邱天异

北京大学

信息科学技术学院

tianyi.qiu@stu.pku.edu.cn

2023 年 1 月 17 日

摘要

20 世纪末以来，预印本平台在学术界中的使用日益增多，这些平台允许研究者不经期刊发表而直接公布自己的研究成果。这是否意味着，那些并非顶级的论文，也能得到更多的曝光度？本文构建了 19 个学科在 40 年中论文发表情况和预印本平台普及情况的数据集，并采用连续变量双重差分法 (Callaway et al., 2021) 分析预印本平台普及度对论文引用量不均衡程度的效应，填补了文献中关于“预印本平台对领域整体的效应”这一类问题的空白。分析中没有发现显著且稳健的效应，但不排除是模型和数据中的局限性掩盖了实际存在的效应。

1 引言

1.1 预印本平台概述

1991 年，科研工作者建立了预印本平台 arXiv(Ginsparg, 2021)。从那时以来，以 arXiv 为首的预印本平台，已在许多学科领域站稳脚跟，成为最主要的发表途径之一。预印本平台允许科研工作者通过网络公布自己已发表或未发表的论文，并允许其他研究者检索和阅读这些论文。在一些研究领域中，将自己的论文第一时间公布在预印本平台上，并在预印本平台上关注该领域内的最新进展，已经成为了研究者的日常。

自 1991 年以来的约 30 年间，预印本平台的普及程度呈现增长态势，不论是预印本的绝对数量 (图1) 还是占各学科论文的比例 (图2)。

在预印本平台逐渐普及的 30 年中，同时发生的还有学术论文数量的大幅增长。建立预印本平台的初衷，是为了更高效、更广泛地传播研究进展，从而弥补学术期刊反应迟缓且容量

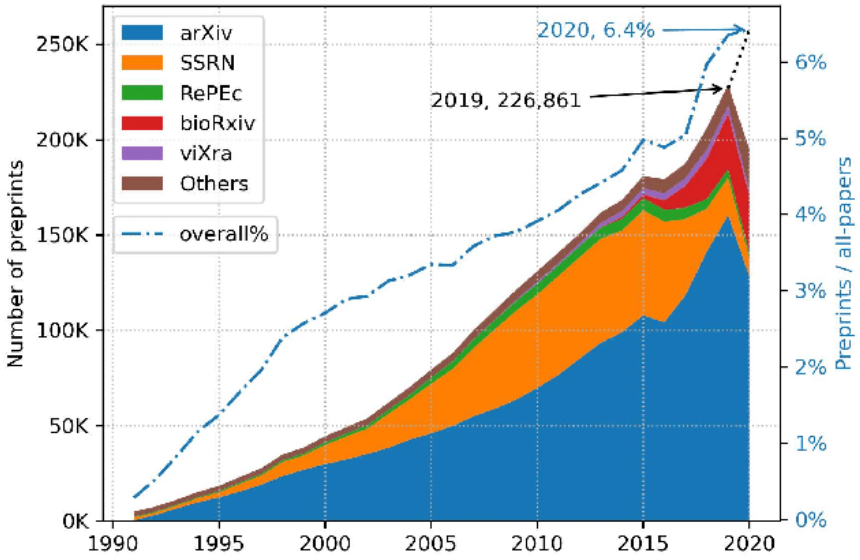

图 1: 各预印本平台上的论文数量，随时间的趋势 (Xie et al., 2021)

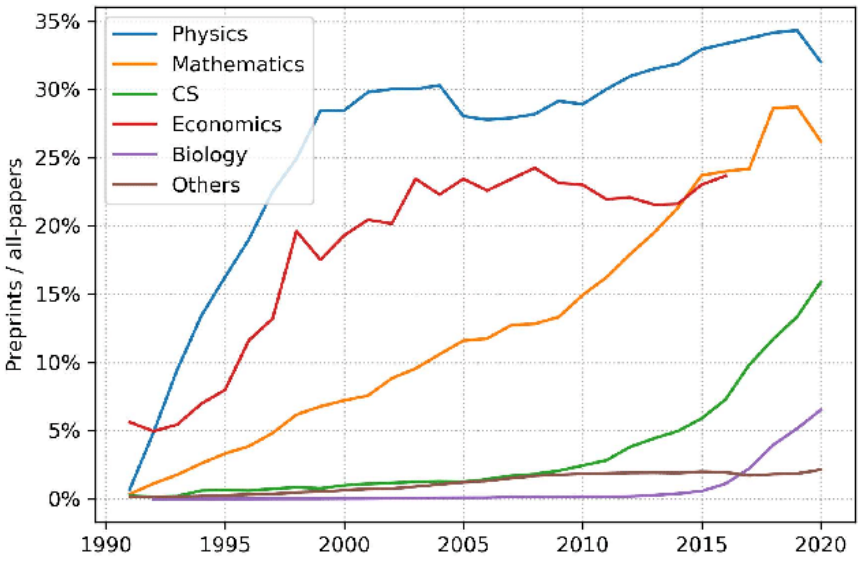

图 2: 各学科中公开在预印本平台上的论文比例，随时间的趋势 (Xie et al., 2021)

有限的缺点。预印本平台 “大容量” 的特性是否意味着，除了那些发表在顶级期刊上的 “明星成果” 之外，其余更为普通的论文也受益于预印本平台的普及，而得到了更多的曝光度？本文通过分析领域内预印本平台普及度对该领域中论文引用量不均衡程度的影响，来尝试回答这一问题。分析得到否定的答案 (未发现显著且稳健的影响)，但不排除是模型和数据中的局限性掩盖了实际存在的效应。

1.2 相关研究和文献

对学术动态的定量研究，主要集中在科学计量学 (Scientometrics)、文献计量学 (Bibliometrics)、科学学 (Science of Science，也称 Metascience) 等领域，其中前两者通常被认为是信息科学 (Information Science，也称情报学) 的子学科。对学术动态的其他实证研究 (包括定性研究和定量研究) 分布在科学社会学 (Sociology of Science) 和科学技术研究 (Science and Technology Studies) 等领域，它们与社会学、科学哲学、科学史等学科密切相关。

在这些研究工作中，只有较少的工作关注预印本平台。这些工作中的绝大部分可以分为以下类别：

- **预印本平台所发布论文与其他论文对比**：很多此类研究对比了有预印本的论文和无预印本的论文在引用量等评价指标上的差异，这些研究将在下面详述。在此之外，还有研究对比了预印本和期刊的发表速度的差异：Johansson et al. (2018) 发现在埃博拉和寨卡病毒爆发期间，与其相关的研究工作的预印本发布比期刊论文发表早 100 天以上，对应对病毒爆发起到有益作用；但是，对这些病毒的研究工作中，仅有 5% 发布了预印本。Tsunoda et al. (2022) 对比了通过预印本平台向期刊投稿的文章和直接向期刊投稿的文章，发现通过预印本平台投稿提高了审稿流程的速度。

- **预印本平台间的对比及其发展趋势**：Chaleplioglou and Koulouris (2021) 和Balaji and Dhanamjaya (2019) 整理并对比了各个预印本平台的学科范围、平台政策和运行情况，后者更进一步指出预印本平台带来的学术交流基础设施和评价指标的改变。Tennant et al. (2018) 回顾了 2012 年起预印本平台的 “爆炸式增长” 和围绕预印本平台的政策改变，并指出了预印本平台的发展趋势以及其中的问题。Xie et al. (2021) 发现预印本平台的文章数量在 30 年间增长 63 倍，但仍只包含所有论文的 4%。

在第一类研究中，很大一部分工作将有预印本的论文和无预印本的论文在引用量等评价指标上进行对比。这些工作与本文较为相关，其结论可对回答本文的问题起到参考作用：如果发现预印本平台使得论文获得更多的引用，且这一效应对非顶级的论文显著，则提示预印本平台的普及有可能降低了引用量 “贫富差距”。

这些工作又可分为以下类别：

- **直接作引用量对比或其他指标对比**

  - Davis and Fromerth (2007) 发现在 arXiv 平台上有预印本的论文比无预印本的论文平均多 35% 的引用量，而在出版商网站上的下载量少 23%。作者认为两类论文本身的质量差距 (而不只是预印本平台直接产生的因果效应) 可能是引用量差距的主要原因之一。
  - Wang et al. (2020b) 发现在图书馆与信息科学 (Library and Information Science) 顶级期刊中，有 arXiv 预印本的论文比其他论文具有更高的引用量、Altmetric 分数和更快的社交媒体报导。
  - Aman (2013) 发现在除生物学以外的各学科中，有 arXiv 预印本的论文都比其他论文拥有更高的引用量和更快的被引速度。
  - Larivière et al. (2014) 发现 arXiv 预印本的引用量低于期刊文章，且在许多学科中前者的引用量随时间衰减更快。注意到该工作对比的是预印本和期刊论文，而非有预印本的期刊论文和无预印本的期刊论文；这是它与本节中列举的其他工作的差异。
  - Serghiou and Ioannidis (2018) 发现 bioRxiv 预印本的引用量显著低于期刊论文，但有 bioRxiv 预印本的期刊论文的引用量和 Altmetric 分数显著高于其他期刊论文。

- **处理文章质量自选择等混杂因素**

  - Fu and Hughey (2019) 在控制了所在领域、期刊影响因子等混杂因素后，发现在 bioRxiv 平台上有预印本的期刊论文，比无预印本的期刊论文平均多 36% 的引用量和 49% 的 Altmetric 分数。
  - Fraser et al. (2020) 在控制了期刊、作者相关的多个混杂因素后，发现拥有 bioRxiv 预印本的文章比其他文章有更高的引用量和 Altmetric 分数。预印本在社交媒体和博客上分享较为广泛，但在主流媒体和百科条目中出现少于期刊论文。
  - Fraser et al. (2022) 通过问卷调查发现，受调查的研究者认为自己在 bioRxiv 平台上发布预印本的的动机与该篇文章的质量无关。
  - Metcalfe (2006) 发现太阳物理学论文中有预印本的期刊论文比平均水平高出 70% 的引用量，且类似的效应对会议论文 (质量通常低于期刊论文) 同样存在，作者认为这提示文章质量的自选择效应并非导致这一差距的原因。

- **区分预印本平台的 “可开放获取性” 和 “获取及时性”**

  - Wang et al. (2020a) 发现在 arXiv 平台上有预印本的数学论文，其对应的期刊论文在引用量、阅读量、社交媒体关注度等方面高于无预印本的期刊论文，且这一差距的原因包括预印本平台的 “可开放获取性” 和 “获取及时性”。
  - Gentil-Beccot et al. (2010) 发现高能物理学中，有预印本的论文因预印本的 “获取及时性” 而拥有引用量优势。在开放获取期刊上发表的论文相比其他论文没有显著的引用量优势，提示 “可开放获取性” 并非导致引用量优势的原因。

  - Moed (2007) 在控制文章质量的混杂因素后，发现拥有 arXiv 预印本的凝聚态物理学文章的引用量优势，来自预印本平台的 “获取及时性” 而非 “可开放获取性”。

整体上，上述研究提示：拥有预印本的论文比其他论文在包括引用量的各指标上表现都更佳，导致这一现象的主要原因包括预印本平台的 “获取及时性” 和文章质量高低对是否发布预印本的自选择效应。

但是，上述研究对于 “顶级和非顶级论文从预印本平台的受益程度差别” 没有给出令人满意的回答，且对混杂因素的处理在整体上也不够令人信服。本文用双重差分法结合控制变量来处理混杂因素，并直接关注引用量的 “贫富差距”，希望解决这两个问题。

另一方面，本节中列举的所有研究工作，或者着眼于论文对比，或者着眼于对预印本平台本身的分析，而没有尝试分析**预印本平台在领域内的普及度对领域整体的效应**。在文献调研中没有发现着眼于此类 “整体效应” 的工作，故本文希望填补这一空白。

2 定性的理论框架

2.1 主要变量

本文的目标是回答：一个领域内预印本平台普及程度的提高，是否会带来论文引用量不均衡程度的降低。

这意味着，我们所关注的核心是以下两个变量：

- 领域 d 在时期 t 的**引用量不均衡度** Y_{dt}
- 领域 d 在时期 t 的**预印本平台普及度** P_{dt}

更具体地说，我们关注的是 P_{dt} 对 Y_{dt} 所产生的因果效应。并且，由于平台普及所产生的效果可能存在滞后，需额外将 P_{dt} 的滞后项纳入考虑。

2.2 控制变量

为进行因果推断，需要控制 P_{dt} 和 Y_{dt} 以外的变量。

所控制的变量分为两个类别：

- **固定效应**
  - **领域固定效应**：不同领域在研究性质、合作方式、学术交流机制等方面可能有着很大不同，而这些因素可能对引用量不均衡性产生影响。这意味着需要加入领域固定效应项以控制这些因素。
  - **时期固定效应**：研究性质、合作方式、学术交流机制等因素往往随着时间演变，故需要加入时间固定效应项以控制这些因素。

- **与 P_{dt} 相关且影响 Y_{dt} 的因素**
  - **领域规模**：规模更大的领域更可能建立自己的预印本平台、规模更小的领域更可能受少数人影响而采用预印本平台，这意味着领域规模与 P_{dt} 相关。同时，领域规模显然可能影响 Y_{dt} 。
  - **相关领域总规模**：即该领域与邻近领域的总规模。由于领域划分并非泾渭分明，且领域间的嵌套关系使得我们难以确定 "领域" 应在哪个嵌套层级上定义，因此将该领域的邻近领域也纳入考虑，作为 "领域规模" 变量的补充。
  - **领域内平均自引时间差**：即该领域内的学者在引用自己的论文时，引文和被引文的发表时间差。这一变量反映了一个研究工作所用的时长 (自引时间差与研究周期正相关) 和领域内整体的引用时间差 (自引时间差与一般的引用时间差正相关)，而这两者越短，该领域越有动机采用预印本平台，意味着它们与 P_{dt} 相关。同时，这两者可能对 Y_{dt} 产生影响。自引时间差可以看作是研究周期和引用时间差的一个不完美的代理变量。之所以采用这一代理变量而非直接采用引用时间差，是因为后者会受到 P_{dt} 的直接的因果影响 (预印本平台的普及会降低引用时间差) 且可能是 P_{dt} 影响 Y_{dt} 的主要途径之一，但前者不会受到 P_{dt} 的直接因果影响 (任何学者无需预印本平台也能知道自己的未发表工作，故预印本平台的普及不会降低自引时间差)。
  - **领域内学者的人均近期文章数、领域内近期文章的篇均作者数**：这两个因素合起来反映了每篇文章需花费的时间周期和人力，且 "篇均作者数" 反映了领域内学者合作的广泛程度。每篇文章耗费精力越大，作者越有动机确保其得到广泛关注；每篇文章工作周期越短，领域进展越快，作者越有动机确保自己的工作在第一时间发布；学者合作越广泛，通过社会网络传播研究工作的速度越快，对预印本平台的需求越低。因此，这些因素与 P_{dt} 相关，且它们可能通过影响引用总量等途径影响 Y_{dt} 。

此外，由于可能存在的时间滞后，需额外将各个控制变量的滞后项纳入考虑。

3 定量模型和估计方法

3.1 回归方程

采用回归分析的方法，回归方程如下。

$$Y_{dt} = \alpha_d + \beta_t + \gamma P_{dt} + \sum_{factor} \delta^{[factor]} C_{dt}^{[factor]} + \epsilon_{dt}$$

这是一个连续变量双重差分 (Callaway et al., 2021) 模型 (DID with continuous treatment)。其中：

- d 表示领域，t 表示时期。

- Y_{dt} 为引用量不均衡度指标。
- P_{dt} 为预印本平台普及度。
- α_d 为领域固定效应，β_t 为时间固定效应。
- $C_{dt}^{[*]}$ 为与 P_{dt} 相关的若干个需要控制的变量。
- ϵ_{dt} 为随机误差项。

对于固定效应，由于不同领域随时间的演变速度和演变方向可能不同，故理想情况下应当加入领域虚拟变量和时期虚拟变量的交叉项。但这使得变量数过大而无法进行后续回归，并且由于领域演变中可能存在的非线性性，时间趋势项在这里可能并不合适。未能加入交叉项，是该模型的局限性之一。

虽然后续回归采用双重差分法，但与 P_{dt} 相关的遗漏变量可能破坏平行趋势的假设，所以依然需要控制 $C_{dt}^{[*]}$ 这些变量。

3.2 变量定义详述

3.2.1 引用量不均衡度指标 Y_{dt}

An et al. (2004) 发现计算机科学论文的引用量近似服从指数为 1.71 的帕累托分布 (也称幂律分布)，提示学术论文的引用量分布是重尾分布，且很可能与计算机科学论文这一子类一样，服从指数小于 2 的帕累托分布。因此，本节的目标是设计一组指标来衡量一个重尾分布 (以及特别地，指数小于 2 的帕累托分布) 的不均衡度，其中不均衡度定义为 “分布的期望受尾部支配的程度”。

本文所采用的指标有：

- **对数期望与对数中位数之差：** $Y_{dt} = \mathrm{E}[\log X] - \mathrm{Med}[\log X]$ ，其中随机变量 X 是一篇等概率随机的论文的引用量，$\mathrm{Med}[\log X] = \mathrm{Q}_{0.5}[\log X]$ 表示 $\log X$ 的 0.5 分位数。
  - 当 X 服从以 α 为指数的帕累托分布时，$\mathrm{E}[\log X] = \frac{1}{\alpha}, \mathrm{Med}[\log X] = \frac{\log 2}{\alpha}$ ，故该指标等于 $\frac{1-\log 2}{\alpha} \approx \frac{0.3}{\alpha}$. 此时 $\log X$ 服从指数分布，故方差 $\mathrm{Var}[\log X] = \alpha^{-2}$ 有限，从而可以通过抽样来有效地估计 $\mathrm{E}[\log X]$.
  - 当 X 服从对数正态分布时，该指标等于 0 .
  - 实际数据中 $X = 0$ 处可能有非零的概率质量，故用 $\log(X+1)$ 代替 $\log X$.
- **对数分位数与对数中位数之差：** $Y_{dt} = \mathrm{Q}_r[\log X] - \mathrm{Med}[\log X]$ ，其中 $r = 0.9$ 或 0.99 ，$\mathrm{Q}_r[\log X]$ 表示 $\log X$ 的 r 分位数。
- **“二八定律” 指标：** $Y_{dt} = -q$ ，其中 $q \in [0, 1]$ 满足引用量最高的的 $q \times 100\%$ 文章，占据了 90% (或 $70\%, 50\%, 30\%$) 的总引用。

上述指标都随引用量不均衡度增大而增大。

值得注意的是，之所以没有采用基于分布偏度或期望值的指标，是因为帕累托分布在指数低于 3 时偏度无定义，且在指数低于 2 时方差无穷 (从而无法通过抽样来有效估计期望值)。

3.2.2 预印本平台普及度 P_{dt}

即该领域在该时期内所有期刊和会议发表的文章中，上传到预印本平台上的文章比例。

3.2.3 需控制的变量 $C_{dt}^{[*]}$

这些变量包括：

- $C_{dt}^{[Size,Size']}$ ：时期 t 领域 d 的规模 ($C_{dt}^{[Size]}$)，及其滞后项 ($C_{dt}^{[Size']}$)
- $C_{dt}^{[RelSize,RelSize']}$ ：时期 t 领域 d 以及相关领域的总规模，及其滞后项
- $C_{dt}^{[CiteLag,CiteLag']}$ ：时期 t 领域 d 的自引时间差，及其滞后项
- $C_{dt}^{[\#Paper,\#Paper']}$ ：时期 t 领域 d 的人均文章数，及其滞后项
- $C_{dt}^{[\#Author,\#Author']}$ ：时期 t 领域 d 的篇均作者数，及其滞后项

这些变量随着 d, t 的变化而有着巨大的变化幅度，故这些变量都采用对数尺度。

理想情况下，变量 $C_{dt}^{[factor]}$ 应关于领域 d 有变截距，即系数 $\delta^{[factor]}$ 应为 $\delta_d^{[factor]}$ 。但由于样本量有限，故去掉下标 d 来降低变量个数；这是该模型的局限性之一。

3.3 参数估计方法

本文会分别采用 Y_{dt} 的不同指标，进行多次回归尝试。这些回归尝试将使用相同的设定，本节给出这一套设定。

回归均采用 LSDV 模型，包含领域维度固定效应 α_d 和时间维度固定效应 β_t .

存在以下数据问题：

- 异方差：由于样本量可能存在的差异，和领域/时期本身的不同，很可能有异方差。
- 序列相关：由于可能的遗漏变量、蛛网现象、持续不止一期的扰动等，导致很可能存在时间维度上的序列相关。由于不同领域同受扰动，导致很可能存在领域维度上的序列相关。

因此采用对异方差和序列相关稳健的标准误。

为决定采用何种稳健标准误，先进行横截面相关性 (即不同领域的相关性) 的 Pesaran's CD 检验 (Pesaran, 2004)，根据检验结果决定采用以下两种稳健标准误中的一种。

- 双重聚类 (Double-Clustering, 下称 DC) 稳健标准误 (Thompson, 2011; Cameron et al., 2011)：针对同时有横截面相关性和时间序列相关性的情形。
- Newey and West (下称 NW) 稳健标准误 (Newey and West, 1986)：针对无横截面相关性、只有时间序列相关性的情形。

4 数据收集和处理

4.1 数据概览

(数据集和代码等公开在 `https://github.com/TianyiQ/preprint-and-disparities`)

构建的数据集包含 19 个领域 (对应 OpenAlex 数据库中的 19 个零级概念，见表1) 在 40 个时期 (从 1979 年到 2018 年，每年作为一个时期) 中的统计数据。统计数据中包含 20 个变量，其含义和概况见表2 。760 个数据点中，有 3 个数据点因样本量不足而缺少部分变量数值。

图3展示了各领域中预印本普及度 P_{dt} 和引用量不均衡度 Y_{dt} 随时间的变化趋势。

OpenAlex ID	**领域名称**	**OpenAlex ID**	**领域名称**
C17744445	Political science	C185592680	Chemistry
C138885662	Philosophy	C142362112	Art
C162324750	Economics	C144024400	Sociology
C144133560	Business	C127413603	Engineering
C15744967	Psychology	C205649164	Geography
C33923547	Mathematics	C95457728	History
C71924100	Medicine	C192562407	Materials science
C86803240	Biology	C121332964	Physics
C41008148	Computer science	C39432304	Environmental science
C127313418	Geology		

表 1: 19 个领域的名称和 OpenAlex ID

4.2 数据思路

数据收集流程基于 OpenAlex (Priem et al., 2022) 提供的数据库和 API，收集过程中使用了 openalexR (Aria and Le, 2023) 实现的接口。数据收集和处理基于 R (R Core Team, 2022) 和 Python (Van Rossum and Drake, 2009) 实现，并使用了 plm (Croissant and Millo, 2008) 、stargazer (Hlavac, 2022) 、NumPy (Harris et al., 2020) 和 Pandas (Wes McKinney, 2010) 等软件包。

数据收集和处理分为采样和回归两个步骤。

变量	说明	N	Mean	St. Dev.	Min	Max
nsamp_paper	时期 t 领域 d 采样的文章数	760	949.671	1,206.067	0	5,587
ex_log	$\mathrm{E}[\log X]$	758	2.806	0.777	0.000	4.212
ex	$\mathrm{E}[X]$	758	59.665	45.786	0.000	481.950
log_med	$\mathrm{Med}[\log X]$	758	2.839	0.856	0.000	4.322
med	$\mathrm{Med}[X]$	758	21.729	15.261	0.000	74.365
repo_ratio	P_{dt}	758	0.035	0.084	0.000	0.847
log_field_sz	时期 t 领域 d 规模的对数	758	6.923	2.539	1.829	12.792
log_rel_field_sz	时期 t 领域 d 相关领域规模的对数	758	8.466	2.135	4.209	13.369
n_top_author_d	领域 d 高产学者总数	760	7,472.632	11,338.300	221	45,597
nsamp_author_d	领域 d 采样的学者总数	760	254.263	81.519	63	347
nsamp_author_dt	时期 t 领域 d 采样的学者总数	760	168.645	99.561	5	344
self_cite_lag	时期 t 领域 d 的自引时间差	758	856.258	103.782	653.656	1,096.000
log_self_cite_lag	时期 t 领域 d 自引时间差的对数	758	6.747	0.120	6.484	7.000
log_papers_per_author	时期 t 领域 d 人均文章数的对数	760	1.373	0.648	0.000	2.974
log_authors_per_paper	时期 t 领域 d 篇均作者数的对数	758	1.270	0.609	0.000	4.408
log_cite_top_1_in_10	$\mathrm{Q}_{0.9}[\log X]$	758	4.625	0.879	0.000	7.120
log_cite_top_1_in_100	$\mathrm{Q}_{0.99}[\log X]$	758	6.078	1.179	0.000	9.415
cite_30pc_dominator	“二八定律” 指标 q (30%)	757	0.049	0.085	0.0004	1.000
cite_50pc_dominator	“二八定律” 指标 q (50%)	757	0.101	0.088	0.005	1.000
cite_70pc_dominator	“二八定律” 指标 q (70%)	757	0.205	0.095	0.006	1.000
cite_90pc_dominator	“二八定律” 指标 q (90%)	757	0.439	0.107	0.030	1.000

表 2: 数据集中的变量概况

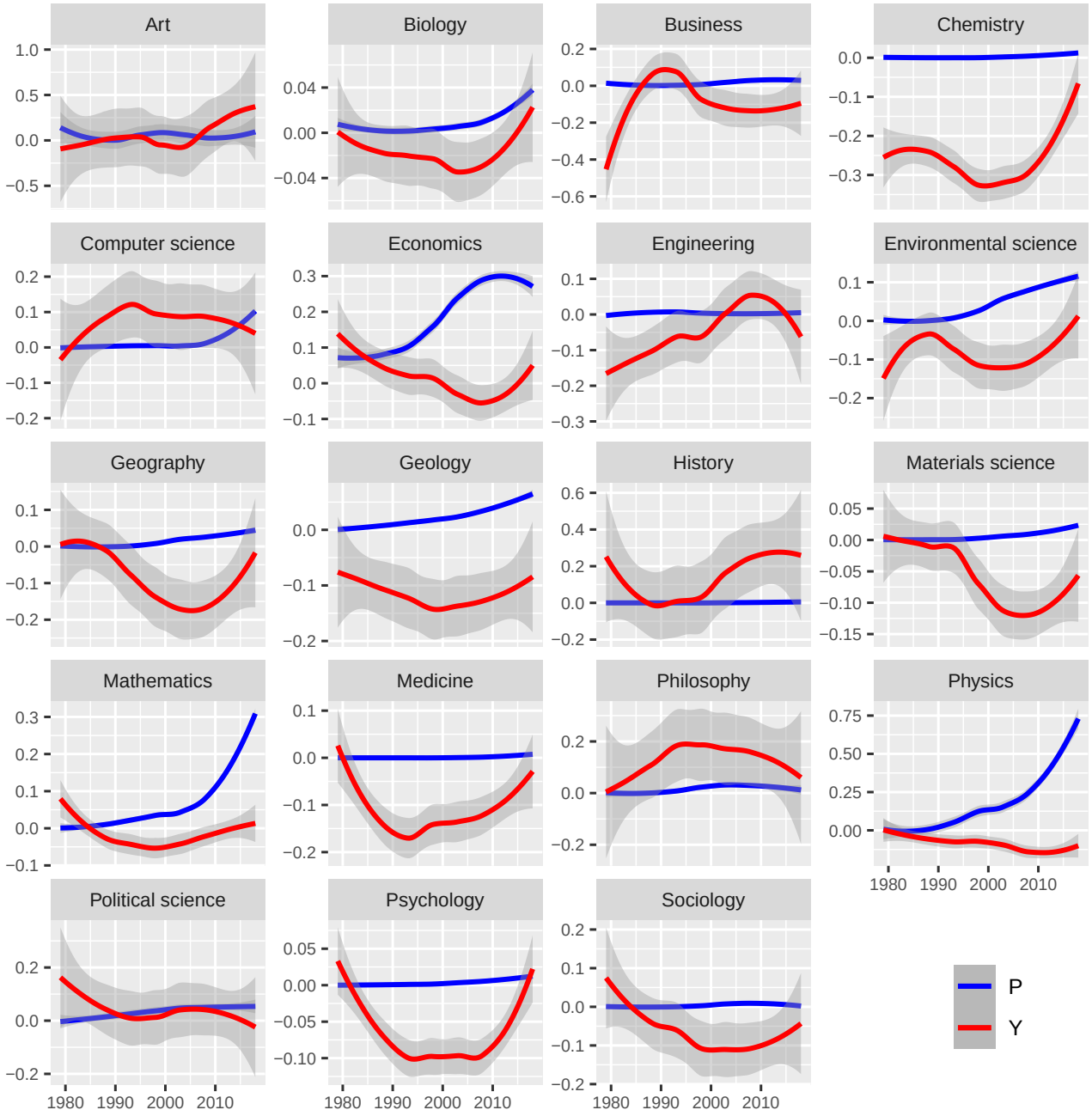

图 3: 各领域中预印本普及度 (P_{dt}，蓝色) 和引用量不均衡度 ($Y_{dt} = \mathrm{E}[\log X] - \mathrm{Med}[\log X]$，红色) 随时间的变化趋势，$Y_{dt}$ 越大表示越不均衡，曲线经过平滑处理

- **采样**：对每个领域，找到其中最高产 (发表作品数最多) 的一部分作者，并在这部分作者中等概率随机抽样 (“对作者采样”)，并且把这些作者所发表的所有满足特定条件的文章作为采样的文章集合 (“对文章采样”)，然后基于作者和文章样本计算每个数据点处各变量的取值 (“变量取值计算”)。

- **回归**：对 Y_{dt} 指标的不同选择和解释变量的不同选择，分别进行回归，以判断是否存在显著且稳健的效应。这将同时起到稳健性检验的作用。

其中，采样步骤的主要瓶颈在于 API 调用频率的限制。因此，采样流程被设计为以尽可能少的调用次数获取尽可能多的数据。

4.3 采样流程

4.3.1 对作者采样

对作者采样，需对每门领域找到其中最高产的一部分作者，并在其中等概率随机抽样。这一过程分为以下步骤。

- **划定作品数阈值**：对每个领域，考察所有在 OpenAlex 数据库中与其相关的学者，计算这些学者中作品最多的千分之一的学者中最少的作品数，将这一数量作为作品数阈值。

- **学者归类**：对于所有达到作品数阈值的学者 (称为 “高产学者”)，从 OpenAlex 数据库获取其简要信息 (不含作品列表)。对每名学者，找到与其相关度最大的领域，并将其归入该领域。

- **本地抽样**：对每个领域，在高产学者列表中随机抽样。给不同的领域分配不同的抽样个数，以使得后续对文章采样时，每个领域所调用的 API 次数大致相同。

上述过程中，还需额外对一些异常情形 (比如作品数超过 5000 且大部分不是期刊和会议论文的作者) 进行剔除。

4.3.2 对文章采样

对于采样获得的所有作者，批量查询其发表的所有文章，筛选其中同时满足以下条件的，作为采集的文章样本。

- **条件 1**：该文章的 OpenAlex 条目中有至少一条参考文献。

- **条件 2**：该文章的 OpenAlex 条目中，至少一个发表途径 (Venue) 满足类型为期刊或会议，或至少一个发表途径为 DOAJ 或 DergiPark (这是两个收录正式出版的学术文章的平台) 。

注意到这些文章样本并非从所有文章中等概率随机抽取，这是这种采样方式的一大缺陷。为了弥补这一缺陷，本文还设计了一种基于马尔可夫链蒙特卡洛和两阶段抽样的方法，能够按照自定义的分布进行抽样，但是会需要较多的 API 调用次数，且会在一定程度上增大序列相关性。本文的实验没有实现这一方法，该方法细节见附录A 。

此外，由于解释变量和被解释变量从同一组样本算得，故二者的测量误差有相关性。理想的做法是独立地获取两组样本，并用两组样本分别计算解释变量和被解释变量取值。为弥补这一问题，可以将 $t-1$ 期和 $t+1$ 期的 repo_ratio 取平均 (记作 repo_ratio_avg)，用以在 t 期的回归方程中代替解释变量 repo_ratio .

4.3.3 变量取值计算

获得文章样本后，还需计算所需各变量的取值。大部分变量的计算方式比较直接，以下仅列举非平凡的那些。

- P_{dt} **的计算：**在时期 t 领域 d 的文章样本中，直接计算拥有预印本的文章比例。一篇文章被识别为拥有预印本，当其 OpenAlex 条目中至少一个发表途径属于表3 。PhilArchive，E-LIS 和 PsyArXiv 等平台未列在其中，因为 OpenAlex 所收录的这些平台的文章很少。

- **nsamp_paper 和 nsamp_author_dt 的计算：**对每名作者，计算其最早发表和最晚发表的文章年份，并认为该名作者在这两个年份之间都处于活跃状态。nsamp_author_dt 定义为领域 d 在年份 t 中活跃的被采样作者数。nsamp_paper 即所有采样文章中，作者属于领域 d 且发表年份为 t 的数量。

- **人均文章数的计算：**人均文章数即 nsamp_paper 与 nsamp_author_dt 的商。后续还需加一后取对数。

- **篇均作者数的计算：**篇均作者数，即所有采样文章中，作者属于领域 d 且发表年份为 t 的文章中平均的署名作者数，后续还需加一后取对数。计算过程中为节省 API 调用次数，若发现有太多文章的作者数超过 100 ，将对部分作者数超限的文章，用近年的数据对其作者数进行估测。

- **平均自引时间差的计算：**对每一名作者的每一篇文章，找到其首次被同一名作者引用的时间，计算发表与被引的时间差；在作者属于领域 d 且发表年份为 t 的文章中，这一时间差的平均值即为平均自引时间差 (后续还需加一后取对数)。为保证 2018 年发表的文章不会因时间太近而来不及被引，在平均自引时间差的计算中忽略一切时间差超过 3 年 (1095 天) 的自引；若不存在 3 年以内的自引，则认为该文章的自引时间差为 1095 天。为保证公平性，对所有年份的文章都采用这一标准。

- **领域规模的计算：**取领域 d 在年份 t 中的所有文章样本，计算与领域 d 相关度平方的均值，并假设该领域中未被采集的文章的均值与之相同 (未被采集的文章数量按高产作者数量估算)，从而估算该领域中所有文章的相关度平方和。

- **相关领域总规模的计算**：在领域规模的计算中，取消按相关度平方的加权，认为所有文章等权，转而考虑文章总数。由于 OpenAlex 会给一篇文章标上所有相关的领域，因此这样做可以得到相关领域的总论文数。

发表途径 (Venue) 名称	OpenAlex Venue ID
arXiv	V4306400194
Zonodo	V4306400562
HAL	V4306402512 V4306402144 (and more)
bioRxiv	V2734324842 V4306402567
RePEc	V4306401271 V3121261024
medRxiv	V4306400573
SSRN	V4210172589 V2751751161

表 3: 用于识别预印本文章的发表途径列表

4.4 回归流程

回归时，将对 Y_{dt} 采用不同的指标，并同时尝试解释变量的不同选择，对所有不同的组合都进行回归并比较结果，以起到稳健性检验的作用。

解释变量有 2 种不同的选择：

- **解释变量方案 A**：repo_ratio_avg + 需控制的 10 个变量 $C_{dt}^{[*]}$
- **解释变量方案 B**：repo_ratio_prev + 需控制的 10 个变量 $C_{dt}^{[*]}$

其中 t 期的 repo_ratio_avg 如4.3.2小节所述，是 $t-1$ 和 $t+1$ 期的 repo_ratio 取平均而得。repo_ratio_avg 用于代替 repo_ratio ，可以降低其测量误差与被解释变量 Y_{dt} 测量误差的相关性，因为 repo_ratio_avg 计算所用的文章样本与 Y_{dt} 计算所用的文章样本不同。

repo_ratio_prev 是 repo_ratio 的滞后项，t 期的 repo_ratio_prev 等于 $t-2$ 期的 repo_ratio . 之所以考虑滞后项，是因为预印本平台普及所产生的效应可能有所滞后。

被解释变量 Y_{dt} 有 7 种不同的选择 (详见3.2.1小节)：

- **被解释变量方案 A**：$Y_{dt} = \mathrm{E}[\log X] - \mathrm{Med}[\log X]$

- **被解释变量方案 B1-B2：** $Y_{dt} = \mathrm{Q}_r[\log X] - \mathrm{Med}[\log X]$ ，其中 r 分别取 $0.9, 0.99$.
- **被解释变量方案 C1-C4：** $Y_{dt} = -q$ ，其中 $q \in [0,1]$ 满足引用量最高的 $q \times 100\%$ 的文章，分别占据了 $30\%, 50\%, 70\%, 90\%$ 的总引用。

后续将对所有 14 种组合都分别进行 LSDV 回归，并对比结果。首先会对 14 种组合都进行 Pesaran's CD 检验 (Pesaran, 2004)，根据检验结果判断是否存在横截面相关性。为保证 14 种组合之间统一，如果相当一部分组合呈现出横截面相关性，则对所有组合都统一采用 DC 稳健标准误 (Thompson, 2011; Cameron et al., 2011)，否则采用 NW 稳健标准误 (Newey and West, 1986) 。

5 结果和讨论

横截面相关性 CD 检验的结果见表4 。可以看到，在相当一部分情况中存在显著的横截面相关性，因此下面采用对横截面相关性、时间序列相关性、异方差性同时稳健的 DC 标准误 (Thompson, 2011; Cameron et al., 2011)。

解释变量方案	被解释变量方案	p 值 (CD 检验)
A	A	0.402
A	B1 (0.9)	0.869
A	B2 (0.99)	0.001***
A	C1 (30%)	0.224
A	C2 (50%)	0.070*
A	C3 (70%)	0.003***
A	C4 (90%)	0.003***
B	A	0.058*
B	B1 (0.9)	0.514
B	B2 (0.99)	0.004***
B	C1 (30%)	0.008***
B	C2 (50%)	0.468
B	C3 (70%)	0.045**
B	C4 (90%)	0.021**
Note:	*p<0.1; **p<0.05; ***p<0.01	

表 4: CD 检验结果，p 值越低表示横截面相关性越显著

表5是所有组合分别的回归结果。以 "解释变量方案 A + 被解释变量方案 A" 这一组合为例，回归结果见表6 。

解释变量方案	被解释变量方案	$\hat{\gamma}$
A	A	-0.100
		($p = 0.570$)
A	B1 (0.9)	0.319
		($p = 0.479$)
A	B2 (0.99)	-0.570
		($p = 0.558$)
A	C1 (30%)	-0.077
		($p = 0.120$)
A	C2 (50%)	-0.048
		($p = 0.253$)
A	C3 (70%)	-0.047
		($p = 0.534$)
A	C4 (90%)	-0.070
		($p = 0.543$)
B	A	-0.992^{**}
		($p = 0.025$)
B	B1 (0.9)	0.014
		($p = 0.975$)
B	B2 (0.99)	-0.825
		($p = 0.359$)
B	C1 (30%)	0.010
		($p = 0.817$)
B	C2 (50%)	0.022
		($p = 0.719$)
B	C3 (70%)	0.031
		($p = 0.761$)
B	C4 (90%)	0.110
		($p = 0.388$)
Note:		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

表 5: 回归结果中解释变量 P_{dt} (即预印本平台普及度) 的系数 $\hat{\gamma}$

Independent variable	*Statistics*
repo_ratio_avg	−0.100
	(p = 0.570)
log_field_sz	0.043
	(p = 0.445)
log_field_sz_prev	0.082
	(p = 0.159)
log_rel_field_sz	0.226***
	(p = 0.002)
log_rel_field_sz_prev	−0.287***
	(p = 0.003)
log_self_cite_lag	0.529*
	(p = 0.066)
log_self_cite_lag_prev	−0.406***
	(p = 0.0002)
log_papers_per_author	−0.262
	(p = 0.191)
log_papers_per_author_prev	0.088
	(p = 0.333)
log_authors_per_paper	0.216
	(p = 0.260)
log_authors_per_paper_prev	−0.165
	(p = 0.164)
Observations	693
R^2	0.369
Adjusted R^2	0.304
F Statistic	5.645*** (df = 65; 627)
Note:	*p<0.1; **p<0.05; ***p<0.01

表 6: 解释变量方案 A + 被解释变量方案 A 组合下的回归结果，省略了固定效应项和常数项

从表 5 中可以看到，在 14 种组合中，预印本平台普及度 P_{dt} 的系数正负不一，且仅在一种组合下显著。这表明，回归分析没有发现预印本平台普及度 P_{dt} 对引用量不均衡度 Y_{dt} 的效应。

6 结论

6.1 结果总结和局限性

本文中经过回归分析，没有发现预印本平台普及度 P_{dt} 对领域内引用量不均衡度 Y_{dt} 存在显著影响。

这一否定的结果可能意味着这样的影响确实不存在，也可能是模型和数据中的局限性所导致的结果。本文的方法中有以下潜在的局限性：

- **模型的局限性：**
  - 未能以合适的方式，向回归方程中加入时间固定效应和领域固定效应之间的交互效应。
  - 为降低变量个数，未能对所控制的变量加入领域维度的变斜率。
  - 连续变量双重差分模型需要比较强的平行趋势假设 (Callaway et al., 2021)，这一假设在此处是否成立仍有待考察。
  - 尚不确定所控制的变量 $C_{dt}^{[*]}$ 是否囊括了所有与 P_{dt} 相关且影响 Y_{dt} 的因素。
- **数据的局限性：**
  - 数据的个体数 (领域个数) 和时期数可能不够大，可能不足以用作推断，尤其是在采用了序列相关稳健标准误之后。增大个体数，可以通过把一级概念也纳入考虑来实现。增大时期数，可以通过细分时段 (即把一个时期定义得比一年更短) 来实现。但考虑到 API 调用次数是瓶颈，要实现这两点可能需要在本地搭建 OpenAlex 数据库，这在技术上有一定挑战。
  - 所有变量用同一组样本计算，导致不同变量间测量误差可能存在相关性。通过采用 repo_ratio_avg 可以缓解这一问题，但是否完全解决了问题仍有待考察。
  - 对变量测量误差和对代理变量未做细致讨论，因此尚不清楚它们对结果产生了怎样的影响。

6.2 展望

目前为止，围绕预印本平台的研究主要着眼于微观层面 (不同类型文章之间的对比)，或着眼于平台本身。本文尝试另辟蹊径，研究预印本平台对领域整体的宏观影响，希望起到抛砖引玉的作用。

在这一问题上，后续的研究可以尝试克服本文的局限性，例如设计更可靠的模型，或扩大数据集规模。另一方面，类似的研究不必局限于“引用量分布”这个单一的话题，而可以扩展到一个领域的各个方面。本文的研究问题和研究思路可以作为这一方向的“概念验证”，但远未探索完所有的可能性。

参考文献

Aman, V. (2013). The potential of preprints to accelerate scholarly communication-a bibliometric analysis based on selected journals. *arXiv preprint arXiv:1306.4856*.

An, Y., Janssen, J., and Milios, E. E. (2004). Characterizing and mining the citation graph of the computer science literature. *Knowledge and Information Systems*, 6(6):664–678.

Aria, M. and Le, T. (2023). *openalexR: Getting Bibliographic Records from 'OpenAlex' Database Using 'DSL' API*. https://github.com/massimoaria/openalexR, https://massimoaria.github.io/openalexR/.

Balaji, B. P. and Dhanamjaya, M. (2019). Preprints in scholarly communication: Re-imagining metrics and infrastructures. *Publications*, 7(1):6.

Callaway, B., Goodman-Bacon, A., and Sant'Anna, P. H. (2021). Difference-in-differences with a continuous treatment. *arXiv preprint arXiv:2107.02637*.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29(2):238–249.

Chaleplioglou, A. and Koulouris, A. (2021). Preprint paper platforms in the academic scholarly communication environment. *Journal of Librarianship and Information Science*, page 09610006211058908.

Croissant, Y. and Millo, G. (2008). Panel data econometrics in R: The plm package. *Journal of Statistical Software*, 27(2):1–43.

Davis, P. and Fromerth, M. (2007). Does the arxiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics*, 71(2):203–215.

Fraser, N., Mayr, P., and Peters, I. (2022). Motivations, concerns and selection biases when posting preprints: a survey of biorxiv authors. *Plos one*, 17(11):e0274441.

Fraser, N., Momeni, F., Mayr, P., and Peters, I. (2020). The relationship between biorxiv preprints, citations and altmetrics. *Quantitative Science Studies*, 1(2):618–638.

Fu, D. Y. and Hughey, J. J. (2019). Meta-research: Releasing a preprint is associated with more attention and citations for the peer-reviewed article. *Elife*, 8:e52646.

Gentil-Beccot, A., Mele, S., and Brooks, T. (2010). Citing and reading behaviours in high-energy physics. *Scientometrics*, 84(2):345–355.

Ginsparg, P. (2021). Lessons from arxiv' s 30 years of information sharing. *Nature Reviews Physics*, 3(9):602–603.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Hlavac, M. (2022). *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Social Policy Institute, Bratislava, Slovakia. R package version 5.2.3.

Johansson, M. A., Reich, N. G., Meyers, L. A., and Lipsitch, M. (2018). Preprints: An underutilized mechanism to accelerate outbreak science. *PLoS medicine*, 15(4):e1002549.

Larivière, V., Sugimoto, C. R., Macaluso, B., Milojević, S., Cronin, B., and Thelwall, M. (2014). arxiv e-prints and the journal of record: An analysis of roles and relationships. *Journal of the Association for Information Science and Technology*, 65(6):1157–1169.

Metcalfe, T. S. (2006). The citation impact of digital preprint archives for solar physics papers. *Solar Physics*, 239(1):549–553.

Moed, H. F. (2007). The effect of "open access" on citation impact: An analysis of arxiv's condensed matter section. *Journal of the American Society for Information Science and Technology*, 58(13):2047–2054.

Newey, W. K. and West, K. D. (1986). A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix.

Pesaran, M. H. (2004). General diagnostic tests for cross section dependence in panels (iza discussion paper no. 1240). *Institute for the Study of Labor (IZA)*.

Priem, J., Piwowar, H., and Orr, R. (2022). Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Serghiou, S. and Ioannidis, J. P. (2018). Altmetric scores, citations, and publication of studies posted as preprints. *Jama*, 319(4):402–404.

Tennant, J., Bauin, S., James, S., and Kant, J. (2018). The evolving preprint landscape: Introductory report for the knowledge exchange working group on preprints.

Thompson, S. B. (2011). Simple formulas for standard errors that cluster by both firm and time. *Journal of financial Economics*, 99(1):1–10.

Tsunoda, H., Sun, Y., Nishizawa, M., Liu, X., and Amano, K. (2022). How preprint affects the publishing process: Duration of the peer review process between biorxiv and journal papers. *Proceedings of the Association for Information Science and Technology*, 59(1):505–509.

Van Rossum, G. and Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

Wang, Z., Chen, Y., and Glänzel, W. (2020a). Preprints as accelerator of scholarly communication: An empirical analysis in mathematics. *Journal of Informetrics*, 14(4):101097.

Wang, Z., Glänzel, W., and Chen, Y. (2020b). The impact of preprints in library and information science: an analysis of citations, usage and social attention indicators. *Scientometrics*, 125(2):1403–1423.

Wes McKinney (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Xie, B., Shen, Z., and Wang, K. (2021). Is preprint the future of science? a thirty year journey of online preprint services. *arXiv preprint arXiv:2102.09066*.

Appendices

A 附录：基于马尔可夫链蒙特卡洛的文章采样方法

这里的方法是对 Metropolis-Hastings 算法修改而得，用于在仅支持局部查询 (即查询一篇文章的参考文献和引用其的文章的情况) 的文献互引图上，实现可控且灵活的随机抽样。采样方法分为以下步骤：

- **规定目标分布和转移图**

  - 马尔可夫链蒙特卡洛 (Markov Chain Monte Carlo，下称 MCMC) 算法通过在转移图上随机游走来实现对目标分布 (分布的支持集是转移图的结点集合) 的抽样。因此，需要规定该算法的转移图和目标分布。
  - 转移图是一张带权有向图，以文章为结点，包含引用边和被引边，引用边和被引边互为反向边。每个点的所有引用边需要等权、所有被引边需要等权，在此前提下具体的权重分配没有要求，可以视算法需要 (如减小序列相关性、降低 MCMC 算法的热身用时) 而随意设定。
  - 目标分布的选定，采用两阶段方法，用一阶段的抽样结果确定二阶段的目标分布。一阶段的目标是了解随机文章与各领域相关度平方的期望，二阶段的目标是使得抽样结果中各领域的相关度平方和尽可能平均。注意到，二阶段中会取所访问的所有结点的全体邻居 (包括结点自身) 作为采样结果，故应平均化的是邻居集合与各领域的相关度平方和。
  - 在一阶段中，所有文章等概率分布，即目标分布为均匀分布。在二阶段中，对一篇文章 i ，设其与每个领域的相关度平方构成的向量为 $\mathbf{r}_i \in \mathbb{R}^{19}$ ，则其权重 w_i 正比于 $\mathbf{r}_i \cdot \mathbf{w}$ ，其中 $\mathbf{w}$ 是某个非负的权重向量。
  - $\mathbf{w}$ 的选取方式是，尽可能使得一阶段抽样结果的 $\sum_i \sum_{j \in N(i) \cup \{i\}} w_j \mathbf{r}_j$ 向量在各维度上平均，其中 $N(i)$ 为文章 i 的出度集合。即 $\text{normalize}(\mathbf{M^T R w}) \approx (1, \cdots, 1)^T$ ，其中矩阵 $\mathbf{M}$ 中第 i 行是文章 i 的所有邻居 (包括 i 自身) 的相关度平方向量之和；$\mathbf{Rw}$ 即等于向量 $(w_i)_{i=1}^{19}$. 这样的 $\mathbf{w}$ 可以通过解矩阵方程或线性规划 (后者可以保证非负性等性质) 求得。

- **一阶段抽样**
  - 在引用-被引图上，从随机起点运行 Metropolis-Hastings 算法，以实现均匀抽样。
- **二阶段抽样**
  - 先按照目标分布的前述定义，计算二阶段的目标分布。
  - 然后在引用-被引图上，从随机起点运行 Metropolis-Hastings 算法，以实现按二阶段目标分布抽样。
  - 由于 OpenAlex API 的特性，查询一篇文章时可以获得关于其邻居的很多信息，因此在访问一篇文章时可将它和它的所有邻居都加入所获得的样本。

上述算法可以通过 $O(M)$ 次 API 调用，获得由大约 $\bar{d}M$ 篇文章构成的随机样本，其中 $\bar{d}$ 为转移图 (互引图) 的平均度数，M 为 MCMC 的马尔可夫链长度。

最后，一个需要注意的问题是连通性。如果转移图分为多个连通分量且不存在一个几乎包含所有点的巨片 (Giant Component)，则需要多次运行上述算法并将抽样集合取并集，否则抽样的方差可能会非常大。