论文、撤稿和预印本的状态：来自 CrossRef 数据库的证据（2004 – 2024）

Khalid M. Saqr[1,*]

[1]College of Engineering and Technology, AASTMT, 1029 Alexandria, Egypt
[*]Corresponding author: *k.saqr@aast.edu*

arxiv:2506.21232v1 中译本

摘要

一项为期 20 年的 CrossRef 元数据分析表明，全球学术产出——包括出版物、撤稿和预印本——表现出显著的惯性增长，可以用几乎无法区分拟合优度的指数模型、二次模型和逻辑斯谛模型来很好地描述。特别是撤稿动态保持稳定，并且受到 COVID-19 冲击的影响极小，其贡献不足总通知数的 1%。自 2004 年以来，出版物每 9.8 年翻一番，撤稿每 11.4 年翻一番，预印本则以最快的速度，每 5.6 年翻一番。研究结果强调了一个系统在不变的结构性瓶颈下持续承受压力的状态。尽管模型预测在 2024 年后出现分歧，但证据表明学术交流的未来轨迹将由持久的系统惯性而非偶发性中断决定——除非通过政策或人工智能驱动的改革有意进行重新定向。

介绍

过去二十年见证了学术产出的空前增长，这得益于数字基础设施、预印本仓库以及诸如 COVID-19 之类的全球危机。长期以来被视为科学诚信守护者的同行评审制度现在正面临巨大的压力。多项近期研究表明了这一“同行评议者危机”，发现编辑们必须比以往任何时候都发出更多的邀请才能找到审稿人，从而延长出版延迟并威胁传统编校流程的可持续性 [2–5]。尽管在发现和通信技术方面有所改进，但这些系统瓶颈仍然存在甚至恶化，并对整个研究生态系统产生了连锁反应。

大规模、数据驱动的分析已经开始详细地描绘这一转变。Davidson 等人。[1] 发现预印本和经过同行评审的 COVID-19 试验文章在很大程度上报告了一致的效果估计，支持快速分发模型的可靠性，尽管在结果报告方面仍存在一些差异。Whitaker 等人。[7] 表明，在疫情期间，出版动态和引用行为发生了根本性的变化：预印本激增，其重要性波动，并且公众和科学界的注意力更加集中地关注 COVID-19 研究。Tsunoda 等人。[6] 证

明了通过预印本传播的文章一旦发表，其获得的引用次数显著高于直接提交给期刊的文章，证实了在危机和快速变化时期预印本的受众放大效应。

然而，这一生态系统的一个显著特征是其惯性：即使在大流行规模的冲击下，科学进步与错误校正之间的核心同步关系仍然保持。之前的元研究通常依赖于单一的回归形式（通常是指数型），可能忽略了转折点和非线性变化，这些信号表明现实世界的限制。迫切需要并行测试多种增长模型——尤其是在系统从扩张阶段过渡到饱和阶段时。

在这里，我们通过利用全面的 CrossRef 元数据来建模从 2004 年到 2024 年的年度出版物、撤稿和预印本的趋势，并系统地比较指数拟合、二次拟合和逻辑斯蒂拟合。通过对科学系统的惯性和适应性进行量化，我们识别了当前增长的边界、纠错机制的韧性以及学术交流新结构时代早期预警信号。

结果

在 2004 年至 2024 年间，CrossRef 索引的学术出版物数量增加了四倍多，从 228 万增加到 1012 万。所有三种模型类型——逻辑回归（$R^2 = 0.920$）、二次（$R^2 = 0.919$）和指数（$R^2 = 0.918$）——对历史轨迹的拟合精度几乎无法区分（图 1）。然而，对未来至 2030 年的预测显示出明显的差异：只有逻辑回归模型预示饱和状态，而指数和二次拟合则预测无限加速。这些结果显示，截至 2024 年，学术系统仍然受惯性驱动力的支配，并且模型机制的重大转变尚未实现。

撤回趋势

撤回通知从 6,099 增长到 30,689。拟合几乎相同：$R^2 = 0.896$（逻辑斯蒂），0.894（二次），0.892（指数）。尽管存在 COVID-19 大流行和间歇性的元数据修正峰值（2010 – 2011，2023 年），总体趋势依然保持惯性平滑（图 2）。COVID-19 撤回通知在顶峰时达到 165 例，并且从未超过年度总数的 1%，这代表了一种至少比主要趋势低两个数量级的冲击。主要时间序列的韧性以及所有模型拟合的高度一致，表明撤回的增长最好被理解为一个惯性过程，对外部干扰具有相对较强的抵抗力。

预印本 超级增长与体制转型

预印本显示最快的倍增时间（5.6 年）和更陡峭的非线性。逻辑斯蒂拟合（$R^2 = 0.907$）略优于二次拟合（0.853），远超指数拟合（0.721），但所有拟合在 2024 年前都紧密吻合（图 3）。未来预测（2024 年后）的收敛表明，新的预印本服务器浪潮可能超过传统控制——尤其是随着生成式 AI 加速生产而同行评审滞后的情况下。

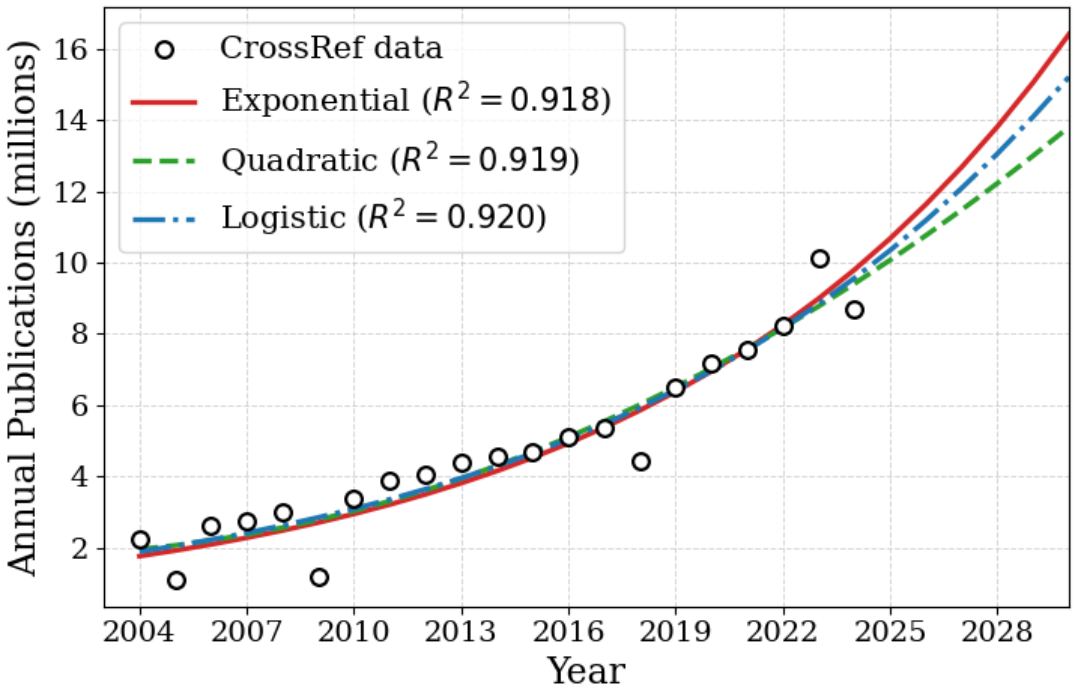

图 1: 年度 CrossRef 索引的研究出版物（2004–2024，圆圈）与指数（红色）、二次（绿色虚线）和逻辑斯蒂（蓝色点划线）模型拟合的未来预测（至 2030 年）对比。所有模型均紧密匹配过去的数据；仅未来趋势有所分歧。

系列间相关性分析

为了评估学术出版物、撤稿和预印本趋势之间的统计关联程度，我们计算了所有时间序列对的皮尔逊相关系数（r）、斯皮尔曼等级相关系数（ρ）和肯德尔 tau（τ）。相关性既针对完整年度时间序列（2004–2024），也针对减去与 COVID-19 相关的撤稿的撤稿系列进行了计算，以控制疫情特定冲击。

表 1 总结了结果。该出版物-预印本对在所有测量指标上表现出极强的正相关性（皮尔逊 $r = 0.992$，斯皮尔曼 $\rho = 0.974$，肯德尔 $\tau = 0.882$，全部 $p < 0.001$），表明了同步的指数增长和紧密耦合的扩张动态。出版物与总撤稿之间的相关性中等（皮尔森 $r = 0.855$），但对于基于排名的指标则进一步下降（斯皮尔曼 $\rho = 0.779$，肯德尔 $\tau = 0.623$），反映出非线性和间歇性的趋势（例如，元数据更正）。移除 COVID-19 撤稿对这些系数的影响微乎其微，强化了疫情期间事件并非系统同步性主要驱动因素的发现。预印本和撤稿显示出最低的相关性，但仍然是正值（皮尔森 $r = 0.742$），这与预印本作为部分独立子系统的操作一致。

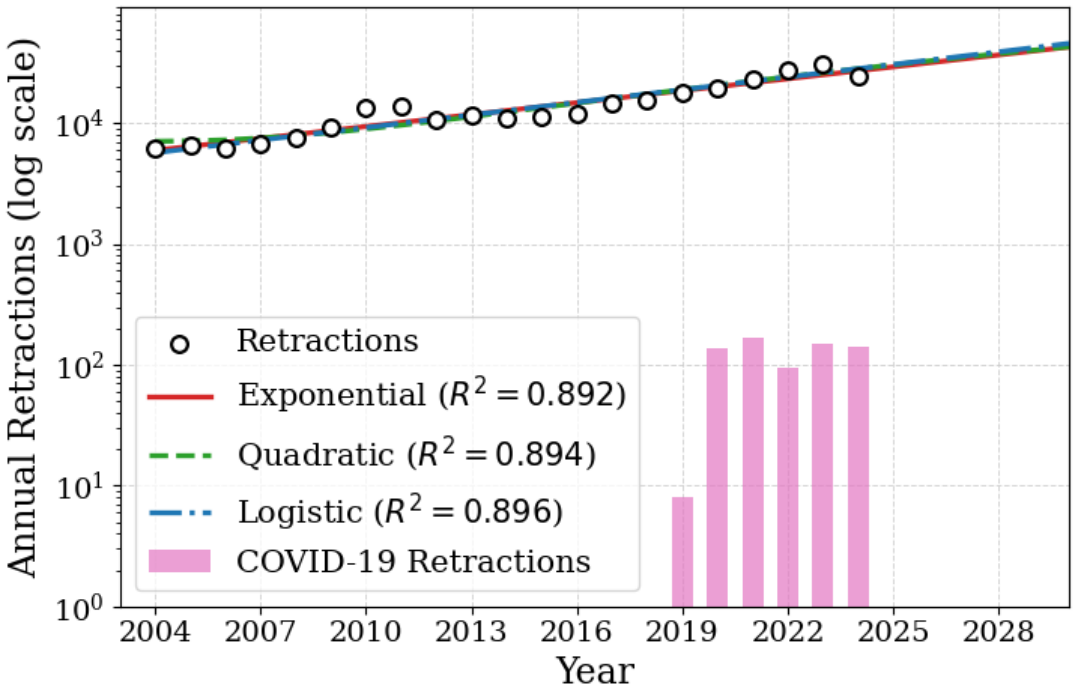

图 2: 撤稿（2004–2024，圆圈）与 COVID-19 相关的撤稿（条形图，2019–2024）。模型拟合和未来预测（至 2030 年）分别展示了指数函数、二次函数和逻辑斯蒂函数的结果。

讨论

表 2 概述了所有三个模型的紧密拟合以及系统的惯性翻倍时间。事实证明，尽管全球经历了重大干扰，但 20 多年来撤稿、发表文章和预印本都与逻辑斯蒂、二次方和指数模型相匹配，这是深层次系统惯性的有力证据。COVID-19 冲击——在总体数据中几乎看不见——表明偶发事件本身不足以改变长期趋势。

这种惯性行为既是机遇也是警告。一方面，这意味着研究生态系统是稳健的：即使产出大幅增加或出现临时错误，也不会立即破坏其完整性的机制。另一方面，这种惯性也预示着风险：随着生成式 AI 进一步加速出版进程，纠错和撤稿程序可能会继续滞后，导致未被发现的错误或不当行为不断增加，除非采取积极干预措施。

多个回归模型的紧密匹配也表明，至少在下一个周期内，对未来趋势（无论是增长还是危机）的预测应被视为概率性的“惯性锥”：除非出现真正的结构性创新，否则当前体制将简单地继续向前发展，仅受其自身倍增时间的限制。由 AI 驱动的写作、评审和错误检测的兴起可能是新过渡的催化剂，但截至 2024 年，这仍然是一个前景情景。

出版物、撤稿和预印本之间的强相关性和统计显著性关联——即使在移除与 COVID-19 相关的撤稿后仍然不变——表明这些学术生态系统维度受共同的结构和系统力量支配。这种“惯性同步”反映在不同回归模型（逻辑、二次和指数）在整个历史时期达到近

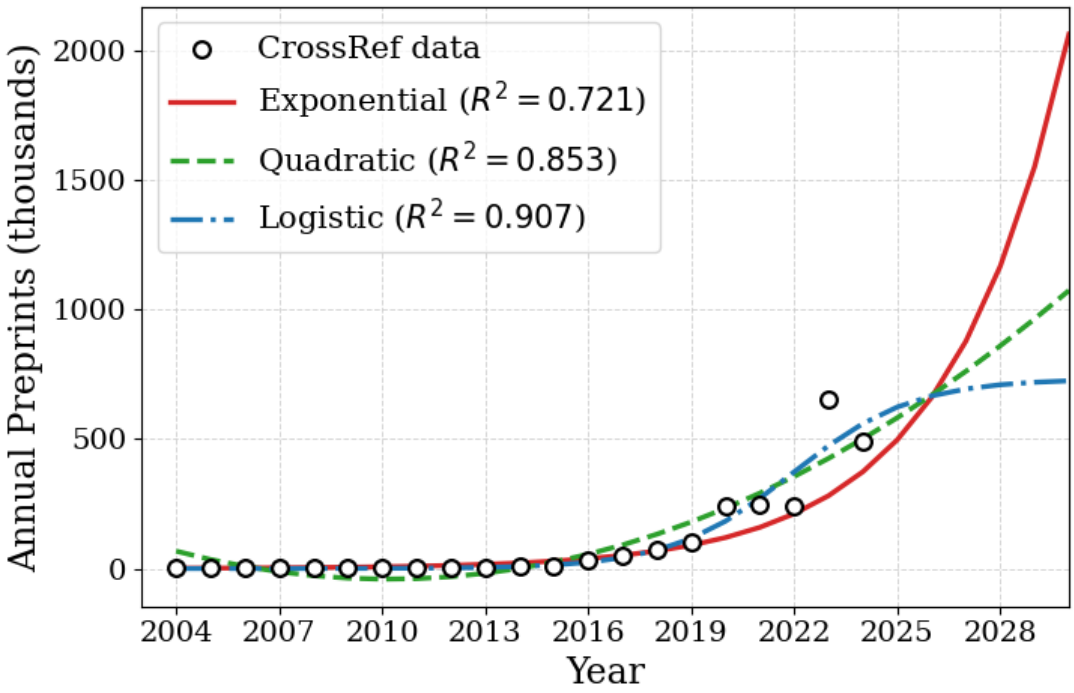

图 3: 年度交叉引用索引预印本（2004–2024，圆圈）以及拟合的指数、二次和逻辑斯谛模型及向前预测至 2030。

乎等效的 R^2 值上，并且进一步得到这样的发现支持，即即使是像 COVID-19 大流行这样重大的冲击也只是产生了短暂的局部影响，而没有从根本上重新排列时间序列关系。

这些结果加强了将撤稿增长主要解释为一种惯性、系统驱动现象的观点——与产出的扩展紧密相关，但受到其自身的滞后和平台动态影响。这种同步也突显了“休克疗法”干预措施的局限性，例如在疫情期间旨在加快科学研究以满足监管目标的那些措施，未能扰乱整体轨迹或改变基础校正率。因此，观察到的倍增时间和平台预测不应仅被视为统计预测，而应视为科学系统内深层次嵌入、相互强化反馈持续存在的证据。

展望未来，这种同步性及其对外部干扰的抵抗力表明，要提高研究诚信和错误纠正方面的实质性改进，需要采取系统性的干预措施，能够改变整个学术基础设施的“惯性”轨迹，而不是对周期性危机的孤立反应。

科学治理应专注于在时间序列数据中出现之前识别拐点，主动弥补输出与质量控制之间的滞后，并为可能由人工智能或政策创新触发的体制转变做好准备。直到那时，惯性仍是主要法则——无论是进展还是错误。

表 1: **统计相关性（Pearson$r$, Spearmanρ, Kendallτ）以及关键 CrossRef 时间序列（2004–2024）之间的 p 值**。“撤回*” 表示撤回的项目减去与 COVID-19 相关的项目。

序列配对	皮尔逊 r	p	斯皮尔曼 ρ	p	肯德尔 τ	p
Publications–Preprints	0.992	<0.001	0.974	<0.001	0.882	<0.001
Publications–Retractions	0.855	<0.001	0.779	<0.001	0.623	<0.001
Publications–Retractions*	0.851	<0.001	0.776	<0.001	0.618	<0.001
Preprints–Retractions	0.742	0.001	0.699	0.002	0.547	0.003

表 2: 年度 CrossRef 出版物、撤稿和预印本系列的倍增时间及模型拟合，2004–2024。

	Publications	Retractions	Preprints
Fitted Model	Logistic	Logistic	Logistic
R^2 (logistic)	0.920	0.896	0.907
Doubling Time (years)	9.8	11.4	5.6

方法

通过 CrossRef REST API 获取数据

本研究中的所有数据均直接从官方 CrossRef REST API（`https://api.crossref.org/works`）编译而来，该 API 提供了在 CrossRef 会员出版商注册的学术成果权威且最新的元数据。为了确保全面覆盖和可重复性，每年都会针对四个类别发出查询：（1）总发表量，（2）撤稿，（3）与 COVID-19 相关的撤稿，以及（4）预印本。对于从 2004 年到 2024 年的每一年，通过在 *自发表日期* 和 *直到出版日期* 参数上使用日期范围过滤器来检索总发表量。撤稿是通过将 *更新类型*：*撤回* 过滤器与特征标题字符串（如 “Retraction”、“Retracted”、“Retraction Notice” 等）相结合提取出来的，而 COVID-19 撤稿还需要包含诸如 “COVID-19”、“SARS-CoV-2” 或 “Coronavirus” 之类的标题术语。预印本则是使用 *类型*：*已发布内容* 过滤器来隔离的。每次查询返回一个 JSON 对象，从中从 *消息*.*总数结果* 字段中提取年度数量。用于从 CrossRef 数据库重现结果的计算机代码可在 Github 上获得（参见代码可用性）。

实现与脚本架构

数据收集通过开源的 *交叉引用获取器* Python 脚本（可在 GitHub 上获取，参见代码和数据可用性）实现了自动化。该脚本通过 `urllib3.util.Retry` 建立了一个持久的 `requests.Session`，并采用强大的重试逻辑来缓解服务器错误和速率限制问题，同时记录所有时间及失败事件以最大化完整性和可审计性。对于每种数据类型和年份，脚本会组装适当的 REST API 查询参数并发出请求，在必要时自动重试。所有输出按年份索

引，并保存到 CSV 文件中。

数据输出结构

主要输出是一个时间序列 CSV 文件（分析时间序列.csv），包含每个类别的逐年计数：出版物、撤回、COVID_撤稿和预印本。额外的输出包括回归拟合详细.csv，其中包含时间序列中所有回归模型的拟合值，以及回归汇总.csv，它汇总了每个拟合的最佳参数值、R^2 得分和倍增时间。相关性分析结果存储在相关性总结.csv 中，报告了所有成对比较的 Pearson、Spearman 和 Kendall 系数及关联的 p 值。

回归模型与统计分析

每个时间序列都拟合了三种典型的生长模型：指数型、二次型和逻辑斯谛型。指数模型遵循公式 $y = ae^{bx}$，二次模型为 $y = ax^2 + bx + c$，逻辑斯谛模型为 $y = \frac{K}{1+e^{-b(x-x_0)}}$。非线性拟合使用了 SciPy 的曲线拟合和线性回归分析程序，并计算了每个拟合的判定系数 (R^2)。对于每个序列，从最佳拟合参数中分析计算了倍增时间 T_2。

模型选择基于比较 R^2 和视觉合理性，特别是关于向前预测（至 2030 年）。所有回归参数、拟合度和 R^2 值都报告在回归摘要.csv 中，并在结果部分进行了讨论。

相关性分析

为了评估出版、撤稿和预印本趋势之间的相互依赖性，使用了三种方法计算成对相关性：皮尔逊的 r、斯皮尔曼等级相关以及肯德尔的 τ。这些方法被应用于原始年度序列以及去除了与 COVID-19 相关的撤稿后的序列中，以区分惯性耦合和偶发效应。得到的相关系数及其显著水平在相关性汇总.csv 中报告，并构成了本文讨论的系统级同步解释的基础。

所有脚本、原始数据、拟合结果和分析程序均公开可用且完全可重复。

限制条件

虽然这项研究利用了最大的开放访问元数据语料库（CrossRef）并使用了多种回归模型，但仍有一些限制需要提及。首先，不同出版商和年份的元数据完整性存在差异，特别是对于较早的撤稿和预印本而言。其次，依赖基于标题的过滤器（例如，“撤稿通知”）可能会遗漏非标准标签的项目。第三，逻辑模型假设了能力上限，但没有直接证据表明系统已经饱和。最后，虽然模型预测在 2024 年后会有所不同，这些预测应谨慎解读，并需通过 2025 年以后的实际数据进行验证。

代码和数据可用性

所有代码（包括完整的 API 查询、分析脚本和绘图程序），以及基础的时间序列和回归结果，均可在 `https://github.com/khalid-saqr/CrossRef` 公开获取。该仓库包含文件 `analysis_time_series.csv` 和 `回归拟合详细.csv`，以实现完全可重复性。

参考文献

[1] M. Davidson, T. Evrenoglou, C. Graña, A. Chaimani, and I. Boutron. Comparison of effect estimates between preprints and peer-reviewed journal articles of covid-19 trials. *BMC Medical Research Methodology*, 24(1), 2024. doi: 10.1186/s12874-023-02136-8.

[2] F. Fernandez-Llimos. Peer review and publication delay. *Pharmacy Practice*, 17(1), 2019. doi: 10.18549/PharmPract.2019.1.1502.

[3] D. E. Parrish. From the editor—a peer review crisis or new normal? *Journal of Social Work Education*, 58(4):619 – 621, 2022. doi: 10.1080/10437797.2022.2138070.

[4] D. E. Parrish. The peer review crisis continues: What comes next? *Journal of Social Work Education*, 60(2):171 – 173, 2024. doi: 10.1080/10437797.2024.2363167.

[5] C. Tropini, B. Brett Finlay, M. Nichter, M. K. Melby, J. L. Metcalf, M. G. Dominguez-Bello, L. Zhao, M. J. McFall-Ngai, N. Geva-Zatorsky, K. R. Amato, E. A. Undurraga, H. N. Poinar, and J. A. Gilbert. Time to rethink academic publishing: the peer reviewer crisis. *mBio*, 14(6), 2023. doi: 10.1128/mbio.01091-23.

[6] H. Tsunoda, Y. Sun, M. Nishizawa, X. Liu, K. Amano, and R. Kominami. The impact of preprints on the citations of journal articles related to covid-19. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 15494 LNCS:57 – 69, 2025. doi: 10.1007/978-981-96-0868-3_4.

[7] M. Whitaker, S. Rodrigues, G. Cooke, B. Virlon, C. A. Donnelly, H. Ward, P. Elliott, and M. Chadeau-Hyam. How covid-19 affected academic publishing: a 3-year study of 17 million research papers. *International Journal of Epidemiology*, 54(3), 2025. doi: 10.1093/ije/dyaf058.