别盲从领头者：独立思考者推动科学创新

Sean Kelty，[1] Raiyan Abdul Baten，[2] Adiba Mahbub Proma，[2]

Ehsan Hoque，[2] Johan Bollen，[3] 以及 Gourab Ghoshal[1,2,*]

[1]物理与天文学系，

美国纽约州罗切斯特市罗切斯特大学，邮编 *14607*

[2]美国纽约州罗切斯特市罗切斯特大学计算机科学系，邮编 *14607*[3]*Luddy*信息学、计算与工程学院，东第十街*919*号，印第安纳州布卢明顿市，邮编 *47408*

arXiv:2301.02396v1 [cs.DL] 6 Jan 2023

摘要

学术成功的分布并不均衡；少数顶尖科学家获得了大部分关注、引文和资源。然而，这些“超级明星”是否推动了科学创新领域的领导力？我们提出三种信息论度量，用于量化学术引用网络中的新颖性、创新和影响力，并比较与超级明星科学家毫无关联或联系紧密的科学家的学术产出。我们发现，虽然与超级明星有联系的科学家确实发表更多论文、获得更多引文并产出更多样化的内容，但代价是创新性较低、思想冗余度较高。此外，剔除与超级明星合著的论文后，这些有联系的科学家的学术产出会减少。相比之下，不借助与科学界超级明星合作而产出创新性内容的作者，发表的论文能够连接更多样化的概念，发表数量也更多；在控制超级明星转移的声望后，其引用率也不逊色。总体而言，我们的研究结果表明，学术界将关注和资源集中于超级明星是要付出代价的。

* 联系邮箱：gghoshal@pas.rochester.edu

一、引言

“要从零开始真正做出一个苹果派，你必须先创造宇宙”——这句被归于卡尔·萨根名下的话 [1]——阐明了这样一种观点：个体的创造过程取决于其创作所依托的要素。无论是创作一段新音乐、处理日常事务，还是从事科学研究，人们的行动都建立在自身及通过他人所建立的信息、经历和关系之上 [2–5]。每个人都有源自自身生活经历的知识基础，同时也处于一个关系网络之中；人们在其中彼此分享经历和知识，从而在相互关联的个体网络中形成集体认知 [6]。在此类网络中，可能会出现等级结构：一些行动者对网络施加更大的社会影响力，进而影响网络所支撑的创造过程；另一些行动者则可能只影响与自己最亲近的人，或根本影响不到任何人 [7]。这种社会等级结构常见于政府和政治的社会动态之中，在这些领域，一些个人和机构对系统中的信息流动和意见形成施加着很大影响力 [8–10]。

学术界并不能免于社会等级的形成；一些学者凭借其工作所体现的学术价值和影响力，以及他们在学术合作网络中的突出地位，能够成为权威人物。作为学术影响力指标的引文 [11]长期以来一直被认为分布极不均衡[12], ，少数科学家获得了大部分引文。这种不平等在全球范围内可能正在加剧[13], ，至少就引文数量而言是如此。在学术出版领域，这类偏向效应一直从马太效应的角度加以研究：成功带来更多成功，早期成功则会累积成一种累积优势，正如“富者愈富”[14]。有人认为，这种效应有益于学术界；顶尖研究人员获得的回报与其贡献成正比，从而保障该领域的“认识论安全” [15]。这种观点与科学应当实行精英主义的理念相符：贡献最大的人也最受重视，因此影响力也最大。事实上，人们非常信任最成功的学者，也看重他们的指导。例如，在职业生涯早期与顶尖科学家合作的初级研究人员，很可能自己也会成为高被引科学家，

尤其是在声望较低的大学任职者 [16]。经验不足的学者可以通过跟随顶尖科学家接受学徒制培养而受益；对早期职业阶段科学家的“引导”会提高他们在高影响力期刊上发表论文的比率 [17]。这些关系往往互惠互利。知名度较低的作者有更多机会在能够吸引更广泛读者的高质量期刊上发表论文，而顶尖科学家则能获得具备独特技能的合作者，从而产出更多高质量作品 [18]。知名度较低的学者与上层学者密切合作，还可能创造先发优势，形成正反馈循环，并促使创新思想在早期出现从众效应 [19]。

尽管顶尖学者（有时被称为“超级明星”）可能持续做出高影响力的贡献，从而使其所在领域和合作者受益，但由于随后出现的资源与关注的集中，他们作为超级明星的地位也可能产生有害影响。例如，研究表明，在某位超级明星突然去世后，学术明星的合作者的发表率会下降 5% 至 9%， [20],这凸显了他们对该超级明星合作的依赖。事实上，目前尚不清楚与超级明星合作是否真正促进了独立的职业发展 [21, 22]此外，由于资助中的马太效应，超级明星可能导致科研经费分配高度不平等。那些获得资助的人，其后累积的科研经费是另一类人的两倍；后者提交了价值相近的提案，却恰好因偶然因素略低于资助门槛。没有证据表明，这种科研经费的累积源于先前资助所促成的实际成就 [23, 24]。如果与超级明星的成功合作能够带来早期资助方面的成功，就可能形成由超级明星推动的资助循环，使科学界“富者”与“贫者”之间的差距不断扩大。

科学合作网络的拓扑结构、结构和特征可能在这些效应中发挥重要作用，因为它们会影响思想的产生与传播，并可能带来相互冲突的结果。紧密连接的网络能够更有效地传播和利用知识，从而提高生产力，但同时也可能导致多样性下降，减少探索与发现 [25–27]。尽管可能会出现一些溢出效应，即备受赞誉的作者的合作者会间接受益 [28], ，但尚不清楚将关注和资源集中于

超级明星是否会带来更多新颖且富有创新性的研究。随着跨学科研究的兴起，这一问题尤为重要，因为跨学科研究依赖科学家能够在公平的团队中开展合作，从而推动不同研究领域的创造力与创新 [29]。

为探究超级明星影响力对学术生产力、影响力和创新的影响力，我们对美国物理学会语料库进行了全面分析。继 [20], 我们将超级明星定义为h指数位居前0.1%的学者 [30, 31]。我们提取了超过25万份摘要的语义内容，并定义了一系列信息论度量，以量化每篇论文的新颖性和创新性。在此基础上，我们进一步分析出版物和引用率，并考察与超级明星合作或频繁引用超级明星论文的研究人员，与几乎不与此类超级明星产生联系的研究人员之间的学术产出差异。我们发现，在个体层面，与超级明星合作以及经常引用其论文的研究人员，相较于其他学者发表更多论文、获得更高引用量，并产出内容更加多样化的论文。然而，他们的工作并不比语料库中的其他论文更具创新性，且其内容更加冗余。此外，排除与超级明星共同署名的论文后，他们的出版物和引用产出与语料库中的其他研究人员并无差异，在某些情况下产出反而更低。

聚焦于职业生涯早期研究人员，我们发现，那些在职业生涯初期频繁与超级明星合作的人，最终确实会取得出色的学术产出；然而，一旦不再合作，他们在出版率、引用影响力和创新方面的产出就会显著下降。另一方面，职业生涯早期研究人员若不依靠早期与超级明星合作便能产出创新性内容，那么在职业生涯的其余阶段也会持续产出此类内容。控制合作本身的影响后，他们发表的论文比超级明星的早期合作者更多，获得的引用次数则相近。

II. 结果

A. 数据

我们使用美国物理学会（APS）语料库 [32]，其中收录了自1893年以来发表在APS 期刊上的文章。该数据集包含完整的引文数据，即从一篇文章的参考文献指向另一篇文章的引文，因此可以重建所有文章之间的完整引文网络。数据还包括文章特有的字段，例如数字对象唯一标识符、期刊、卷、期、首页和末页，或文章编号及页数、标题、作者、所属机构、出版历史、PACS代码、目录标题、文章类型和版权信息。由于数据不包含文章摘要，我们使用网页抓取算法 [33]，收集了250,628篇文章的摘要，这些文章约占APS 各期刊已发表论文总数的35%至40%（图 S1）。值得注意的是，这些文章中约有1%的参考文献未收录在 APS 引文网络中；平均而言，我们抓取了38%的论文参考文献所对应的摘要。引文分布和 h指数分布都呈重尾特征（图 S2），平均引用次数为14.4，平均h指数为1.74。我们使用基于规则的评分方法 [34] 进行作者消歧（参见第 S1.2 节）。我们考察了1970年及以后首次发表论文的作者，并将语料库中 h指数排名前0.1%的作者定义为超级明星，对应的 h指数阈值为21。由此，我们在292,394位作者中确定了303位超级明星。描述性统计见表 S1。

为了从收集到的摘要中提取主题，我们对短语采用无监督的潜在狄利克雷分配（LDA）算法（P-LDA） [35]，为语料库中的短语和文档建立向量嵌入。语料库中的停用词已被移除，所有词语均经过词形还原；短语则根据显著性得分确定，该得分用于判断短语的出现是否由随机因素造成。这些向量嵌入的维度为 k，对应于语料库中定义的主题数量。P-LDA 利用吉布斯采样生成短语和文档上的主题分布 [36],，并据此根据主题扩散提取新颖性得分。我们根据 UMass 一致性度量（[37]）选择主题数量 k；该度量的值在主题数量达到 $k = 25$ 时首次趋于稳定（图 S3）。表 S2 展示了每个主题的前 10 个术语。每篇文档 u 的输出是一个 k 维向量$\mathbf{v}^u$，其元素 v_i^u

对应于 o t从其摘要中提取的主题 i 的频率（示例见表 S3）。

B. 新颖性、创新与冗余

文献中的新颖性检测已通过多种方式实现 [38], 例如，将机器学习中的新颖性置于信息检索的语境中 [39, 40], 通过引文关系识别思想的远距离组合 [41], 首次组合此前从未关联的概念 [42], 社交网络中概念的知识图谱 [26], 以及社会与个体学习的基于智能体的模拟 [27]。

在此，我们依赖文档级嵌入来表示给定论文摘要中所包含的所有主题的分布，并据此可以从论文、其参考文献以及引用该论文的文章等层面定义主题多样性。基于此，我们定义了多种指标，用于捕捉新颖性与创新的不同方面。

随后，将作者之间的联系与其作品的内容结合起来，可以阐明超级明星对其他学者所取得的成功以及产生的新颖性的影响。熵： 对于给定文档 u，我们将香农熵定义为

$$I_u^{(S)} = -\sum_{i=1}^{k} v_i^u \ln v_i^u, \tag{1}$$

该表达式量化了随机变量 [43] 各种结果的平均“意外程度”或不确定性。在此语境下，摘要仅聚焦于少数主题的论文，其 $I_u^{(S)}$ 值较低；而涵盖主题多样性较广的论文，其熵值则较高。

参考文献与引文多样性：虽然 $I_u^{(S)}$衡量的是相对于论文内容（此处指其摘要）的“意外程度”，但参考文献和引文所指的是某篇论文中的观点受到其他论文启发（参考文献），或启发其他论文（引文）的程度。因此，我们可以通过评估该论文参考文献或其收到的引文所涉及主题的离散程度，来衡量论文的新颖性，或其信息多样性[44], 。主题分布的方差越大，信息多样性就越高。对于集合 X^u，它可以代表论文 u中的参考文献，也可以代表引用论文 u的引文，我们

定义该量：

$$I_u^{(X)} = \frac{1}{|X^u|} \sum_{l \in X^u} \left[1 - \cos\left(\mathbf{v}^l, \overline{X^u}\right)\right] \tag{2}$$

其中，$\cos\left(\mathbf{v}^l, \overline{X^u}\right)$ 表示特定参考文献/引文 $\mathbf{v}^l$ 的向量嵌入与集合 X^u中所有参考文献/引文的向量嵌入均值之间的余弦相似度。因此，我们可以将来自一篇论文的参考文献以及引用该论文的引文中的信息多样性分别定义为参考文献多样性和引文多样性。

创新： 截至目前定义的指标，基于主题模型；这些主题模型根据文档内容（摘要）中的词语，表示为每个文档的主题分布。这些指标能够捕捉论文自身或其影响的主题多样性，但无法表达论文通过创新扩展文献的程度。换言之，它们表达了文档本身的主题，却没有说明这些内容是否增加了文献的多样性。因此，我们将创新定义为文档以新的组合方式向文献中增加主题的程度 [45, 46]。具体而言，在此语境下，创新度用于衡量某些术语首次在语料库中被引入或组合的情况（参见第 S1.4 节和图 S4）。结合新颖性度量，我们可以追踪思想多样性与术语的新概念重组及共现之间的相关性。遵循这一逻辑，我们将论文 u 的创新度定义为

$$I_u^{(I)} = \frac{1}{2} \sum_{w_1 \neq w_2 \in u} \mathcal{I}(w_1, w_2; u) \tag{3}$$

其中，w_1和 w_2是论文 u中的不同术语，$\mathcal{I}(w_1, w_2; u)$ 是一个指示函数：如果术语 w_1 和 w_2 首次在论文 u 中出现于语料库，则其值为 1；否则为 0，而 1 2前缀用于避免重复 _ 计数。为排除偶然或极为罕见的情况造成的虚假概念联系，我们计算所有联系的点互信息，即共现概率与各个概念单独出现概率之比的对数 [46]。在图 S5 中，我们计算了各项指标之间的皮尔逊 r 相关系数，发现它们之间的相关性很弱，这表明每项指标都反映了学术产出的不同方面。

冗余： 最后，在一个相关背景下，有研究报告称，在创意构思领域，受到极具创意的互动对象启发的受启发者，往往会产生更多创意 [47–49]。然而，从组的层面来看，受启发者的想法被发现彼此相似

随着时间推移，导致组层面生成的想法出现冗余。为检验学术出版中是否会出现类似效应，我们计算了论文 u， u' 在集合 $P(G,\ s,\ t)$ 中的余弦相似度得分，因此

$$\mathrm{Sim}(G,s,t) = \frac{2}{|P(G,s,t)|\,(|P(G,s,t)|-1)} \sum_{u,u' \in P(G,s,t)} \cos(\mathbf{v}^u, \mathbf{v}^{u'}). \tag{4}$$

C. 超级明星统计

接下来，我们考察超级明星所产生的新颖性和创新，与学术语料库中其他成员相比是否存在显著差异。在图 1 中，我们绘制了超级明星学者群体与语料库中其他作者群体的参考文献与引文多样性（公式 (2)）、香农熵（公式 (1)）和创新（公式 (3)）的比较结果。在参考文献多样性、引文多样性和香农熵方面，超级明星分别比其他学者高出 20%、15% 和 2%。换言之，与非超级明星相比，超级明星受到更多样化内容的启发，发表的作品在概念上更加多样，也启发了种类更广的出版物。最显著的差异体现在创新方面：超级明星与其他学者之间相差十倍，这表明前者在引入新的术语组合方面产出更为丰富。我们注意到，对所有学者而言，这些指标均随出版物数量单调变化，不过这种效应在超级明星中更为明显（图 S6）。此外，论文 u 获得的引文数量也与新颖性/创新指标呈单调关系（同样在超级明星中更为明显），这表明概念多样性的提升以及首次建立概念联系的能力，会因该论文获得更多关注而得到回报（图 S7）。

D. 超级明星影响力

在确认超级明星在我们的各项指标上优于其他学者后，我们接下来考察超级明星在多大程度上影响其合作者和“受启发者”（即受到其作品启发的人）的学术产出。受启发者是引用超级明星论文的作者，我们根据引文频次来衡量他们受到启发的程度

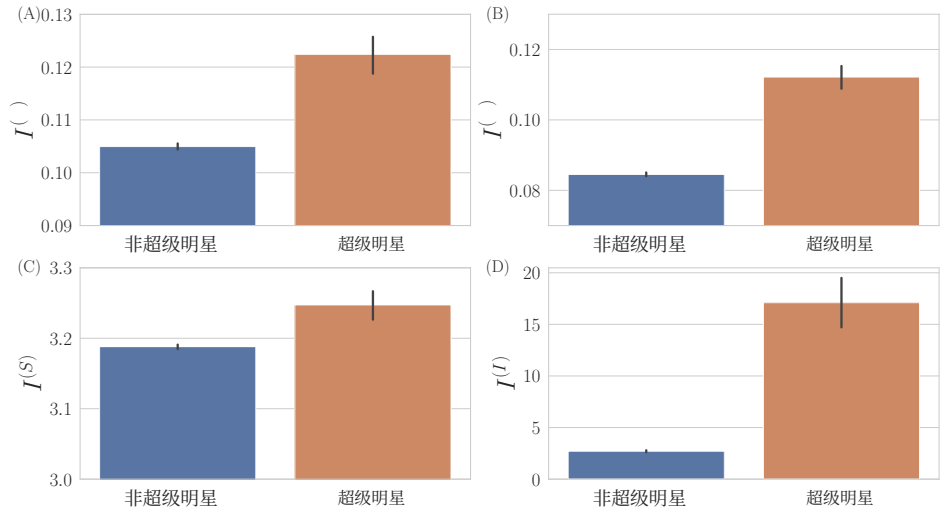

图 1. **新颖性和创新的作者层面的平均统计数据 A** 参考文献多样性，**B**引文多样性，**C** 香农熵，**D** 创新。橙色柱表示超级明星（h指数≥ 21），蓝色柱表示语料库中的所有其他作者。

。我们从组层面和个体层面考察受启发者。在组层面，我们以超级明星为中心构建受启发者网络，其中受启发程度指研究人员引用该超级明星的次数。随后，我们根据受启发程度将受启发者分组，各分档的上限分别为受启发者前 10%、20%、30%、50% 和 100%。这些组与某位超级明星的联系逐渐减弱；前 10% 的受启发者受到的启发最为显著，而后 50% 的受启发者通常只引用该超级明星一次。请注意，某位超级明星后 50% 组中的一些受启发者，可能同时属于另一位超级明星的前 10% 组。逐步增大的分档规模是为了考虑到受启发程度最低的受启发者中，受启发引用频次不断下降的情况，从而确保各组之间有足够数量的论文可供比较。

鉴于我们关注超级明星影响力对受启发者新颖性与创新的时间演变，我们将超级明星首次发表论文的年份记为 $t_0 = 0$，并针对后续每一年 $t > t_0$，考察引用该超级明星的受启发者所发表的论文集合。对于每个分组，我们计算各分组在年份 t 发表的所有论文的平均新颖性。将受超级明星启发的论文集合记为

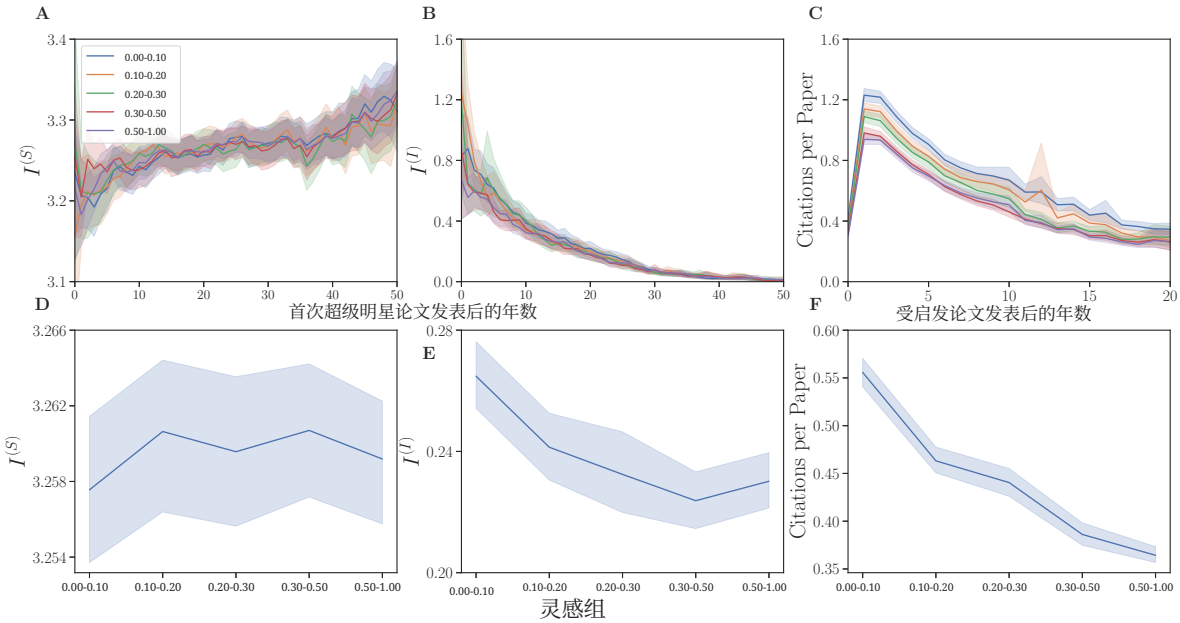

图2. **群体层面的新颖性与创新统计数据**论文层面平均统计数据的时间轨迹。**A**：香农熵，**B**：创新，**C**：单篇论文引用数。汇总的群体层面统计数据：**D**：香农熵，**E**：创新，**F**：单篇论文引用数。曲线表示平均值，阴影区域表示95%置信区间。

s 针对参与者 tion G 在年份 t 的集合记为 $P(G,\ s,\ t)$，平均新颖性得分的计算方式如下

$$\langle I_u^{(l)} \rangle_{G,s,t} = \frac{1}{|P(G,s,t)|} \sum_{u \in P(G,s,t)} I_u^{(l)} \tag{5}$$

其中，$l = S,\ X,\ I$ 是论文 u 的新颖性或创新得分。

我们在图2中绘制了分析结果。从香农熵的时间演变来看，尽管熵呈单调上升趋势，反映出知识体系随时间不断扩展（图 S8），但如图2**A**所示，各组之间几乎没有差异。对整个时间范围取平均后，趋势也同样平缓（图2**D**）。参考文献多样性的时间演变（图 S9**A**,**B**上图）及其时间平均值（下图）也呈现出类似趋势。与熵或参考文献多样性不同，引文多样性随时间呈下降趋势。前 10% 群体的这一指标比后 50% 群体低 5%。图2**B**,**E**显示，创新也呈现相同趋势：所有组的创新都随时间下降，这反映出作者在职业生涯中不断组合的新术语组合数量趋于饱和

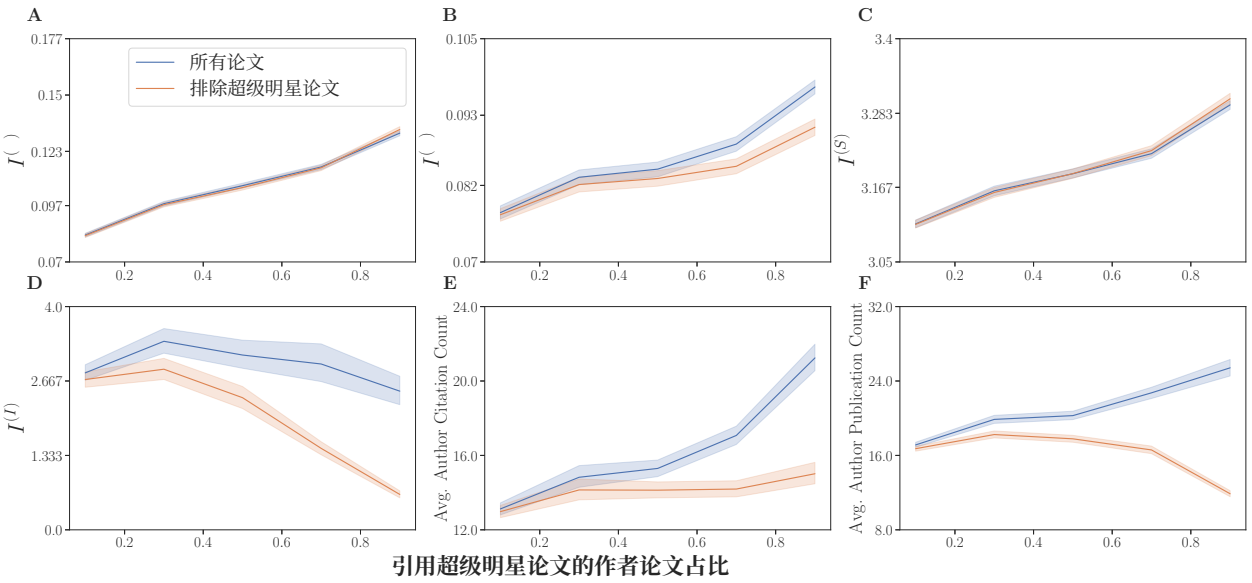

图 3. **作者个体层面的新颖性与创新统计数据. A**参考文献多样性，**B** 引文多样性，**C** 香农熵，**D** 创新，**E** 平均引文数量，**F**平均出版物数量。

。随着作者职业生涯的推进，他们组合这些术语。前 10% 与后 50% 群体之间的差异此时约为 15%。最后，论文的引文数起初有所增加，随后随时间下降，如图2C所示，此时各组之间的差异也明显得多。事实上，如图2F所示，最受启发组与最少受启发组的单篇论文引用数相差 40%。

就冗余而言，我们在图 S9C 中绘制了余弦相似度（公式 (4)）。如图所示，所有组的相似度均随时间呈下降趋势，但组间存在明显差异：与后 50% 相比，前 10% 发表的论文在内容上的平均相似度高出 8%。综合来看，结果表明，相较于其他作者群体，经常引用超级明星论文的作者群体确实获得了引用优势。然而，他们的成果创新性略高，而新颖性与语料库中的其他部分相当。相比之下，他们的内容比其余作者群体的内容更具冗余性。

接下来，我们将群体层面的结果分解开来，考察超级明星影响力在个体作者层面的程度。图 3 展示了作者的内容新颖性与创新指标的平均值，以及引文和发表数量，并以作者论文中引用超级明星的比例为依据进行比较。鉴于许多作者会与超级明星合作发表

的论文，蓝色曲线表示纳入此类论文后的结果，橙色曲线则表示排除这些论文后的结果。图 3A-C 表明，无论是否纳入作者与超级明星合作的论文，随着作者引用更多超级明星，其参考文献与引文多样性以及香农熵都会增加。虽然我们没有发现群体层面的内容新颖性受到超级明星影响力驱动的迹象，但在个体层面，益处显而易见。另一方面，考察创新（图 3D）时，纳入所有论文后趋势基本持平，而排除合著出版物后则呈下降趋势。事实上，作者引用的超级明星越多，他们自己的出版物（即未与超级明星合著的出版物）似乎就越缺乏创新性。考察引文（图 3E）和出版物数量（图 3F）时，与超级明星合作的益处就更加明显。前者在纳入合作论文时，引文收益显著（平均引文约增加 67%），排除合作论文后则大幅下降。事实上，引文方面的收益似乎主要源于与超级明星合作；按定义，超级明星的论文拥有最多的引文。后者似乎也是如此：纳入合作论文时，出版物数量增加；排除后则减少。

E. 早期合作者与早期创新者

迄今为止的结果表明，受到超级明星启发的学者能够产出内容多样化、并通过引文获得可见度的成果，但从将新概念联系起来的意义上说，这些成果未必具有创新性。另一方面，也有证据表明，与超级明星直接合作会显著提升这些特征；而当这些学者自行开展工作时，他们的出版产出、新颖性和创新性都低于语料库中的其他人。事实上，这不禁引发一个问题：超级明星究竟是在促进独立的个人成功，还是反而在抑制这种成功？例如，如前所示，在总体层面上，最常引用超级明星的作者组往往主要围绕相同的主题发表成果。

为进一步探究这一点，我们将分析范围限定在早期职业阶段科学家。鉴于既往研究的发现表明，与成功的科学家合作会为

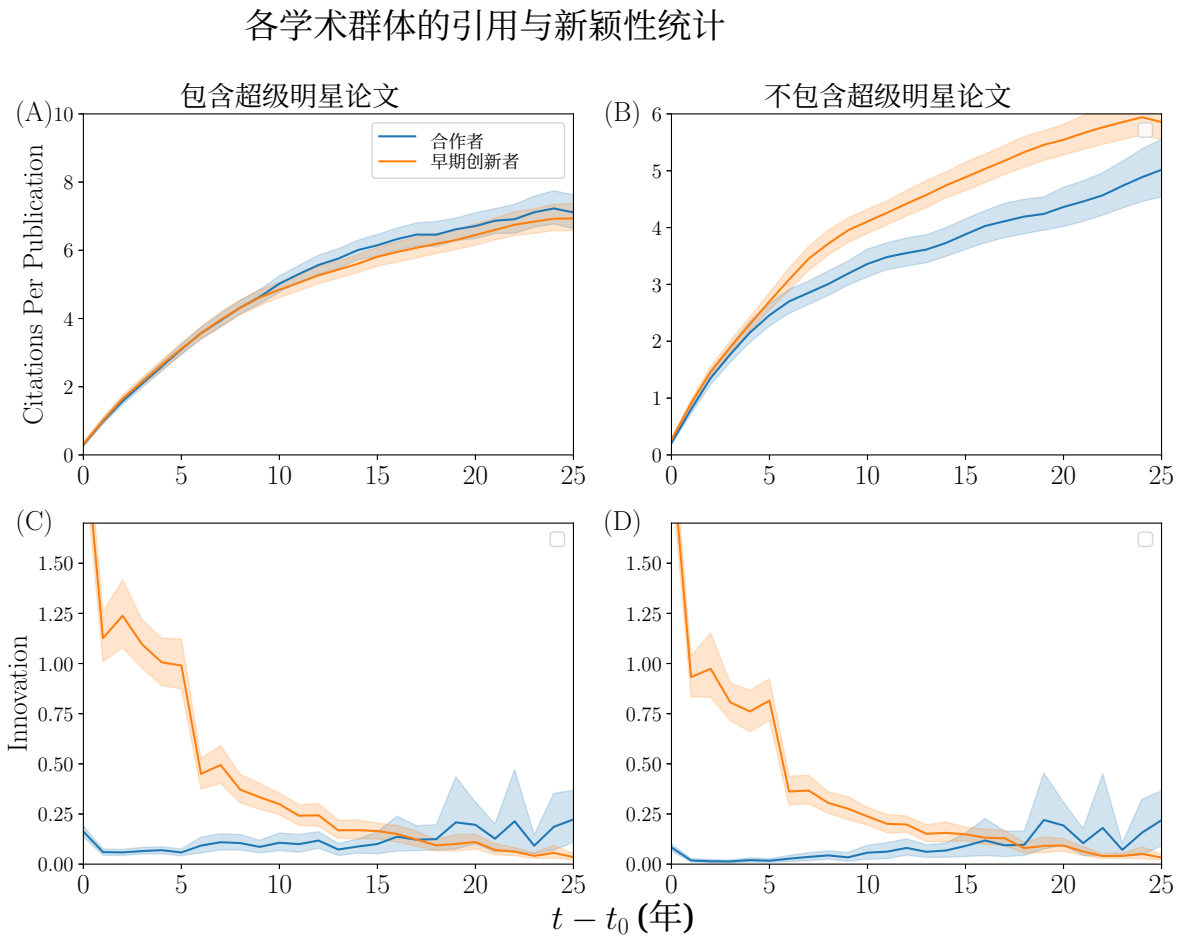

图 4. **频繁合作者和早期创新者的引文与创新 A**纳入超级明星论文时的篇均引用次数，**B** 排除超级明星论文时的相同指标。**C** 创新的时间演变。**D** 排除超级明星论文时的相同指标。横轴 $t - t_0$ 表示两组作者自 t_0 首次发表以来经过的时间。

职业生涯早期研究人员带来助益 [16], ，且早期成功会产生长期职业成功的累积优势 [14], ，我们将早期合作者 定义为：在职业生涯的前五年中，至少有一半论文与超级明星合作的作者。作为比较，我们还定义了另一组作者：他们不与超级明星合作，也不引用超级明星论文，但根据其职业生涯前五年的出版物所衡量的创新，在语料库中位列前 10%。我们将这些作者称为早期创新者。我们使用创新作为指标，是因为超级明星在这一指标上的表现超越其他学者最多（图 1D），因此创新可能是学术潜力的可靠指标。

对于每个组中的学者，我们追踪其篇均引用次数、出版物数量以及创新的时间演变；创新以首次出版物日期为基准进行测量 t_0 ，适用于任一组的作者。与早期创新者相比，早期合作者获得的篇均引用次数更高（图 **4A**），发表的出版物也更多（图 **S10A**），尤其是在职业生涯的前十年中。然而，剔除明星学者论文后，这一趋势发生逆转：早期创新者发表的出版物更多（图 **S10B**），其获得引文的速率也与另一组相当（图 **4B** ）。此外，与早期合作者相比，早期创新者在整个职业生涯中都保持着更高程度的创新（图 **4C,D**），无论是否将与超级明星的合作纳入其中。因此，证据表明，尽管职业生涯早期的科学家确实能从与超级明星合作中获益，但与其他在没有此类合作助力的情况下产出创新成果的职业早期科学家相比，他们自身的学术产出创新性较低，而在引文方面同样具有可见度。

III. 结论与讨论

在学术界知识库呈指数级增长、可见度和资助日益向顶尖学者与机构倾斜的背景下，我们考察了超级明星学者对整个学术共同体的影响，以及他们在新颖性和职业成功方面的表现。超级明星学者提供了不可替代的新颖思想与贡献来源，其产出率超过语料库中的其他学者；我们的指标表明，他们获得的荣誉实至名归，应当得到相应奖励。我们发现，超级明星学者具有很高的新颖性，并能激发追随者和合作者产生更高的概念多样性。然而，他们确实会抑制创新潜力。最受某位超级明星学者启发的学者，其个人论文的多样性更高；但在组层面，与受该超级明星学者启发较少的组相比，他们对内在新颖性的贡献较小，尽管获得了更多引文。

此外，我们发现了马太效应显著存在的迹象：尽管在概念多样性方面的收益高于其他学者，但高度引用超级明星的学者与其合作时获得的引文，比不与其合作时更多。虽然与成功学者合作能够助力职业道路顺利发展，但我们发现，这些合作

可能会抑制创新，也未必能最准确地反映长期的独立职业成功。

合作是应对日益复杂的跨学科问题的必要条件。超级明星在推动跨学科研究方面处于有利地位，可以为职业生涯早期的研究人员提供支持。尽管后者与超级明星合作时会获得引用优势，但这并不意味着他们比那些与顶尖学者联系较少的同事开展了更具新颖性的研究。事实上，我们的结果表明，最接近超级明星的人创新潜力最低。鉴于文献曾指出，与超级明星合作的初级研究人员更有可能在高质量期刊上发表文章，也更有机会与其他顶尖科学家开展高质量研究，这一发现多少令人意外。然而，综合来看，我们发现这种合作并不能促进长期的独立职业成功。这可能表明，有些人会在超级明星的光环下被淹没，也就是说，这些研究人员会借助各自超级明星的观点和可见度来“跟风”，并在超级明星的工作基础上反复迭代。尽管在既有研究问题的基础上继续推进具有价值，但这或许无法促成创新性工作，也无法推动个人职业发展。事实上，最近的研究显示，科学出版物和专利中的颠覆性观点都在减少 [50]。作者将此归因于研究人员日益依赖范围更窄的现有科学知识来构思新观点；这一发现与我们的观察十分一致：作为一个组，超级明星的追随者产出的内容更加冗余，创新性也更低。

观察到的效应可能是超级明星牢牢掌控各自领域的结果。已有研究表明，超级明星突然离世后，思想领域会发生范式转变。超级明星的合作者在超级明星离世后，发表率会下降；而该领域可能会涌现大量由局外人作出的贡献，其中被高度引用的可能性远高于一般情况 [51]。由此可以推断，超级明星的合作者之所以成功，是因为他们与超级明星合作。在考虑为这些门生提供资助或进行学术招聘时，应谨慎评估。如果目标是促进高度新颖的研究，就应考虑声望和社会关系之外的因素，例如效能、公平与创新。

我们的研究发现并不局限于早期创新者、合作者和受启发者。尽管我们以早期创新者为例，但也可以用我们在此采用的方法分离并研究许多其他组 [52] ，

从而根据新颖性的早期迹象或一系列社会参数，识别有潜力的学者。我们在引言中概述了新颖性的多种不同定义，但本研究并未进一步探讨这些定义。运用不同定义并区分不同类型的新颖性，有助于阐明不同社会配置会抑制或促进哪些类型的新颖性。

我们尚未探讨、但与本文讨论直接相关的一个议题是资助问题。近年来，资助越来越倾向于顶尖机 [53], 构：2018年，美国国家科学基金会的资助中，90%流向了22%的受资助机构；这些机构服务于全部机构的43%和代表性不足少数群体的34% [54]。此外，资助分配在种族和代表性不足的社群方面长期存在差异 [55–57]。另外，代表性不足的群体以更高的比率产出新颖成果，但与性别和种族多数群体的新颖贡献相比，这些成果被其他学者采纳的比率更低 [46]。研究表明，公平资助项目能够加强科研基础设施、提升研究人员的能力，并促进较不知名机构的大学内部及大学间合作[58]。正如我们所展示的，受超级明星影响最小的人创新最多，因此引用率也更高。结合这些结果，并进一步关注公平资助实践 [59]，我们认为这将有助于缩小日益加剧的学术界不平等，并推动新颖且创新的研究。

最后，我们注意到，本研究不可避免地存在局限性。鉴于我们仅关注 APS 的文献体系，因此不应轻易将这些结论外推至其他学科。此外，这也只是整本期刊的不完整子集；如果拥有包含完整引文网络的完整语料库，将能呈现更准确的图景。

---

[1] Cliff, H. 如何从零开始制作苹果派：寻找宇宙的配方（Picador，伦敦，2021）。[2] McAndrew, S. & Everett, M. 音乐作为集体创造：作曲家的社交网络分析。《文化社会学》**9**, 56–80（2014）。网址：https://doi.org/10.1177/ 1749975514542486。[3]Muller, E. & Peres, R. 社会网络结构对创新绩效的影响：综述与研究方向。国际研究与营销期刊

《*Keting*》 **36**，3–19（2019年）。网址：
https://www.sciencedirect.com/science/article/pii/ S0167811618300284。[4] Hazarie, S.、Barbosa, H.、Frank, A.、Menezes, R. 和 Ghoshal, G. 揭示物理移动与虚拟移动之间的差异与相似性。《英国皇家学会界面》**17**，20200250（2020年）。网址：
https://doi.org/10.1098/rsif.2020.0250。[5]Chen, Z. 等。 人类移动中社会性与非社会性可预测性来源的对比。《自然·通讯》 **13**，1922（2022年）。网址：https://doi.org/10.1038/
s41467-022-29592-y。[6] Nathaniel Rodriguez, Y.-Y. A.、Johan Bollen。在认知一致性与社会从众作用下的信念演化集体动力学。《公共科学图书馆·综合》 **11**，e0165910（2016年）。[7] Holme, P. 和 Ghoshal, G. 争夺高中心性与低度的网络代理动力学。《物理评论快报》 **96**，098701–（2006年）。网址：https://link.aps.org/doi/ 10.1103/PhysRevLett.96.098701。[8] Ghoshal, G. 和 Newman, M. E. J. 构建具有任意度分布的分布式网络。《欧洲物理学杂志*B*辑》 **58**，175–184（2007年）。网址：https://doi.org/10. 1140/epjb/e2007-00208-2。[9] Recuero, R.、Zago, G. 和 Soares, F. 运用社会网络分析与社会资本识别推特上极化政治对话中的用户角色。《社交媒体》 + 与社会**5**，2056305119848745（2019年）。网址：
https://doi.org/10.1177/2056305119848745。[10] Dubois, E. 和 Gaffney, D. 影响力的多个层面：识别推特上的政治影响者与意见领袖。《美国行为科学家》 **58**，1260–1277（2014年）。网址：
https://doi.org/10.1177/0002764214527088。[11] Radicchi, F.、Weissman, A. 和 Bollen, J. 量化科学出版物的感知影响力。《信息计量学杂志》 **11**，704–712（2017年）。网址：
https://www.sciencedirect.com/ science/article/pii/S1751157717300846。[12] Hirsch, J. E. 用于量化个人科学研究产出的指数。《美国国家科学院院刊》 **102**，16569–16572（2005年）。[13] Nielsen, M. W. 和 Andersen, J. P. 全球引文不平等日益加剧。《美国国家科学院院刊》 **118**，e2012208118（2021年）。[14] Merton, R. K. 科学领域的马太效应。《科学》 **159**，56–63（1968年）。

[15] Runco, M. & Pritzker, S.《创造力百科全书》。《创造力百科全书》（爱思唯尔科学，2011年）。[16] Li, W., Aste, T., Caccioli, F. & Livan, G. 与顶尖科学家早期共同署名可预测学术职业生涯中的成功。《自然·通讯》 **10**，5170（2019年）。[17] Sekara, V. 等。科学出版中的伴随效应。《美国国家科学院院刊》 **115**，12603–12607（2018年）。[18] Xie, Q., Zhang, X., Kim, G. & Song, M. 探究与顶尖科学家共同署名对研究人员隶属机构、研究主题、生产力和影响力的影响。计量学杂志 **16**，101314（2022年）。网址 https://www.sciencedirect.com/science/article/pii/ S1751157722000669。[19] Abrahamson, E. & Rosenkopf, L. 社会网络对创新扩散程度的影响：一项计算机模拟。《组织科学》 **8**，289–309（1997年）。网址 http://www.jstor. org/stable/2635149。[20] Azoulay, P., Graff Zivin, J. S. & Wang, J. 超级明星消亡。《经济学季刊》 **125**，549–589（2010年）。网址 https://doi.org/10.1162/qjec.2010.125.2.549。 https://academic.oup.com/qje/article-pdf/125/2/549/5319678/125-2-549.pdf。[21] Clauset, A., Arbesman, S. & Larremore, D. B. 教职招聘网络中的系统性不平等与等级结构。《科学进展》 **1**，e1400005 （ 2015 年）。网址 https://doi.org/10.1126/ sciadv.1400005。[22] Janosov, M., Battiston, F. & Sinatra, R. 创意职业生涯中的成功与运气。《EPJ数据科学》 **9**，9（2020年）。网址 https://doi.org/10.1140/epjds/s13688-020-00227-w。[23] Bol, T., de Vaan, M. & van de Rijt, A. 科学资助中的马太效应。《美国国家科学院院刊》 **115**，4887–4890（2018年）。网址 https://www.pnas.org/doi/ abs/10.1073/pnas.1719557115。 https://www.pnas.org/doi/pdf/10.1073/pnas.1719557115。[24] Petersen, A. M., Jung, W.-S., Yang, J.-S. & Stanley, H. E. 通过职业生涯长度研究，对马太效应进行定量和实证验证。《美国国家科学院院刊》 **108**，18–23（2011年）。网址 https://www.pnas.org/doi/abs/10.1073/pnas.1016733108。https://www.pnas.org/doi/pdf/10.1073/pnas.1016733108。[25] Lazer, D. & Friedman, A. 探索与开发的网络结构。管理科学季刊 **52**，667 – 694（2007年）。

[26] Rodan, S. & Galunic, C. 网络结构之外：知识异质性如何影响管理绩效与创新度。战略管理期刊 **25**，541–562（2004年）。网址 http://www.jstor.org/stable/20142143。[27] Chang, M. & Joseph E. Harrington, J. 内生社会网络中的知识发现与扩散。《美国社会学杂志》 **110**，937–976（2005年）。网址 http://www.jstor. org/stable/10.1086/426555。[28] Trapido, D. 知识的新颖性如何赢得认可：一致身份的作用。《研究政策》 **44**，1488–1500（2015年）。网址 https://www.sciencedirect.com/science/ article/pii/S0048733315000839。[29] Xu, F. & Evans, J. 扁平团队推动科学创新。《美国国家科学院院刊》 **119**（2022年）。[30] Hirsch, J. E. h指数是否具有预测能力？《美国国家科学院院刊》 **104**，19193–19198（2007年）。[31] Hirsch, J. E. 用于量化个人科学研究产出的指标。《美国国家科学院院刊》 **102**，16569–16572（2005年）。网址 https://www.pnas.org/doi/ abs/10.1073/pnas.0507655102。 https://www.pnas.org/doi/pdf/10.1073/pnas.0507655102。[32] 美国物理学会。 https://journals.aps.org/datasets。[33]Richardson, L. https://sethc23.github.io/wiki/Python/Beautiful_Soup_Documentation.pdf。[34] Caron, E. & van Eck, N.-J. 使用基于规则的评分与聚类进行大规模作者姓名消歧。载于 Noyons, E.（编） *2014*年科学与技术指标会议论文集，79–86（莱顿大学，2014年）。网址 http://sti2014.cwts.nl。国际科学与技术指标会议，STI 2014；会议日期：2014年9月3日至2014年9月5日。[35] El-Kishky, A., Song, Y., Wang, C., Voss, C. R. & Han, J. 从文本语料库中可扩展地挖掘主题短语。《VLDB基金会会议录》 **8**，305–316（2014年）。网址 https://doi.org/10.14778/ 2735508.2735519。[36] Lee, S. Y. 吉布斯采样器与坐标上升变分推断：集合论综述。《统计学通讯：理论与方法》**51**，1549–1568（2022年）。网址 https: // doi.org/10.1080/03610926.2021.1921214。

[37] Mimno, D., Wallach, H., Talley, E., Leenders, M. & McCallum. 主题模型中语义连贯性的优化。载于《*2011*年自然语言处理实证方法会议论文集》，262–272页（计算语言学协会，苏格兰爱丁堡，英国，2011年）。网址：https://aclanthology.org/D11-1024.[38] Ouafae, B., Oumaima, L., Mariam, R. & Abdelouahid, L. 新颖性检测综述：现状及主要应用领域创新探讨。载于*2020*年第一届应用科学、工程与技术创新研究国际会议（*IRASET*），1–7页（2020年）。[39] Soboroff, I. & Harman, D. TREC 2003新颖性评测赛道概述。载于Voorhees, E. M. & Buckland, L. P.（编）《第十二届文本检索会议论文集：*TREC 2003，2003*年*11*月*18*日至*21*日，美国马里兰州盖瑟斯堡》，第500-255卷，*NIST*特别出版物，38–53页（美国国家标准与技术研究院（NIST），2003年）。网址：http://trec.nist.gov/pubs/trec12/papers/NOVELTY.OVERVIEW.pdf。[40]Ghosal, T., Saikh, T., Biswas, T., Ekbal, A. & Bhattacharyya, P. 新颖性检测：自然语言处理视角。计算语言学**48**，77–117页（2022年）。网址：https://doi.org/10.1162/coli_a_00429。https://direct.mit.edu/coli/article-pdf/48/1/77/2006641/coli a 00429.pdf。[41] Uzzi, B., Mukherjee, S., Stringer, M. & Jones, B. 非典型组合与科学影响力。《科学》**342**，468–472页（2013年）。网址：https://www.science.org/doi/abs/10.1126/science. 1240474。https://www.science.org/doi/pdf/10.1126/science.1240474。[42] Schumpeter, J. A. 《经济发展理论：关于利润、资本、信贷、利息和商业周期的探究（*Theorie der wirtschaftlichen Entwicklung*）》（Transaction，埃迪森，新泽西州，1934年）。Redvers Opie 译。[43] Cover, T. & Thomas, J. A. 《信息论要素》。Wiley 电信与信号处理丛书（Wiley-Interscience，纽约，美国纽约州，2006年）。[44] Aral, S. & Dhillon, P. 网络中的新颖性究竟是什么？剖析经纪人、桥接者和弱关系的愿景优势。运筹学与管理科学研究所（*INFORMS*）（2021年）。网址：http://dx.doi.org/10.2139/ssrn.2388254。https://ssrn.com/abstract=2388254。[45] Kuhn, T. S. 《科学革命的结构》（芝加哥大学出版社，芝加哥，

1962年）。[46] Hofstra, B. 等科学中的多样性与创新悖论。《美国国家科学院院刊》**117**，9284–9291页（2020年）。网址：https://www.pnas.org/doi/abs/10. 1073/pnas.1915378117。https://www.pnas.org/doi/pdf/10.1073/pnas.1915378117。[47] Baten, R. A. 等时间社会网络中的创造力：发散性思维如何受到同伴选择的影响。《英国皇家学会界面》**17**，20200667（2020年）。[48] Baten, R. A., Aslin, R. N., Ghoshal, G. & Hoque, E. 性别和种族认同线索会降低多样化社交网络中的创造力。《科学报告》**11**，10261（2021年）。网址：https://doi.org/10.1038/s41598-021-89498-5。[49] Baten, R. A., Aslin, R. N., Ghoshal, G. & Hoque, M. E. 在社交网络中，新颖创意的生成通过接触“恰到好处”水平的创意变异性而得到优化。*PNAS Nexus* **1**，pgac255（2022年）。[50] Park, M., Leahey, E. & Funk, R. J. 随着时间推移，论文和专利的颠覆性日益减弱。《自然》**613**，138–144页（2023年）。网址：https://doi.org/10.1038/s41586-022-05543-x。[51] Azoulay, P., Fons-Rosen, C. & Graff Zivi n, J. S. 科学是否每次都要等到一位科学家去世后才会进步？《美国经济评论》**109**，2889–2920页（2019年）。网址：https://www.aeaweb.org/articles?id=10.1257/aer.20161574。[52] He, B., Ding, Y., Tang, J., Reguramalingam, V. & Bollen, J. 多学科科学合作网络中的多样性子图挖掘：中观视角。《信息计量学杂志》**7**，117–128页（2013年）。[53] Murray, D. L. 等科研资助评审中的偏见会对小型大学造成严重后果。《公共科学图书馆·综合》**11**，1–19页（2016年）。网址：https://doi.org/10.1371/journal.pone.0155876。[54] 政府事务办公室，O. 建设美国STEM 劳动力：消除障碍，释放优势。技术报告，美国物理学会，马里兰州大学公园市物理椭圆路1号，邮编20740-3844（2021年）。[55]Woodson, T. & Boutilier, S. 影响惠及谁？运用包容性—即时性标准评估美国国家科学基金会资助项目广泛影响中的不平等。《科学与公共政策》**49**，168–178页（2021年）。网址：https://doi.org/10.1093/scipol/scab072。https://academic.oup.com/spp/article-pdf/49/2/168/43395599/scab072.pdf。[56] Chen, C. Y. 等国家科学

基金会（2022年）。网址：osf.io/xb57u。[57] Ginther, D. 等。 种族、族裔与美国国立卫生研究院研究奖项。《科学》（纽约）**333**，1015–9（2011年）。[58] Harris, L. A. 促进竞争性研究既有计划（EPSCoR）：背景与选定议题。技术报告 R44689，美国国会研究处，物理椭圆路1号，马里兰州大学公园市，MD 20740-3844（2017年）。[59] Bollen, J.、Crandall, D.、Junk, D.、Ding, Y. 和 B¨orner, K. 从资助机构到科学机构。《EMBO报告》**15**，131–133（2014年）。网址：https://doi.org/10.1002/embr. 201338068。

arXiv:2301.02396v1 [cs.DL] 6 Jan 2023

补充信息

学术社交网络中的创造力与产出 s

Sean Kelty、Raiyan Abdul Baten、Adiba Proma、Ehsan Hoque、Johann Bollen、Gourab Ghoshal

目录

S1. 数据 S-2
S1.1. 描述性统计 S-2
S1.2. 作者消歧 S-4
S1.3. 主题模型一致性 S-5
S1.4. 示例主题表示 S-6
S1.5. 创新示例 S-8

S2. 新颖性与创新趋势 S-10

S3. 作者与论文层面的新颖性与产出 S-11

S1. 数据

S1.1. 描述性统计

表 S1： **包含摘要和不包含摘要的语料库引文统计汇总。**

DOI 数量	678,916
DOI 数量（含摘要）	250,628
作者数量（消歧后）	307,894
超级明星数量	303
超级明星的 h 指数门槛	21
平均 h 指数	1.74
每篇论文的平均参考文献数	13.5
每篇论文的平均参考文献数（含摘要）	5.57
每篇论文的平均引文数	14.4
每篇论文的平均引文数（含摘要）	6.88

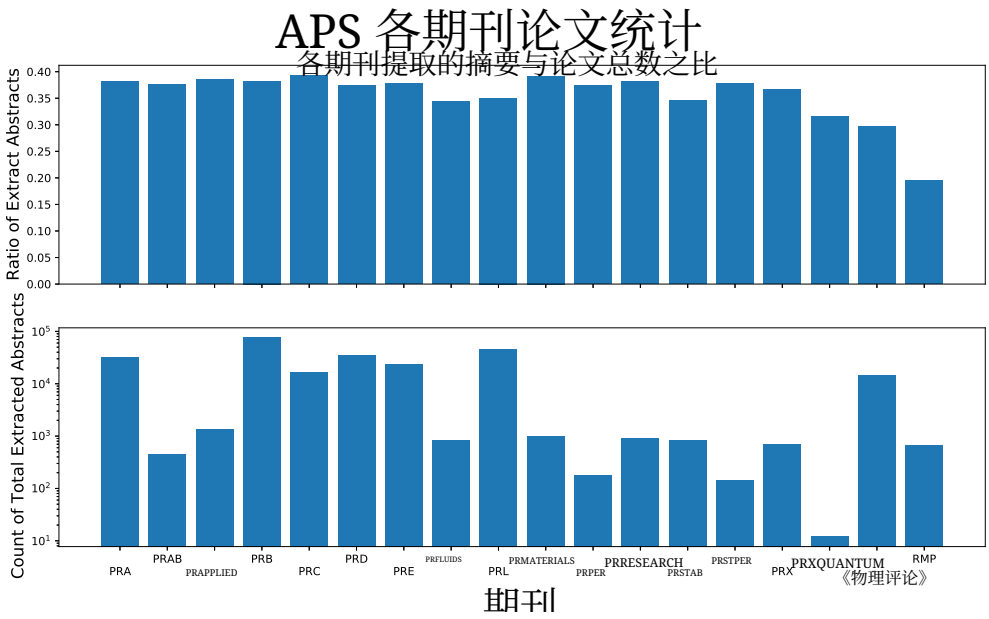

图 S1。上图：APS 各期刊中经分析的论文所占比例。下图：各期刊的论文总数。

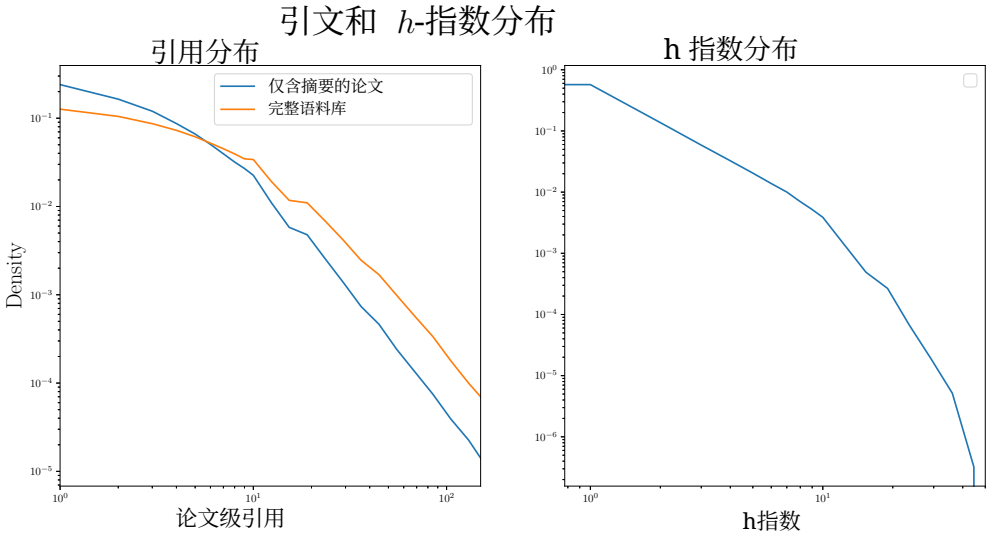

图 S2。左图：所有论文（橙色曲线）和提取了摘要的论文（蓝色曲线）的引文分布。右图：语料库中所有作者的 h指数分布。

S1.2. 作者消歧

我们使用基于评分的方法对数据集中的作者进行消歧。

1. 姓名首字母

- 两个姓名首字母：5
- 超过两个姓名首字母：10
- 姓名首字母冲突：-10

2. 名字

- 常见姓名：3
- 不常见姓名：6（姓名出现次数超过 1000 次才视为常见姓名）

3. 地址/所属机构

- 国家、城市：4
- 国家、城市、组织：7
- 国家、城市、组织、部门：10

4. 共同作者

- 一位：4
- 两位：7
- 超过两位：10

5. 来源

- 期刊：6

6. 自引：10 7. 书目耦合（两部作品引用同一部第三作品）

- 一：2
- 二：4
- 三：6
- 四：8
- 超过四个：10

8. 共被引（第三篇作品同时引用两篇作品的次数）

- 一：2
- 两个：3
- 三个：4
- 四个：5
- 多于4个：6

S1.3. 主题模型一致性

我们使用 UMass 一致性度量，为主题模型确定一个稳定的主题数量。这一一致性得分衡量主题中的高频词彼此之间的相似程度。我们的目标是找到尽可能高的一致性值，并且该值在主题数k附近趋于稳定。图S3 显示，一致性值大约在k = 25个主题时趋于稳定。

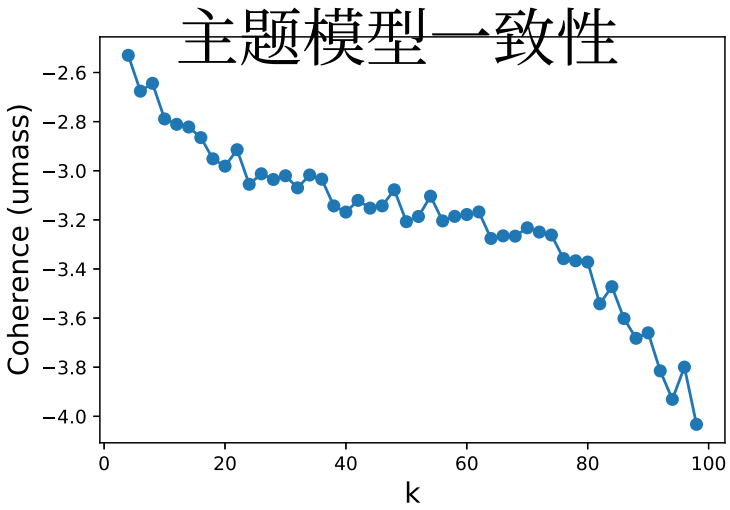

图 S3. P-LDA 主题模型的一致性得分

S1.4. 主题表示示例

语料库中的词语和短语通常统称为“术语”，它们由潜在主题上的分布表示；该分布反映了该术语在整个语料库中的主题分配频率。主题则由与其相关的术语频率来表征。对于每个主题，所有术语都会根据其在该主题上的相对主题频率进行排序。例如，如果某个短语在$k = 3$个主题上的主题分布为 [.1,.2,.7]，，那么该短语就能代表主题 3。预处理术语时，会移除停用词并进行词干提取，使同一个词的不同屈折形式能够以相同的词语表示。

表 S2: **最具代表性的主题模型摘要**

每个主题的术语。

主题数量	代表性术语
主题 1	晶体薄膜，超薄，磁隧道结，扫描隧道显微镜针尖，堆叠结构，自支撑，高分辨角度，Franz，扫描隧道显微镜，力显微镜
主题 2	面心立方结构，温度加成，测量 X，铜酸盐临界温度，温度下降 K，温度 K K，超导临界温度 K，超导临界温度，超导体临界温度，温度 Tc K
主题 3	谱线，ωp，拉曼谱线，吸收部分，吸收线，ND3，电反射。Eliashberg，B1g，电子核双共振
主题 4	轴向磁体，自旋角动量，转动惯量，磁矩，平行磁场，磁体反转，力矩，层间交换，自旋织构，Moriya
主题 5	准直、电子发射、离子产额、离子阱、N4、离子产生、离子等离子体、损伤、壁碳、电子漂移
主题 6	Cauchy、时间反演破缺、Takahashi、Hamilton–Jacobi、自发对称性破缺、快子、Ward恒等式、Polyakov，圈量子宇宙学、库仑规范
主题 7	兴奋性，枢纽，感染，流行病，波动性，精确可解模型，网络模型，突触，突触，积分发放
主题 8	非平衡相变，一级相变，j’，玻璃序，托勒斯跃迁，类玻璃，玻璃形成体，三角晶格，最近邻耦合，最近邻距离

主题9	大小更大、大部分、第四代、尽管、QCD阶、选择定则、第三、马赫-曾德尔 干涉仪、更大、拉曼阶
主题10	准线性、朗之万方程、吉尔伯特方程、状态方程、沙、吸引子、经典混沌、欧拉， 奇美拉态、欧拉方程
主题11	高级激光干涉引力波天文台，麻省理工学院袋模型，目录，背景模型，暗扇区，暗物质，视线，暗物质模型，天空， sno
主题12	巢，范德华力，节点线，螺旋边缘，非费米，态简并，范霍夫，马约拉纳零模， 马约拉纳束缚态，舒布尼科夫-德 哈斯效应
主题13	三维、吸引盆、Fuld-Ferrel、维度平方、LZ、囚禁玻色子、体效应、体力， 硬核玻色子 ，费米原子
主题14	最高占据分子轨道、缀饰球轨道、GaAs1、笼形化合物、Cl2、Cl、六方氮化硼、间隙、凝胶、CI
主题15	脉冲宽度，光学参数，蓝宝石激光器，激子与双激子，光泵浦，二次谐波产生，光学 脉冲，InxGa1−xAs，光学非线性，超强耦合
主题16	Clauser，Horn-Shimoni-Holt，简洁解析表达式，推导，论文部分，解析公式，CB，精确 公式，精确表达式，泡利不相容原理
主题17	一致的原因，发现吻合良好，近期实验数据，发现高度一致，发现良好 一致，理论数据，理论截面，原因与实验结果一致，发现高度一致， 吻合良好，实验结果
主题18	量子比特，寄存器，处理器，研究纠缠，协议，马尔可夫动力学，纯化，诱骗态，量子密钥分发，纠 错
主题19	核子-核子散射、深度非弹性散射、总角动量、包容截面、转移截面 截面、多重碎裂、多重外围反应、微分截面、前向角、 πn
主题20	全势线性增广平面波、波恩波、Wannier函数、冲量、内场、使用路径， 使用蒙特卡洛、在密度泛函中、Jastrow、使用 Hartree
主题 21	自避免行走，非遍历，时间 τ，时间尾部，时间 t2，时间 t2，维数扩散，随机时间，非指数- 幂，均方位移
主题 22	偶偶核，氙-136，富中子核，伽莫夫-泰勒，v’，p0，锕-252， α p， α 反应，p1
主题 23	指向矢场、剪切模量、垂直取向、tλ、反涡旋、湿度、u0、亲水性、剪切带、剪切 应变

主题 24	重要作用、关键作用、Kibble-Zurek、氨基酸、发挥基本作用、发挥主导作用、发挥至关重要的作用、发挥关键作用、发挥核心作用、仍难以捉摸
主题 25	参数 η、ev2、rev c、rev lett、方程、右手、右左、电子-电子碰撞、电子-正电子湮灭、f0

表 S3： **短语-主题分布示例**。

Term	主题嵌入
⋮	
量子化	[1, 0, 0, 0, 0 2259, 0, 0, 560, 0, 0, 882, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 677, 0, 0]
量子	[29, 0, 0, 21, 0, 4304, 1069, 4276, 0, 308, 0, 6008, 454, 46, 14920, 0, 0, 35931, 0, 1828, 0, 0, 1384, 7, 1]
夸克	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 239, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 14542]
夸克偶素	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 125]
夸克偶素	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 299]
四分之一	[0, 0, 0, 0, 30, 0, 0, 0, 0, 0, 0, 321, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
⋮	
量子线	[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 342, 0, 0, 292, 0, 0, 23, 0, 0, 0, 0, 91, 0, 0]
量子芝诺效应	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 12, 0, 0, 31, 0, 0, 0, 0, 0, 0, 0]
夸克-反夸克	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 433]
夸克凝聚	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 107]
夸克衰变	[[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 25]
⋮	

S1.5. 创新示例

创新指标统计某个术语或术语的新组合首次出现在整个语料库中的文章里的情况。图 S4 展示了术语“quantum”和“cosmolog”在语料库中的引入。请注意，“cosmolog”是“cosmology”和“cosmological”等词的词根，这些词在预处理过程中进行了词形还原。我们绘制了这些术语随时间变化的频率，以及表示该术语首次出现年份的竖线

。我们还绘制了短语“quantum cosmolog”的计数，该短语是主题模型中另外考虑的术语。

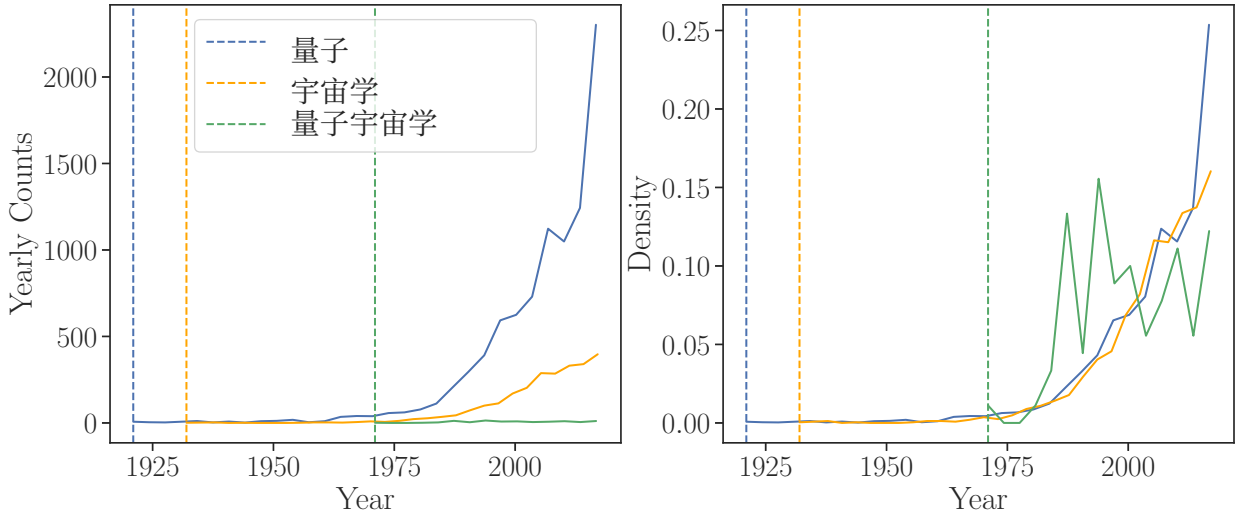

图 S4. 使用“量子”和“宇宙学”术语的创新度量示例 og"

S2. 新颖性与创新的趋势

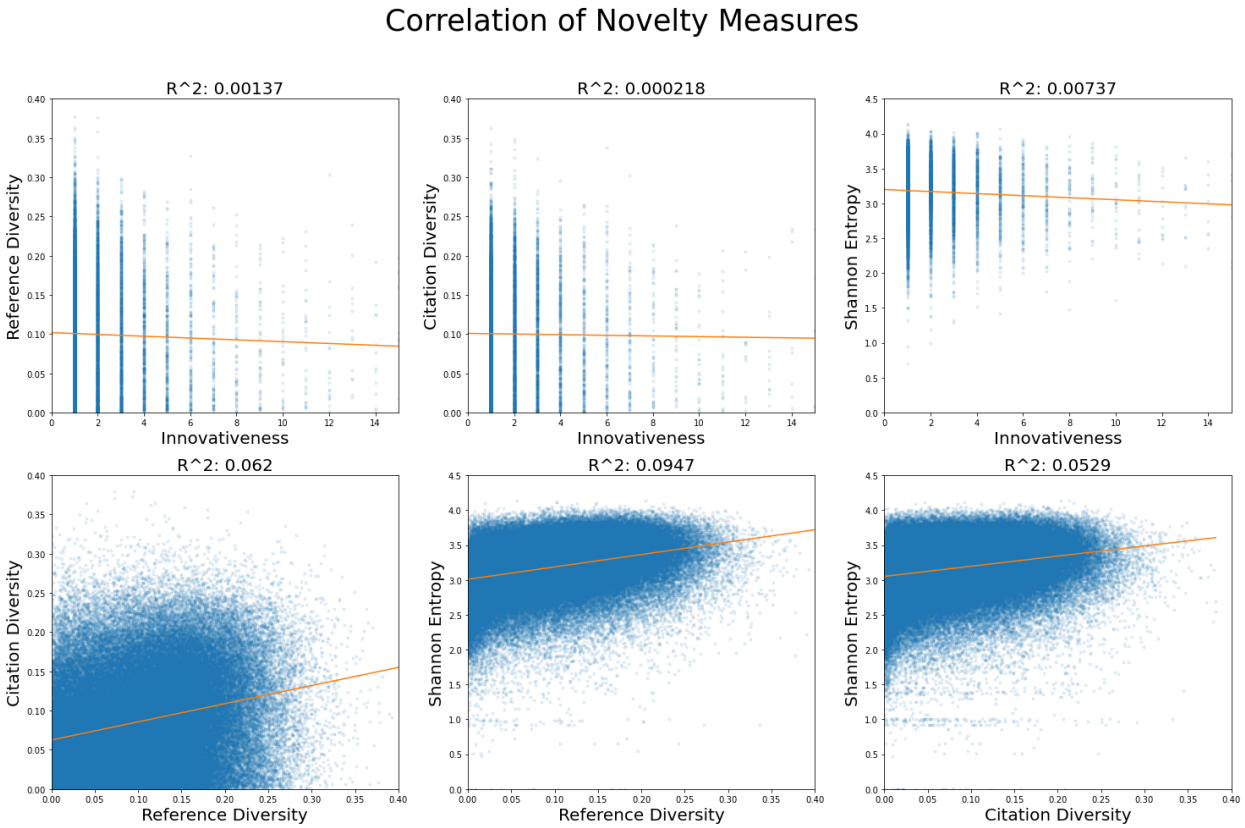

图 S5。 基于皮尔逊相关系数计算的所有新颖性和创新度量之间的相关性 r.

S3. 作者和论文级新颖性与产出

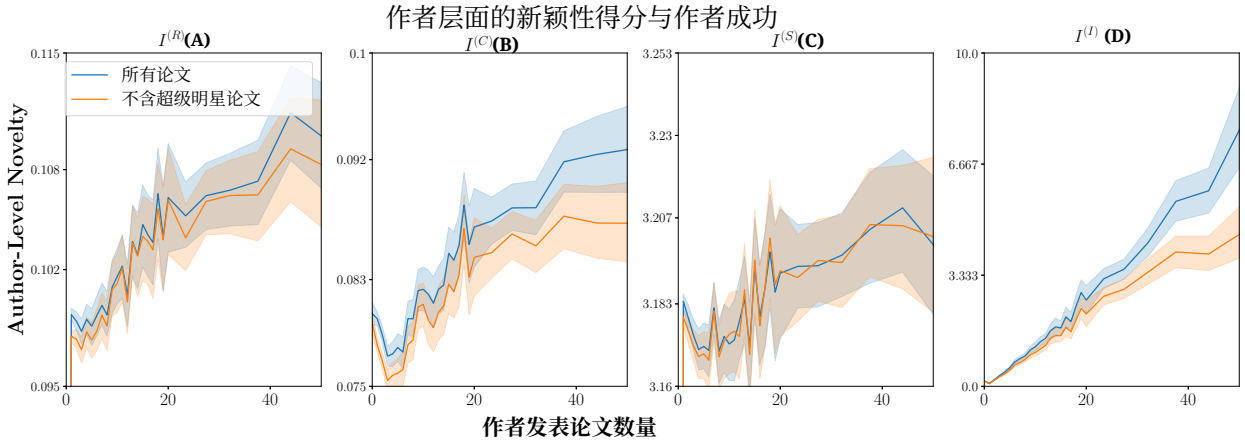

图 S6. 作者发表记录的新颖性和创新指标随其出版物数量的变化。研究对象为语料库中拥有 1 至 50 篇出版物的作者。

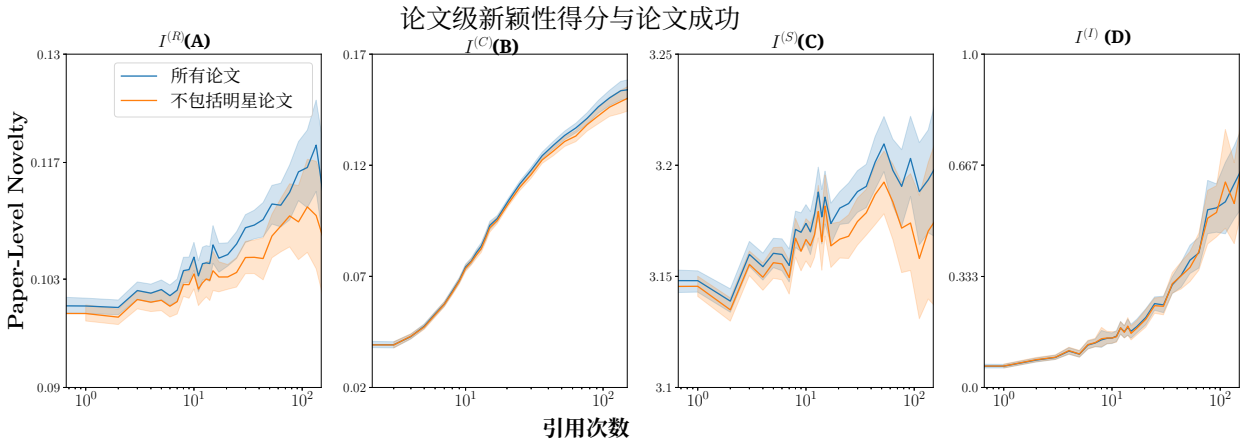

图 S7. 作者发表记录的新颖性和创新指标随其论文获得的引文数量变化。

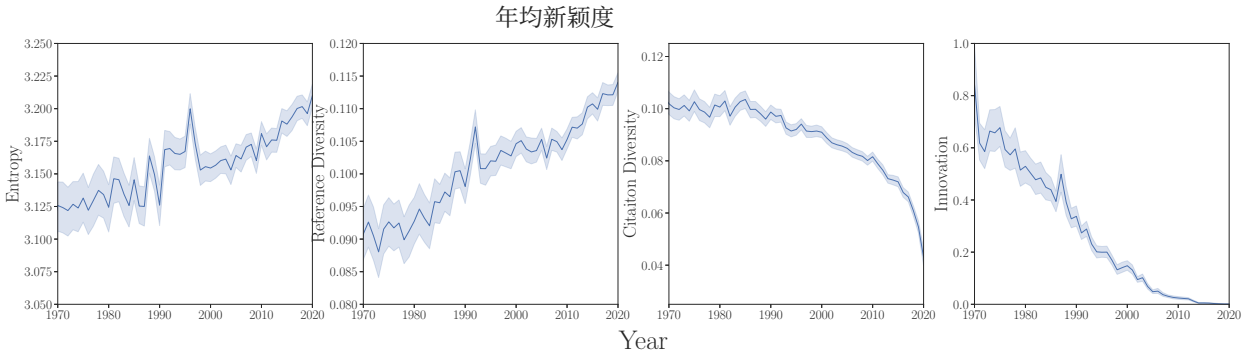

图 S8. 每年的所有新颖性度量。

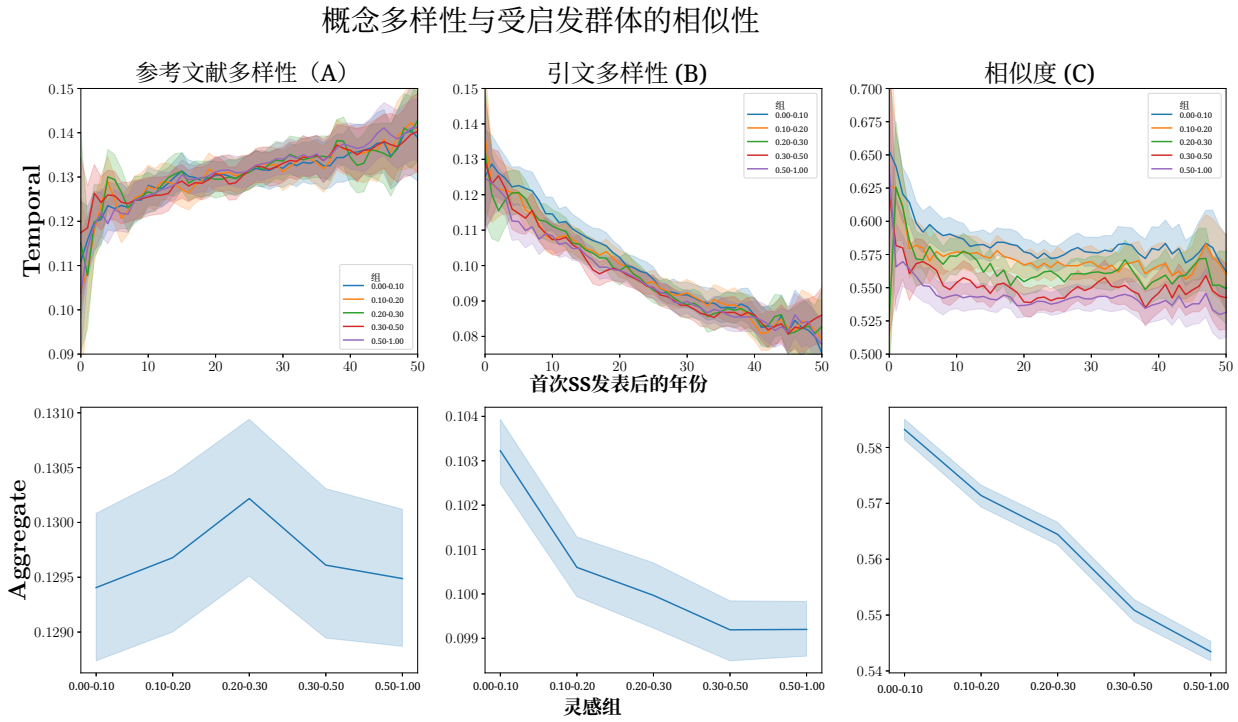

图 S9. (A) 参考文献多样性，(B) 引文多样性，(C) 按启发程度分组的超级明星追随者组内论文相似性。上图：时间演变。下图：时间平均。

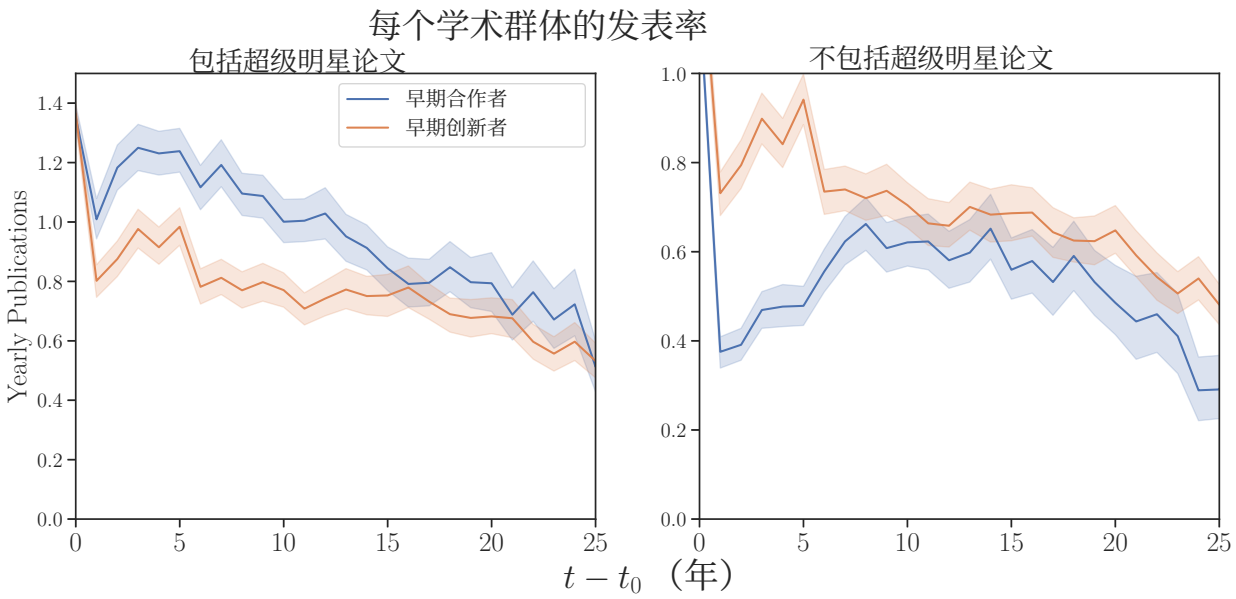

图 S10. 学术群体的发表率，**左**包括超级明星合作，**右**不包括超级明星合作