arXiv:2305.04900v1 [cs.DL] 1 May 2023

个体学者写作风格动态的计算模型

Teddy Lazebnik[1*]、Ariel Rosenfeld[2]

[1] 伦敦大学学院癌症研究所癌症生物学系，英国伦敦[2] 巴伊兰大学信息科学系，以色列拉马特甘

∗ 通讯作者：t.lazebnik@ucl.ac.uk

2023年5月9日

摘要

手稿的写作风格对于决定其读者群、影响力和影响至关重要。以往研究表明，在许多情况下，学者会呈现出独特的写作风格，并体现在其手稿中。本文全面探讨了学者的写作风格在整个职业生涯中的演变，重点关注他们与导师和同行之间的学术关系。我们的结果表明，学者的写作风格往往会在职业生涯早期趋于稳定——大约在发表第 13 篇出版物时。大致在同一时期，学者与导师写作风格之间的差异似乎也会趋于稳定。最后，参与合作的学者人数较少、学者性别相同或来自同一研究领域时，似乎更能促使合著者的写作风格发生变化；其中，年轻学者尤其容易受到影响。

关键词： 学术写作风格、科学传播、计算语言学、时变数学图。

1 引言

学术出版物在知识和信息的推进与传播中发挥着至关重要的作用[1]。除了科学原创性和有效性等许多学科特有因素外，手稿的写作方式，即手稿的风格，对于决定其读者群、影响力和影响效果也至关重要 [2, 3, 4, 5, 6]。具体而言，写得好、表达清晰、易于理解且逻辑连贯、结构合理的手稿，通常能缩短审稿时间、扩大读者群，并受到后续文献更多关注；在某些情况下，也会得到大众媒体和社交媒体的更多关注 [7, 8, 9, 10]。需要指出的是，这种现象并非学术出版物所独有；事实上，新闻报道 [11], 、文献 [12], 以及社交媒体帖子 [13], 等领域也都有充分记录。研究表明，在许多情况下，学者会呈现出独特的写作风格（WS），从而使作者身份分类和画像算法达到令人印象深刻的准确度 [14, 15, 16, 17, 18]。然而，随着学者之间的合作在现代科学中日益普遍 [19, 20, 21], ，协作写作也随之盛行 [22, 23, 24], ，使得许多合著手稿呈现出“混合风格”。也就是说，每位学者的个人风格在最终的合著手稿中体现的方式和程度各不相同 [25, 26]。

这一领域的文献普遍以手稿 为分析单位 [27]。换言之，过去的研究主要是分别考察每篇手稿，分析其内容和/或其作者的身份或特征。然而，将学者 作为分析单位的写作风格研究则少得多。

从社会、文化和职业视角对写作风格变化进行了广泛研究 [28, 29,30, 31, 32]。例如， [33]开发了一种可用于识别在线消息作者的计算模型。作者表明，年龄、性别和母语等参数平均而言是反映个人写作风格的有力指标。类似地， [34] 表明，性别与个人在多种类型文本中的写作风格存在统计关联。此外，[35]表明，在同一时间点，不同年龄的作者具有不同的

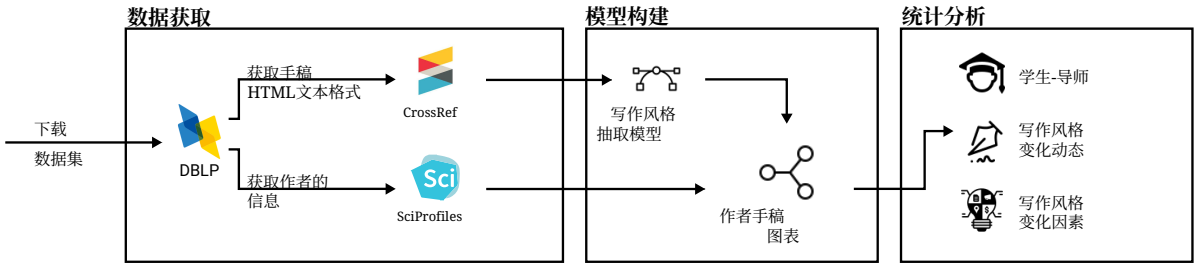


图1：该研究方法论的示意图。

写作风格。 [36] 拓展了这项研究，表明同一个体从儿童成长为成人的过程中，其写作风格也会发生变化。据我们所知，文献尚未考察学者的学术写作风格如何在其职业生涯中演变和形成，尤其是考虑到其与导师及同行的学术关系。

在本研究中，我们旨在回答以下关键问题：

- 一个人的写作风格会如何随时间发生显著变化?
- 研究生（即被指导者）如何与导师的写作风格分道扬镳?
- 学者的写作风格会如何受到其合作的影响?

为回答这些问题，我们提出了一种计算方法，将合著动态的时间图表示、自然语言处理和深度学习技术结合起来。我们将该方法应用于计算机科学（CS）领域真实世界的大规模书目数据，其中包含约57万名CS学者和1,370万篇已编入索引的学术出版物。结果表明，作者的写作风格会随时间变化，并且在发表第13篇手稿前后会出现某种形式的“写作风格趋同”。同样，学者的写作风格似乎在职业生涯早期便开始与其导师的写作风格分离，并在发表第14篇手稿前后趋于稳定。最后，研究发现，性别和研究领域等若干因素在统计上与合作所导致的写作风格变化相关。

本文其余部分安排如下：第2节详细介绍本研究所用的方法和数据。第3节概述分析所得结果，随后第4节对结果进行讨论。最后，第5节总结全文，并指出未来研究可能的方向。

2 方法与材料

我们的方法论包括三个阶段：首先，我们依托广泛收录于热门 DBLP 数据集 [37] 中的计算机科学文献，并分别从 CrossRef[1] 和 SciProfiles[2],获取原始手稿和作者档案。其次，利用所获取的数据构建基于随时间变化的社交图的数学模型。第三，对所得的复杂模型进行分析，以回答上文提出的三个研究问题。图 1 展示了本研究方法论的示意图。

2.1 数据获取

DBLP数据库是一个专业的书目数据库，提供主要计算机科学期刊和会议论文集的开放书目信息，具有较好的覆盖率和准确性 [38]。该数据库于2023年3月20日检索，共获得14,301,639篇出版物。利用其数字对象标识符（数字对象标识符），通过CrossRef将其中13,874,575篇（97%）出版物与其原文进行匹配，并以HTML格式完整恢复。

[1] https://api.crossref.org/ [2] https://sciprofiles.com/

这些出版物由610,281位不同的作者撰写。通过 SciProfiles，我们成功匹配了其中593,513位作者（97.2%），并获取了他们的个人资料。我们筛除了发表手稿超过500篇的学者（占比不足1%）。总体而言，以下分析涵盖了超过1,370万篇计算机科学（CS）出版物，这些出版物由超过587,000位学者撰写。在我们的数据中，每位作者平均发表了 23.42 ± 40.44 篇手稿。

2.2 模型构建

需要注意的是，对任何给定文本的写作风格进行形式化表示都是一项具有挑战性的任务，因为目前尚无清晰且得到一致认可的写作风格定义 [26]。因此，先前处理大量文本的计算研究——这些文本显然无法由领域专家手动标注——采用了数据驱动的方法 [39]。具体而言，基于深度学习的模型被认为是这一领域的最先进方法 [40]。在我们的研究中，我们与既有研究保持一致，使用由[41]提出的最先进模型来表示手稿的写作风格。所采用的模型基于以下假设：大型预训练文本到文本人工神经网络包含可用于通过微调过程 [41]对基于风格迁移的模型 [42] 的解码器进行条件控制的文本写作风格。技术上，我们采用了一个基于T5架构、名为*TextSETTR* [43], 的文本到文本Transformer模型。TextSETTR生成一个1024维向量，并利用Transformer模型中的注意力层接受任意长度的文本。该模型的原始开发者已证明，在广泛的设置下，该模型的性能优于此前的模型。就我们的研究背景而言，研究表明，该模型在超过80%的情况下与写作风格专家的意见一致。这一卓越性能是通过小样本机器学习方法 [44] 获得的；与有监督或无监督学习等其他方法不同 [45, 46], ，该方法在推理过程中只需要极少量带标签的训练示例。综合来看，这些特性使所提出的模型尤其适合应对我们的研究挑战。

将*TextSETTR* 模型应用于给定手稿后，我们便能获得该手稿的写作风格表征。然而，就我们的研究目的而言，我们主要关注的是学者的写作风格表征。因此，如果手稿由一位学者撰写，那么该手稿在当时的写作风格便可完全归于这位学者。相反，如果手稿由多位学者合著，我们便假定其写作风格是各位合著者先前写作风格的混合。为了解析这种写作风格混合，我们假定手稿的各个部分均由一位学者单独撰写，这也是先前文献中的常见假设 [47]。据此，我们采用以下流程：首先，将手稿划分为若干文本连通分量，使每个部分都呈现出独特且一致的写作风格。为此，我们采用了 [14] 提出的最先进模型，该模型在大量文档上完成此任务时的准确率达到 85%。完成文本连通分量划分后，我们使用上文讨论的 *TextSETTR* 模型，分别获取文本中每个连通分量的向量表征。最后，我们通过标准欧几里得距离度量[3]，将得到的每个写作风格向量与当前由最相似写作风格向量表征的共同作者相匹配，从而将该向量映射到其假定来源（即学者）。如果某位学者对应一个连通分量，那么该连通分量在这一时间点的写作风格便视为这位学者的写作风格。若对应多个连通分量，则改用这些连通分量的写作风格向量按比例加权求平均。显然，要执行最后一步（即将连通分量映射到学者），必须先获得每位共同作者先前的写作风格。假设某位学者至少有一篇独著手稿，那么便可将该手稿的写作风格用作迭代过程的起点。具体而言，我们从学者的第一篇独著手稿开始，将该时间点识别出的写作风格按照上述流程，迭代传播到随后及之前的手稿。换言之，以每位学者的第一篇独著手稿为起点，我们使用所得写作风格为紧接着的随后及之前手稿分配连通分量（及其写作风格）；随后，这些结果又用于处理更前及更后的手稿，依此类推。没有任何独著出版物的学者不纳入后续分析（比例低于 3%）。

2.3 作者-手稿图

为捕捉随时间变化的合作和 WS动态，我们定义一个图， $G = (S,\ M,\ E)$，其中每位学者在图 $s \in S$中表示为一个节点，每篇手稿在图中表示为不同类型的节点

[3]不太可能的联系可能导致一个连通分量被分配给多位合著者。

$m \in M$，并且一条有向边 $e \in E \subseteq A \times P$连接每位学者与其每篇手稿。形式上，学者节点 $s \in S$ 由元组 $s := (f,\ g)$定义，其中 f 是学者的主要研究领域，根据其主要关联系的名称确定（例如，计算机科学、数学、物理学）[4]；g 是学者的性别，可取值之一为 $\{male, female, unknown\}$。学者的性别根据对 [48] 提出的模型进行查询而得出；该模型使用约 1 亿组姓名与性别关联数据进行训练，这些数据由 *Yahoo*！收集。只有当模型预测的置信度高于 95% 时，才会记录学者的性别（94.6% 的学者符合这一条件）。手稿节点 $m \in M$ 由与手稿各连通分量相关联的 WS 向量以及手稿的发表时间定义，分别记为 ξ 和 t。最后，一条边 $e = (s,\ m) \in E$ 表示学者 s 是手稿 m 的（共同）作者。总体而言，我们确定了 81.3% 学者的主要研究领域，以及 94.6% 学者的性别。共有 80.5% 的学者档案同时包含这两个属性。

2.4 统计分析

统计分析分为三个部分，分别对应我们的三个主要研究问题。

WS动态： 为量化特定学者的 WS 随时间发生的变化，我们定义函数 $C: \mathbb{R}^{\kappa \times N} \times \mathbb{R}^N \to \mathbb{R}^+$ 其中，κ 是作者每年发表论文数量的年均值（四舍五入到最接近的自然数）。此外，$N \in \mathbb{N}$ 是 WS 表征向量的维度。形式上，C 接受一个列表（L），其中包含 $\kappa \in \mathbb{N}$ 个 WS 向量，这些向量对应于在参考 WS u 之前发表的 κ 个 WS 向量；参考 WS u也会提供给 C:

$$C(L, u) := ||\frac{1}{\kappa} \sum_{v \in L} (v) - u||, \tag{1}$$

其中，$||x||$ 是向量 x的 L_1 范数。直观而言，上述计算量化了当前呈现的 WS 与大约在前一年期间呈现的过往 WS 相比有多大差异。

写作风格的形成： 要捕捉研究生 WS 的形成过程，需要确定学生的导师以及学生的毕业时间。虽然在我们的初步调查中，数学谱系项目[5] 和 Academic Family Tree[6], 提供了一些众包的导师—学生数据，但这些数据并未涵盖我们数据中的很大一部分。因此，我们采用一种启发式方法，该方法已在先前的研究中成功应用（例如，[49]）：将学者在其发表论文的前三年中与其合著手稿数量最多的个人视为其导师。请注意，这种简单的启发式方法可能同时识别出“正式”和“非正式”导师，这似乎更符合我们的研究目的。形式上，将学生与导师在前三年内合著的手稿集合分别记为 A 和 $\rho := |A| \in \mathbb{N}$。因此，将学生接触到的导师 WS 设定为集合 A中手稿所体现的导师平均 WS。由此，学生相对于其导师（A）的风格形成函数定义如下：$\delta_A(i): \mathbb{R}^N \to \mathbb{R}^+$，使得 $\delta_A(u) = ||\frac{1}{\rho} \sum_{v \in A}(v) - u||$，其中 $u \in \mathbb{R}^N$是学生的 WS 向量，我们希望将其与基准 WS 进行比较；基准 WS 由 $\frac{1}{\rho} \sum_{v \in A}(v)$计算得出。由于我们关注学者 WS 随时间的变化，将 u_i 定义为学者的 i_{th} 个 WS 向量，其中 u_0 是该学者在集合 A中最新手稿之后发表的第一篇手稿。直观而言，我们将当前呈现的 WS 与学生在受训期间通过合著手稿接触到的导师平均 WS 进行比较。

写作风格与合作： 对于每篇合著手稿，我们提取以下特征：针对每位共同作者，我们获取其主要研究领域（f）、性别（g）以及已发表手稿的数量。我们还会提取手稿署名栏中列出的合著者人数。这些值被视为

[4]我们通过在 Google 中搜索学者的姓名，从第一个链接中获取数据，然后手动使用正则表达式进行数据标准化。[5] https://www.genealogy.math.ndsu.nodak.edu/[6]https://academictree.org/

学习模型的特征向量 x。我们将因一篇合著手稿而观察到的学者写作风格变化定义为式（1）所示的目标值，并记为 y。由于合著者人数可能不固定，我们将 x 的大小设为数据库中任何手稿所需的最大值，并相应地填充 x 中不需要的位置。

将由样本构成的数据集输入基于树的管道优化工具（TPOT）自动化机器学习模型（AutoML）[50, 51] 。该数据集以每位共同作者的特征为输入，以观测到的写作风格变化为输出。该模型尤其适用于复杂的回归任务，旨在尽量降低预测的平均绝对误差。我们计算并报告特征重要性，以确定各参数对模型预测能力的预期影响 [52]。

此外，我们将每种写作风格变化（即，y）归为以下类型之一：1）趋向质心（即，所有合著者的写作风格都更接近群体的平均写作风格）；2）单向正向变化（即，相关学者的写作风格更接近群体的平均写作风格，但不符合前一项标准）；3）单向负向变化（即，相关学者的写作风格偏离群体的平均写作风格）；4）无明显变化（即，不符合其他任何标准）。据此，我们采用带事后图基校正的方差分析检验进行统计分析，以确定以下分析中详述的特定情况是否与不同类型的写作风格变化存在统计关联。

3 结果

WS动态： 图2展示了作者在不同手稿中的写作风格（WS）变化，其中 $|L| = 3$。正如我们所观察到的，在最初大约12篇已发表的手稿中，学者的写作风格（WS）在相邻手稿之间存在显著差异。从这一阶段开始，学者的写作风格（WS）虽然并非恒定不变（即WS变化不为零），但其变化幅度似乎较小，并且随着时间推移相对稳定。为了从数学上捕捉这一现象，我们引入一个针对WS变化的阈值，以确定学者的写作风格（WS）是否以及何时达到收敛。形式上，对于阈值 $\omega \in \mathbb{R}^{+}$，收敛点 $\alpha \in \mathbb{N}$ 定义为 $\alpha := \min_{j} \left(\frac{1}{z-j} \sum_{i \in [j,\ldots,z]} C(L,\ u_i) \right) \leq \omega$。该分析的结果汇总于表1。需要注意的是，由于并非所有学者都会在给定阈值 ω 下达到收敛，因此我们报告了在指定阈值下实现收敛的学者所占的百分比。

阈值 (ω)	0.01	0.02	0.03	0.04	0.05
收敛点 (α)	13.3 ± 6.7	10.8 ± 4.9	8.5 ± 4.2	7.1 ± 3.6	5.2 ± 3.0
收敛率（%）	89.4	92.3	95.7	96.5	96.9

表1：每个所考察阈值（列）对应的收敛点均值 ±标准差（上行）以及收敛学者的比例（下行）。

有人可能会推测，不同学者的 WS 变化动态可能存在显著差异。为了检验这一假设，我们使用了针对时间序列数据调整的经典 k 均值算法 [53]，并采用了流行的*tslearn* 库 [54]。图 3 展示了我们的数据在不同簇数下的 L_2 簇内指标。通常，当数据本身划分为 $k > 1$ 个簇时，人们会预期图中出现一个“肘部”，表明簇内指标的下降幅度由大变小。然而，如图所示，这里并非如此。因此，数据似乎不支持这一假设。

写作风格的形成： 图 4 展示了学生从毕业时起写作风格的形成过程。该图呈现出一种类似 sigmoid 的趋势：在手稿数量增加的过程中，学生与其导师的写作风格差异逐渐增大。与写作风格随时间变化的分析类似，大约在毕业生发表第 14 篇出版物时，其与导师的差异似乎趋于一个相对稳定的距离。

让我们以本文的两位作者为例进行说明。图5展示了他们各自在毕业后发表的前10篇出版物相对于其在受训期间导师的写作风格（WS）进行的二维标准主成分分析（PCA）降维投影。第二位作者（显示在右侧）呈现出一种相当一致的写作风格形成模式，随着时间的推移，其写作风格沿着同一方向逐渐偏离导师的写作风格。

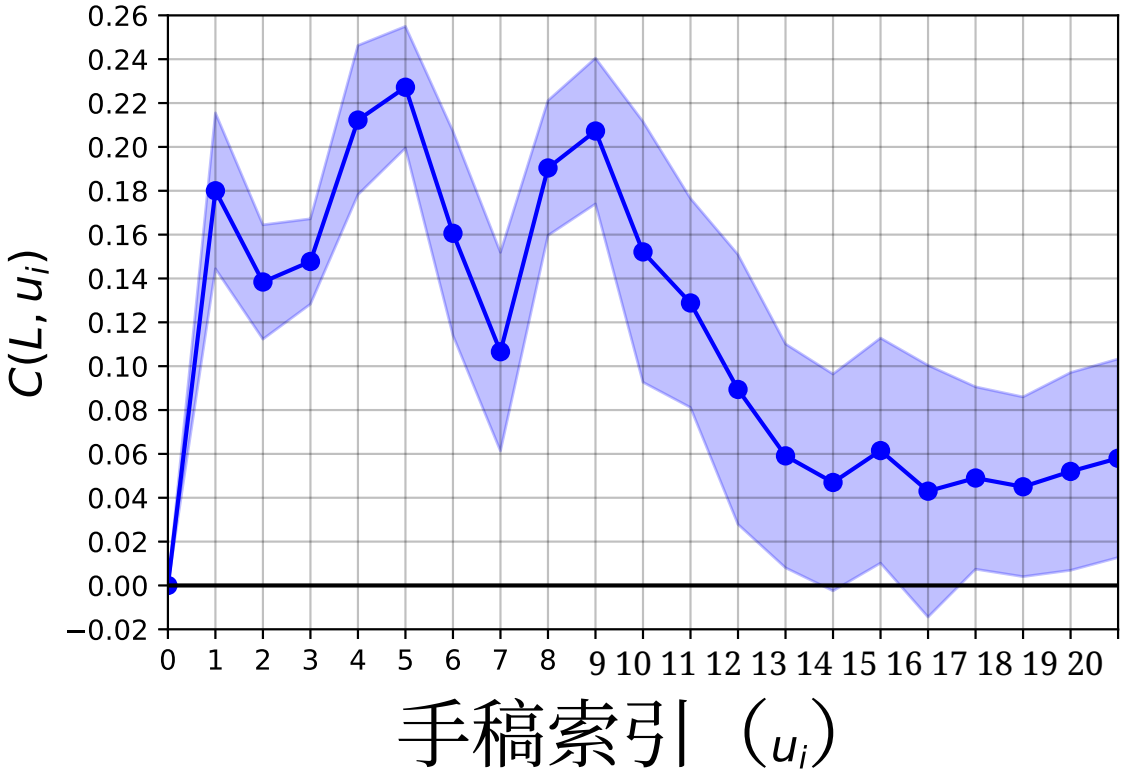


图2：整个研究人群WS变化的均值 ±标准差。

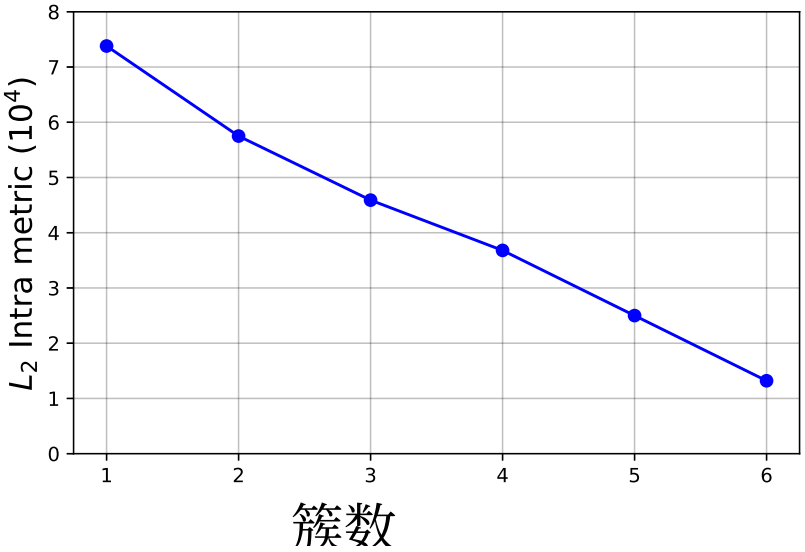


图3：使用 L_2 距离的k均值聚类方法的肘部图。未观察到明显的肘点。

然而，第一作者（显示在左侧）则呈现出一种更为杂乱的模式，且随着时间推移并未表现出明确的方向性。

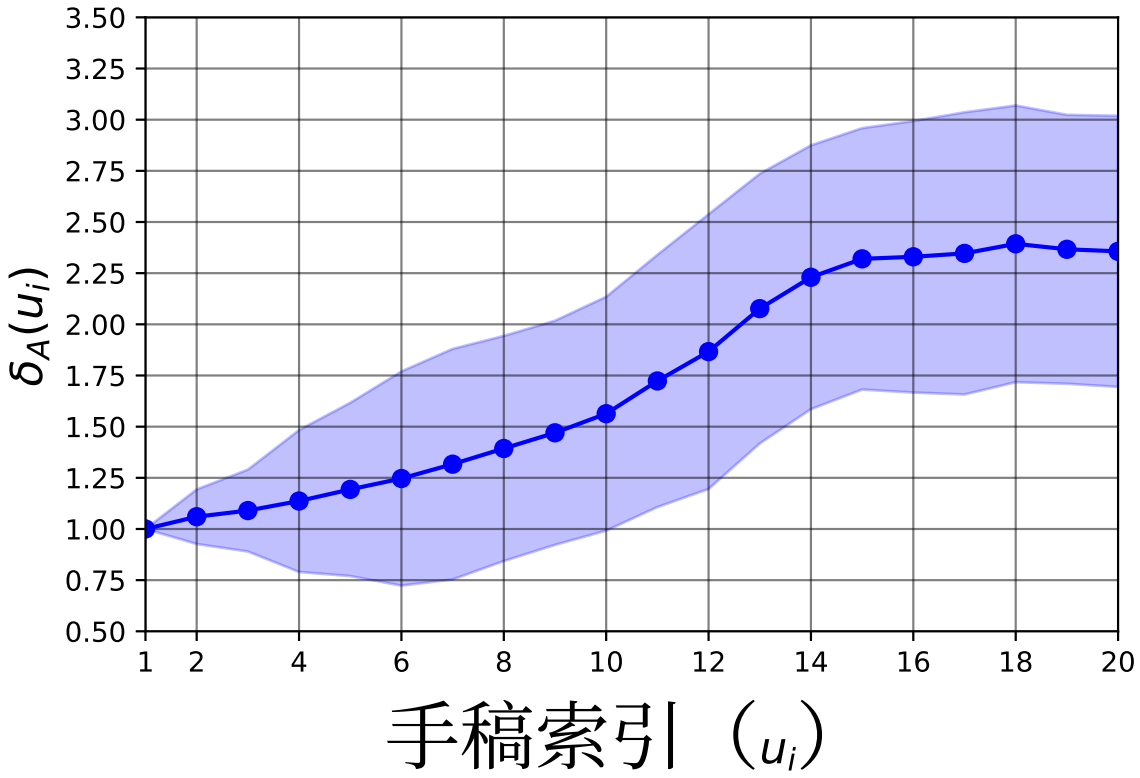
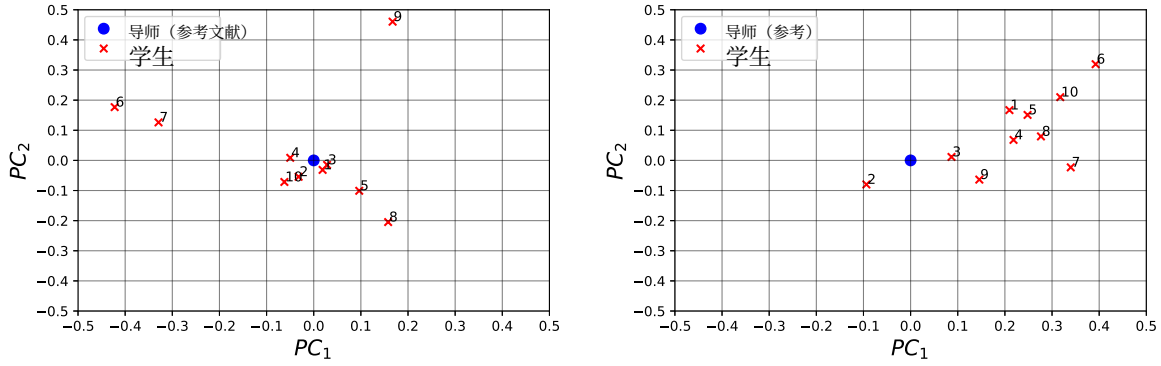


图 4：在多篇手稿中，学者的写作风格与其导师的写作风格之差的均值 ± 标准差。



（a）第一作者（$\rho = 5$）。 （b）第二作者（$\rho = 12$）。

图 5：学者毕业后发表的手稿以“x”标记，并按照发表顺序编号。

WS与合作：由于每篇手稿的合著者人数不定，导致特征数量相当庞大，因此我们重点关注表2中总结的四个因素。具体而言，我们首先考察合著手稿中涉及的性别，并针对每位受考察的学者，将每篇合著手稿分为以下四类：1）全男性（即所有合著者均为男性）；2）男性混合（即该学者为男性，其余合著者中既有男性也有女性）；3）女性混合

（即该学者为女性，其余合著者中既有男性也有女性）；4）全女性（即所有合著者均为女性）。随后，我们考察所有合著者的研究领域是否相同（同时认可词序等细微差异，并忽略“系”或“院系”等通用词语）。“ 合著者人数 ” 这一因素指手稿署名栏中列出的作者人数，而 “ 先前发表的出版物 ” 这一因素指该学者此前发表的出版物数量。对于上述每项因素，文中均列出了改变自身WS的平均重要性。此外，还报告了各个取值与特定类型WS变化之间的统计关系。

考察因素	**平均重要性**	**样本 size**	**典型变化方向**
性别	0.11	15M	全男性：趋向质心*
		32M	男性混合：无变化**
		21M	女性混合：无变化*
		4M	全女性：单侧正向变化*
研究领域	0.23	29M	相同：趋向质心**
		43M	不同：无明显变化**
合著者人数	0.32	9M	*2*：趋向质心
		17M	*3*：单侧正向变化*
		46M	*4+*：无明显变化**
先前发表的出版物	0.34	2M	*1-3*：正向单侧变化*
		4M	*4-13*：朝质心方向变化*
		37M	*14+*：无明显变化*

表2：对于每个考察因素（各行），我们报告其在解释联合出版物后观察到的 WS变化方面的估计重要性，以及该因素与特定类型 WS变化之间的统计关联。具有统计显著性的结果以 * 表示 $p \leq 0.05$ ，以 ** 表示 $p \leq 0.01$。

4 讨论

让我们重新审视本研究提出的最初研究问题。

首先，我们提出了这样一个问题：“一个人的写作风格会随着时间发生怎样的显著变化？”结果似乎表明，绝大多数学者的写作风格都在不断演变：起初表现得杂乱无章，但很快便趋于平缓，在第13篇手稿左右仅有轻微而稳定的变化。一个人的写作风格从一篇手稿到下一篇逐渐趋于细微且稳定的变化，这一点多少是符合直觉的，因为这反映了一个人逐渐形成独特的学术个性、风格与实践的过程，而这些也始终在演变。然而，令我们非常惊讶的是，这种趋稳竟发生在职业生涯的早期。我们的数据显示，平均而言，这种趋稳在发表四年后便会出现。这意味着学者的写作风格“学习曲线”极早就趋于平缓。一种可能的解释是，学者在学术生涯最初几年面临着臭名昭著的高压，必须大量发表论文，部分原因是为了获得长期职位 [55]。具体而言，在这几年里，学者可能会放弃漫长且或许必要的写作风格打磨过程，转而优先充实自己的学术成果。

其次，我们提出了这样一个问题：“研究生（即被指导者）如何逐渐脱离其导师的写作风格（WS）？”结果似乎表明，个人的 WS 与其导师 WS 之间的距离以类似 sigmoid 的方式不断增大，直到在其第 14 篇出版物左右达到收敛。有趣的是，这种收敛似乎也与先前分析所得的收敛结果一致。所观察到的模式似乎与历史上观察到的学徒制动态相吻合 [56]。结果还表明，标准差呈现出不断增大的趋势，如图 4 所示。这些结果表明，一个人偏离其导师 WS 的过程具有很强的个人特征，这与近期一项专门研究计算机科学领域导师与被指导者的合作模式的研究结果一致 [57]。

最后，我们提出了“合作如何影响学者的写作风格？”这一问题。结果表明，有几个具有统计显著性的因素似乎决定了合作影响一个人写作风格的方式。首先来看性别，它是所考察因素中影响最小的。有趣的是，尽管.....之间的相互作用

男性和女性在统计意义上是对称的，即他们没有特定的改变自身写作风格（WS）的方式；相同性别之间的互动会产生不同的结果。这一结果与大量关于合作和性别的既有研究一致 [58, 59, 60]。研究领域似乎发挥着稍微更为核心的作用。具体而言，来自同一研究领域的合著者会显著受到彼此写作风格的影响，而来自不同领域的合著者则不会。这一结果或许有些违背直觉，因为人们可能会预期，来自不同学科的学者由于习惯于略有不同的写作规范和实践，会对彼此产生更大的影响。然而，这一结果与学者如何回应并采纳来自其研究领域内外同行的观点这一现象在本质上是相似的 [61]。此外，合著者人数对于预测合作导致的写作风格变化具有重要作用。尽管在统计上并不显著，但当只有两名合著者时，双方似乎都会相互学习，并略微采纳彼此的写作风格。然而，当涉及三名合著者，尤其是超过三名合著者时，学者较少受到其合著者写作风格的影响。最重要的因素是学者此前发表的手稿数量。与前文讨论的学生—导师动态的结果类似，年轻学者（以发表的手稿数量衡量）往往比更有经验的作者更容易受到他人写作风格的影响；后者已经形成了个人写作风格，因此不太容易发生变化。

5 结论

在本研究中，我们探讨了学者在整个职业生涯中写作风格的演变过程，重点关注他们与导师及同行之间的学术关系。为此，我们提出并实现了一个计算框架，用于捕捉学者的写作风格（WS）如何随着时间的推移，因合著关系和导师—被指导者关系而发生变化。所得结果揭示了若干引人深思的现象。这些现象不仅对于理解学术写作风格动态具有基础性意义，也有助于设计更完善的博士培养项目和青年教师项目。

需要指出的是，所提出的模型和分析并非没有局限性。首先，我们的分析聚焦于计算机科学这一学科。在未来的研究中，我们计划将分析范围扩展至其他学科，而这些学科未必遵循计算机科学领域的实践方式和标准（例如，人文与社会科学）。其次，我们的实现中有若干部分存在不准确之处，这几乎是其本身所固有的，例如共同作者的连通分量映射。因此，用于分析的原始结果不可避免地存在噪声和错误。改进这些组件有望获得更加稳健的结果和结论。最后，所提出的分析并未考虑其他可能影响学者写作风格（WS）及其变化的社会和文化特征 [62]。例如，学者的国籍很可能在塑造其写作风格及其动态变化方面发挥着核心作用。我们计划在未来进一步探索这一因素以及其他社会人口统计特征。

声明

资金支持

这项研究未获得来自公共部门、商业部门或非营利部门资助机构的任何专项资助。

利益冲突/竞争性利益

无。

数据可用性

所使用的数据已在手稿中列出，并注明了相关来源。

参考文献

[1] F. Livesey. 学术手稿市场。《欧洲市场营销杂志》，15(7):52–67，1981年。[2] L. Xiao 和 N. Askin. 学术出版的维基百科：优势与挑战。《在线信息评论》，36(3):359–373，2012年。[3] Y. Sun 和 C. L. Giles. 面向学术数字图书馆的流行度加权排名。载于 G. Amati、C. Carpineto 和 G. Romano 编，《信息检索进展》，第 605–612 页，2007年。[4] M. A. Diego、T. M. Field 和 C. E. Sanders. 学业表现、受欢迎程度和抑郁症可预测青少年物质使用情况。《青春期》，38:149，2003年。[5] P. van den Besselaar 和 C. Mom. 写作风格对资助申请成功率的影响。《计量学杂志》，16(1):101257，2022年。[6] K. Sun、H. Liu 和 W. Xiong. 科学写作中语言的演变模式：皇家学会《哲学汇刊》（1665–1869）的案例研究。《科学计量学》，126:1695–1724，2021年。[7] P. K. Matsuda 和 C. M. Tardy. 学术写作中的声音：盲审手稿中作者身份的修辞建构。《专门用途英语》，26(2):235–249，2007年。[8] J. Hartley、J. W. Pennebaker 和 C. Fox. 摘要、引言和讨论：它们的风格差异有多大？《科学计量学》，57:389–398，2003年。[9] A. Duszak. 学术话语与知识风格。《语用学杂志》，21(3):291–313，1994年。[10] V. Vysotska、Y. Burov、V. Lytvyn 和 A. Demchuk. 为学术环境中的抄袭检测界定作者风格。载于*2018 IEEE* 第二届数据流挖掘与处理国际会议（*DSMP*），第 128–133 页，2018年。[11] B. Hendriks、F. van Meurs、H. Korzilius、R. le Pair 和 S. le Blanc-Damen. 风格一致性与说服力：一项跨文化研究，探讨风格维度差异对英国和荷兰商业通讯说服力的影响。《*IEEE* 专业传播汇刊》，55(2):122–141，2012年。[12] S. Shiyab 和 M. S. Lynch. 文学风格能否翻译？《巴别塔》，52(3):262–275，2006年。[13] C. N. dos Santos、I. Melnyk 和 I. Padhi. 利用无监督文本风格迁移打击社交媒体上的冒犯性语言。*arXiv*，2018年。[14] R. Singh、J. Weerasinghe 和 R. Greenstadt. 多作者文档中的写作风格变化检测。载于*CLEF 2021*——评估论坛会议与研讨会，2021年。[15] M. Koppel 和 Y. Winter. 判定两份文档是否由同一作者撰写。《信息科学与技术协会杂志》，65(1):178–187，2014年。[16] N. Akiva 和 M. Koppel. 识别多作者文档中的不同连通分量。载于*2012*年欧洲情报与安全信息学会议，第 205–209 页，2012年。[17] M. Koppel 和 Y. Winter. 作者身份归属：哪些容易，哪些困难？《法律与政策杂志》，21(2):317–331，2013年。[18] C. Lu、J. Bu、Y.and Wang、Y. Ding、V. Torvik、M. Schnaars 和 C. Zhang. 从语言复杂性的角度考察科学写作风格。《信息科学与技术协会杂志》，70(5):462–475，2019年。[19] T. Amjad、Y. Ding、J. Xu、C. Zhang、A. Daud、J. Tang 和 M. Song. 站在巨人的肩膀上。《计量学杂志》，11(1):307–323，2017年。

[20] S. Wuchty、B. F. Jones 和 B. Uzzi。团队在知识生产中的主导地位日益增强。*Science*，316(5827):1036–1039，2007年。[21] C. Zhang、Y. Bu、Y. Ding 和 J. Xu。理解科学合作：同质性、传递性与优先连接。*Journal of theAssociation for Information Science and Technology*，69(1):72–86，2018年。[22] A. O. Holcombe。贡献者而非作者：用贡献说明谁做了什么。*Publications*，7(3)，2019年。[23] B. Bozeman 和 E. Corley。科学家的合作策略：对科学与技术人力资本的影响。*Researchpolicy*，33(4):599–616，2004年。[24] C. Zhang、Y. Bu、Y. Ding 和 J. Xu。理解科学合作：同质性、传递性与优先连接。*Journal of theAssociation for Information Science and Technology*，69(1):72–86，2018年。[25] J. Hartley、J. W. Pennebaker 和 C. Fox。运用新技术评估男女配对及个人的学术写作风格。*Journal of Technical Writing and Communication*，33(3):243–261，2003年。[26] J. Hartley、M. Howe 和 W. McKeachie。历时写作：新技术对写作影响的纵向研究。*British Journal of Educational Technology*，32(2):141–151，2001年。[27] N. Song、K. Chen 和 Y. Zhao。理解IS-LS领域科学论文的写作风格：基于过去三十年摘要的证据。*Journal of Informetrics*，17(1):101377，2023年。[28] F. Can 和 J. M. Patton。写作风格随时间的变化。*Computers and the Humanities*，38:61–82，2004年。[29] E. L. Snow、L. K. Allen、M. E. Jacovina、C. A. Perret 和 D. S. McNamara。你很有风格：检测随时间变化的写作灵活性。载于第五届国际学习分析与知识会议论文集，第194–202页。计算机协会，2015年。[30] D. S. McNamara、S. A. Crossley、和 P. M. McCarthy。写作质量的语言特征。*WrittenCommunication*，27:57–86，2010年。[31] S. A. Crossley、R. Roscoe 和 D. S. McNamara。什么是成功的写作？探究作者成功撰写文章的多种方式。*Written Communication*，31:184–214，2014年。[32] L. K. Allen、E. L. Snow 和 D. S. McNamara。叙事华尔兹：灵活性在写作能力中的作用。*Journal of Educational Psychology*，108(7):911–924，2016年。[33] R. Zheng、J. Li、H. Chen 和 Z. Huang。在线消息作者身份识别框架：写作风格特征与分类技术。*Journal of the American Societyfor Information Scienceand Technology*，57(3):378–393，2006年。[34] D. L. Rubin 和 K. Greene。书面语言中的性别典型风格。*Research in theTeaching of English*，26(1):7–40，26。[35] S. Rosenthal 和 K. McKeown。博客中的年龄预测：对社交媒体前后世代的风格、内容和在线行为的研究。计算语言学协会第*49*届年会：人类语言技术会议论文集，第19–24页，2011年。[36] W. Haverals、L. Geybels 和 V. Joosen。各年龄皆有其风格：对面向儿童、青少年及成人的跨界写作者进行文体计量研究。*Language and Literature*，31(1):62–84，2022年。[37] N. Aggrawal 和 A. Arora。使用Gephi对DBLP合著网络进行可视化、分析和结构模式注入。载于*2016*年第二届国际下一代计算技术会议（*NGCT*），第494–500页，2016年。

[38] A. Rosenfeld。DBLP是一个优秀的计算机科学期刊数据库吗？*Computer*，56(3):101–108，2023年。[39] M. Gridach。基于（概率）软逻辑和神经网络的自然语言处理框架。*Applied SoftComputing*，93:106232，2020年。[40] D. W. Otter、J. R. Medina 和 J. K. Kalita。深度学习在自然语言处理中的应用综述。*IEEE*神经网络与学习系统汇刊，32(2):604–624，2021年。[41] P. Rileya、N. Constantb、M. Guob、G. Kumarc、D. Uthusb 和 Z. Parekhb。Textsettr：少样本文本风格提取与可调式定向风格重塑。*arXiv*，2021年。[42] F. A. Acheampong、H. Nunoo-Mensah 和 W. Chen。用于基于文本的情绪检测的Transformer模型：基于BERT的方法综述。*Artif Intell Rev*，54:5789–5829，2021年。[43] C. Raffel、N. Shazeer、A. Roberts、K. Lee、S. Narang、M. Matena、Y. Zhou、W. Li 和 P. J. Liu。探索统一文本到文本Transformer的迁移学习极限。*Journal of Machine Learning Research*，21(140):1–67，2020年。[44] P. Xu、J. C. K. Cheung 和 Y. Cao。无监督条件下可控文本表征的变分学习。载于第*37*届国际机器学习会议论文集，第119卷*Proceedings of Machine Learning Research*，第10534–10543页，2020年。[45] H. Jhamtani、V. Gangal、E. Hovy 和 E. Nyberg。使用富含复制机制的序列到序列模型，将现代语言莎士比亚风格化。载于文体变异研讨会论文集，第10–19页，2017年。[46] G. Lample、S. Subramanian、E. Smith、L. Denoyer、M. Ranzato 和 Y. Boureau。多属性文本改写。载于国际学习表征会议，2019年。[47] J. Bevendorff、B. Chulvi、G. L. De La Pe˜na Sarrac´en、M. Kestemont、E. Manjavacas、I. Markov、M. Mayerl、M. Potthast、F. Rangel、P. Rosso、E. Stamatatos、B. Stein、M. Wiegmann、M. Wolska 和 E. Zangerle。PAN 2021概述：作者身份验证、Twitter仇恨言论传播者画像及风格变化检测。载于 K. S. Candan、B. Ionescu、L. Goeuriot、B. Larsen、H. M¨uller、A. Joly、M. Maistro、F. Piroi、G. Faggioli 和 N. Ferro 编，实验信息检索：多语言、多模态与交互，第419–431页。Springer国际出版，2021年。[48] Y. Hu、C. Hu、T. Tran、T. Kasturi、E. Joseph 和 M. Gillingham。名字意味着什么？——使用基于字符的机器学习模型进行姓名性别分类。*Data Mining and Knowledge Discovery*，4，2021年。[49] V. Suresh、N. Raghupathy、B. Shekar 和 C. E. V. Madhavan。从作者合作网络中发现导师关系信息。载于 V. Corruble、M. Takeda 和 E. Suzuki 编，*Discovery Science*，第197–208页，2007年。[50] R. S. Olson 和 J. H. Moore。TPOT：用于自动化机器学习的基于树的管道优化工具。载于自动化机器学习研讨会，第66–74页。PMLR，2016年。[51] T. Lazebnik 和 A. Somech。展示SubStrat：一种在大型数据集上加速自动化机器学习的基于子集的策略。载于第*31*届*ACM*国际信息与知识管理会议论文集，第4907–4911页。计算机协会，2022年。[52] X. Li、Y. Zhou、N. C. Dvornek、Y. Gu、P. Ventola 和 J. S. Duncan。在不确定性下高效地通过Shapley解释估算特征重要性。载于 A. L. Martel、P. Abolmaesumi、D. Stoyanov、D. Mateus、M. A. Zuluaga、S. K. Zhou、D. Racoceanu 和 L. Joskowicz 编，医学图像计算与计算机辅助干预——*MICCAI 2020*，第792–801页，2020年。[53] V. Niennattrakul 和 C. A. Ratanamahatana。使用k均值和动态时间规整对多媒体时间序列数据进行聚类。载于*2007*年国际多媒体与普适工程会议（*MUE'07*），第733–738页，2007年。

[54] R. Tavenard、J. Faouzi、G. Vandewiele、F. Divo、G. Androz、C. Holtz、M. Payne、R. Yurchak、M. Russwurm、K. Kolar 和 E. Woods。Tslearn：用于时间序列数据的机器学习工具包。机器学习研究杂志，21(1):118，2020年。[55] C. J. F. Waaijer、C. Teelken、P. F. Wouters 和 I. C. M. van der Weijden。科学领域的竞争：发表压力、基金申请压力与学术就业市场之间的联系。《高等教育政策》，31:225–243，2017年。[56] A. Fuller 和 L. Unwin。学徒制的变化与延续：一种学习模式的韧性。《教育与工作杂志》，22(5):405–416，2009年。[57] A. Rosenfeld 和 O. Maksimov。年轻的计算机科学家是否应该停止与博士生导师合作？《*ACM*通讯》，65(10):66–72，2022年。[58] G. Abramo、C.A. D'Angelo 和 F. Di Costa。顶尖科学家合作行为的性别分析：来自意大利的证据。《科学计量学》，120:405–418，2019年。[59] P. J. Leman、A. P. Macedo、A. Bluschke、L. Hudson、C. Rawling 和 H. Wright。性别和族裔对儿童同伴合作的影响。《英国发展心理学杂志》，29(1):131–137，2011年。[60] R. Nunkoo、M. Thelwall、J. Ladsawut 和 S. Goolaup。三十年的旅游学术研究：性别、合作与研究方法。《旅游管理》，78:104056，2020年。[61] T. Lazebnik、S. Bunimovich-Mendrazitsky、S. Ashkenazi、E. Levner 和 A. Benis。利用健康传播及早发现并控制下一波流行病：人工智能工具的开发及其在美国COVID-19数据上的验证。《国际环境研究与公共卫生杂志》，19(23)，2022年。[62] E. Savchenko 和 T. Lazebnik。现代俄语文本的计算机辅助功能文体识别与校正。《数据、信息与管理杂志》，4:25–32，2022年。