补充信息：
全球科学领域的性别差异（评论*自然*504、211-213；2013）

Vincent Larivière、倪超群、Yves Gingras、Blaise Cronin 和 Cassidy R. Sugimoto

数据说明

科学网

该项目的数据取自汤森路透的 Web of Science 数据库，涵盖科学引文索引扩展版、社会科学引文索引以及艺术与人文学科引文索引。2008年至2012年的所有文章均纳入分析。原始数据被转换为位于加拿大蒙特利尔魁北克大学科学与技术观察站 (OST) 的 SQL 服务器上的关系数据库，以便执行各种分析。自 2008 年以来，Web of Science (WoS) 包含作者的完整名字，这允许对作者进行性别分类（请参阅下一节）。汤森路透还对每位作者的机构地址（机构、国家、城市等）进行索引，从而可以按性别对文章进行精确的地理分配。

本研究中提出的指标基于各性别作者发表的文章和评论文章的数量。其他类型的文档（例如社论、给编辑的信和书评）被排除在分析之外，因为它们通常未经同行评审，也不被视为对学术知识的原创贡献[1]。这些数字基于论文的分数计数：也就是说，每个作者都获得 1/*X*作者身份计数，其中*X*代表可以在给定论文上分配性别的作者数量。

引用量衡量的是给定论文从其出版年份到 2012 年底收到的所有引用。为了比较不同专业之间的数据，每篇文章的引用次数除以已发表的同一学科文章收到的平均引用次数那年[2,3]。当平均相对引用 (ARC) 高于 1 时，某篇特定文章的引用高于同一领域的世界平均水平。相反，ARC 低于 1 意味着引用次数低于世界平均水平。当然，众所周知的文献计量学局限性也适用于这种分析，因为 Web of Science 并未索引世界上所有的学术文献。这对于社会科学和人文学科来说问题更大，因为 (a) 除了期刊文章之外，媒体几乎没有报道研究成果[4](b) 以英语以外的语言撰写的文章形式的研究成果的覆盖范围非常有限[5]。

姓名 性别 分配

性别姓名信息列表

WoS 作者的性别信息是通过将姓名与通用和特定国家名单进行匹配来确定的。通用列表适用于整个 WoS 作者集，特定国家/地区列表适用于 WoS 作者子集

与相应的国家相关。下表 S1 显示了用于对作者性别进行分类的列表。

表 S1。性别姓名信息列表

列表	列出来源
美国人口普查	https://www.census.gov/genealogy/www/data/1990surnames/names_files.html
维基百科名称	http://wiki.name.com/en/Baby_Names
维基百科	http://en.wikipedia.org/wiki/Category:Given_names_by_gender
法语	http://en.wikipedia.org/wiki/French_name http://en.wikipedia.org/wiki/Category:French_feminine_given_names http://en.wikipedia.org/wiki/Category:French_masculine_given_names
魁北克人口普查	http://www.rrq.gouv.qc.ca/en/enfants/Pages/banque_prenoms.aspx
韩国	http://en.wikipedia.org/wiki/List_of_Korean_given_names http://en.wikipedia.org/wiki/Category:Korean_given_names
立陶宛	http://en.wikipedia.org/wiki/Lithuanian_name
波斯/伊朗	*http://www.top-100-baby-names-search.com/baby-names-persian.html*
罗马尼亚	http://en.wikipedia.org/wiki/Romanian_name http://en.wikipedia.org/wiki/Category:Romanian_given_names
巴西/葡萄牙	http://en.wikipedia.org/wiki/Brazilian_name#Brazilian_names
塞尔维亚	http://en.wikipedia.org/wiki/Serbian_name http://en.wikipedia.org/wiki/Slavic_names
乌克兰	http://en.wikipedia.org/wiki/Ukrainian_names http://en.wikipedia.org/wiki/Slavic_names http://www.top-100-baby-names-search.com/ukrainian-baby-names。 html
泰国	http://www.top-100-baby-names-search.com/thai-first-names.html
印度	http://en.wikipedia.org/wiki/Category:Indian_given_names http://www.studentsoftheworld.info/penpals/stats.php3?Pays=IND www.pkp.in/info/downloads/India%20Baby%20Names。 .xls
日本	http://en.wikipedia.org/wiki/Category: Japanese_given_names http://en.wikipedia.org/wiki/ Japanese_name

美国人口普查

美国人口普查提供了名字列表以及具有特定名字和相关性别的人口百分比。因此，根据从 WoS 数据获得的作者姓名，使用这些列表对每位作者可能的性别进行编码。如果一个名字同时用于两种性别，则只有当其中一种性别的使用频率至少是另一种性别的十倍时，才会将其归因于特定性别。否则它被归类为“男女皆宜”的名字。美国人口普查数据被用作该项目的主要来源，按性别对作者进行分类。其他通用列表仅用于无法使用美国人口普查列表进行分类的名称。

维基百科

该列表提供了 8,155 个女性和男性名字（非排他性）。这用于对与美国人口普查不匹配的姓名进行分类。与之前的程序一样，出现在两个列表中的名字都被归类为男女皆宜的。

维基百科

维基百科的名字列表提供了与 60 多个国家/地区相关的名称。该列表用于对未使用美国人口普查数据和维基名称成功分类的作者进行分类。

魁北克省和法国

这是按性别列出的加拿大大学教授姓名列表，以及按性别列出的魁北克新生儿列表。此列表中的所有非英语欧盟字符都转换为相应的基本英语字符，以便与 WoS 作者集匹配（WoS 提供英文名称）。

韩国

使用一系列规则来匹配通用列表中未匹配的韩国名字。例如，以 -jae 结尾的名字通常是男性名字，而以 -miare 结尾的名字通常与女性名字相关。

立陶宛

还对其余立陶宛名字应用了基于规则的方法：女性名字通常以：-a、-e 或 -ia 结尾；啤酒名称通常以：-s、-as、-is、-ys 结尾，
- 我们和 ius。

日本

规则还用于对剩余的日本名称进行分类。女性名字通常以以下结尾：-a、-chi、-e、-ho、-i、-ka、-ki、-ko、-mi、-na、-no、-o、-ri、-sa、-ya ， 和
-哟。男性名字通常以以下结尾：-aki、-fumi、-go、-haru、-hei、-hiko、-hisa、-hide、
- hiro、-ji、-kazu、-ki、-ma、-masa、-michi、-mitsu、-nari、-nobu、-nori、-o、-rou、-shi、
- shige、-suke、-ta、-taka、-to、-toshi、-tomo、-ya 和 -zou。

俄罗斯及相关国家

以前的作业是基于名字的。然而，对于俄罗斯名字，也使用姓氏。男性的姓氏通常以 -ov、-ev 或 -in 结尾。女性通常以 -ova、-eva 或 -ina 结尾。因此，这些“后缀”适用于俄罗斯作者，以及其他国家，其中 95% 或以上已指定的女性或男性名字以上述后缀之一结尾（捷克共和国、保加利亚、拉脱维亚、哈萨克斯坦、乌兹别克斯坦、立陶宛）和卢森堡）。

波斯语/伊朗、巴西、罗马尼亚、葡萄牙、塞尔维亚、乌克兰、泰国和印度

对于伊朗、巴西、罗马尼亚、葡萄牙、塞尔维亚、乌克兰、泰国和印度，我们根据网上获得的信息编制了每个县的具体姓名和性别列表。编制国家特定列表和命名规则所使用的来源请参阅表 1。

中国

中国有 84,462 个与隶属关系相关的唯一作者姓名，对应于 1,841,748 个作者。唯一作者姓名的作者数量分布遵循幂律分布。也就是说，大多数作者姓名与少量论文相关联，而少数作者姓名与大量论文相关联。其中，与20篇及以上论文相关的作者姓名有12,828个（占总数的15.17%），约占中国总作者的84.25%。因此，我们选择了与至少 20 篇论文相关的作者姓名，并手动分配每个姓名的性别。两名来自中国的母语人士手动编码了这些名字。他们编码

根据他们对中文和中文名字的了解，确定每个名字的性别。网络搜索也被用于模糊的情况，例如，谷歌图片中出现的主要性别，以及与各种 Facebook 帐户相关的性别。

台湾

将中文名字翻译成英文并没有统一的拼音系统——台湾人可以从四种不同的拼音系统中选择一种。因此，我们的作业包括：1）查找用于将名字翻译成英文的拼音系统；2）与注音fuaho比较

(http://www.boca.gov.tw/content?CuItem=5609&mp=1) 以确定正确的标点符号。如果该名字不在拼音系统中，则被标记为未知。如果在系统中，则使用发音来确定性别（由母语人士评估）。任何被认为不明确的名称都被标记为未知。

方法

瓦氧作者姓名预处理

作者姓名列表包含由 WoS 索引的作者的名字。名字是在单独的字段中提供的，但不是以统一的形式。有些名字是缩写而不是完整的名字，或者包含特殊字符，如“()”、“-”、“.”。或一个空间。为了与上面介绍的源列表匹配，对作者姓名集进行了如下预处理：

提取名字中“()”中的所有字符并将其视为昵称；

- 识别缩写：
  - 计算“.” 以给定的名称：
    - 如果没有“.”，则计算空格
    - 如果有“.”，则计算整个字符串的长度 哦
      如果字符串长度小于“.”个数的3倍，则视为首字母。
      哦 如果没有：继续下一步
- 对于不是首字母缩写的名字，用空格将名字分成几个部分；
- 将每个部分中的所有连字符替换为空格：例如，“Jean-Pierre”将转换为“Jean Pierre”。

应该指出的是，我们确定的是作者身份，而不是个人——也就是说，我们有兴趣确定每个作者的性别，但不关心论文之间的匹配作者。也就是说，我们感兴趣的是每篇特定论文中每位作者的性别，而不是该作者撰写了多少篇论文。我们的分析是在总体层面上进行的——有多少篇论文有女性或男性作者，而不仅仅是每个女性或男性作者撰写了多少篇论文。

与性别名单匹配

如上所述，作者名集与通用列表和国家特定列表相匹配，以确定 WoS 作者的性别。比赛按照以下顺序进行：

- 美国人口普查

- 维基百科名称
- 维基百科
- 法国和魁北克名单
- 其他特定国家/地区的列表

美国人口普查名单被用作性别信息的基本来源。因此，所有其他列表（特定于国家/地区的列表除外）仅用于匹配美国人口普查无法匹配的名字。

世界和国家层面的覆盖范围

经过这些步骤后，我们成功地为 56.1% 的不同名字（例如 John、Linda）和 59.5% 的不同完整作者姓名（例如 John Smith、Linda Madden）分配性别（女性或男性（F 或 M））（见表S2）。很大一部分作者姓名仅提供缩写（不同作者姓名的 31.0%）。因此，就提供首字母以外的名字信息的作者百分比而言，性别被分配给 57.3% 的不同名字和 83.0% 的不同全名。

表S2。全名和指定性别的名字的数量和百分比。

性别	全名			姓		
	氮	占全部的%	%（全部 - 缩写）	氮	占全部的%	%（全部 - 缩写）
女性	1,194,340	25.0%	35.0%	209,737	25.3%	25.8%
男性	1,642,066	34.4%	48.1%	256,166	30.8%	31.5%
男女通用的	123,023	2.6%	3.6%	23,919	2.9%	2.9%
未知	456,020	9.6%	13.4%	323,687	39.0%	39.8%
缩写	1,354,802	28.4%	-	16,945	2.0%	-
全部	4,770,251	100.0%	-	830,454	100.0%	-

在不同论文和论文作者的层面（例如文章署名中出现的每位作者的总和），结果是相似的（表S3）。81.3% 的论文至少有一位作者指定了性别，65.2% 的作者论文组合指定了性别。当排除只有姓名首字母的作者时，这一比例会增加到 86.1%。

表 S3。不同论文和指定性别的作者论文的数量和百分比。

性别	不同的论文		作者-论文组合		
	氮	占全部的%	氮	占全部 %（全部 - 缩写）	
女性	2,750,850	50.2%	5,546,226	20.3%	26.8%
男性	4,116,595	75.1%	12,264,088	44.9%	59.3%
任何性别	4,458,622	81.3%	17,810,314	65.2%	86.2%
男女通用的	496,825	9.1%	563,954	2.1%	2.7%
未知	1,542,186	28.1%	2,298,439	8.4%	11.1%
缩写	1,153,640	21.0%	6,657,208	24.4%	-
N篇论文	5,483,841	100.0%	27,329,915	100.0%	-

表 S4（完整提供于 http://dx.doi.org/10.1038/504211a）显示了属于每个类别的不同作者和名字的数量，以及百分比（占全部和全部减去首字母缩写）那些被分配了性别的人，而表 S5 (http://dx.doi.org/10.1038/504211a) 则对不同论文和论文作者组合提供了相同的衡量标准。尽管不尽相同，但不同国家在作者和论文分配比例方面的覆盖范围大体处于同一范围内。

验证研究

为了评估我们分析的准确性，我们随机选择了 1,000 条记录，代表一位作者，该作者被分为以下五个类别：姓名首字母、未知、男女皆宜、男性和女性。这些作者与特定的国家、机构以及在某些情况下的电子邮件地址相关联。此信息用于在网络上查找传记信息或照片，可用于验证分类的准确性。可以识别性别的随机样本的百分比因类别而异（见表 S6），并且取决于许多变量，包括作者的身份。例如，在男性类别中，许多作者是技术人员和工作人员，他们缺乏冗长的传记信息（其中包含代词）或照片。

表 S6。每个类别中男性和女性的百分比

类别	确定的 # 和 %	（已确定的）女性人数和百分比	男性（已确定的）的数量和百分比
缩写	839 (83.9%)	198 (23.6%)	641 (76.4%)
未知	890* (89.0%)	282 (31.7%)	607 (68.2%)
男女通用的	540 (54.0%)	113 (20.9%)	427 (79.1%)
男性	605 (60.5%)	10 (1.7%)	595 (98.3%)
女性	830 (83.0%)	720 (86.7%)	110 (13.3%)

* 由于一位作者自我认定为“其他”，因此此处的数字不是男性和女性的总和。因此，他们既不是男性，也不是女性，也不是身份不明。

数据分析和可视化

R是主要的数据分析和可视化工具，ArcGIS用于显示北美详细信息。还使用了 Tableau 软件和数据驱动文档 (D3) JavaScript 库，主要用于可视化的交互式版本。

最初根据每份出版物提供的作者地址信息从 WoS 中提取了 206 个国家/地区的列表。在对生产力、合作和影响进行分析时，排除了在研究时间段内出版物少于 20 份的国家名单，以减少因样本数量少而可能造成的扭曲。使用 WoS 数据库提供的国家名称（英文）。在制作全球地图时，名称由国际标准化组织 (ISO) 提供*3166标准*被用来代替 WoS 名称。例如，刚果民主共和国在WoS数据库中是扎伊尔，正式指的是1971年至1997年间存在的国家。韩国是WoS中的名称，而根据ISO应该是大韩民国*3122标准*。对于每个国家，出版物的数量及其相应的引用是通过国家层面的汇总获得的。

使用世界地图作为底图，使用 D3 库按国家显示女性和男性研究成果的差异。世界地图和学科地图中按性别统计的论文数量是基于作者身份的细分，即通过统计每篇论文的男性和女性作者的比例来获得。因此，对于一篇有 8 位作者的论文，其中性别可以分配给 6 位作者，每个作者及其相应的性别都被分配为论文的 1/6（无法分配性别的作者被排除在分母之外）。然后，这些性别比例在国家和学科层面进行汇总，并作为世界地图和学科地图中 F-M 比率的基础。每个国家根据女性和男性研究成果的差异进行颜色编码：一个国家越蓝，该国家的男性与女性研究成果越高；一个国家的颜色越深，该国女性与男性的研究产出就越高。值得注意的是，有些国家在我们的 2008-12 年 WoS 数据集中没有任何出版记录。这些国家在地理地图上被涂成灰色。对美国各州和加拿大各省也进行了类似的分析。

第554章 不服输的人*加州大学圣地亚哥分校科学地图*学科类别，在该项目中近似为学科/专业。与世界地图一样，这些比率是根据分散的作者身份编制的（见上文）。因此，男性和女性研究成果之间的差异是通过将每个学科中女性的部分作者权总和除以男性的作者权总和来计算的。为了直观地显示不同学科的差异，使用 UCSD 科学地图作为底图，使用 D3 库将每个学科中每个性别的研究成果的差异叠加在底图之上。每个学科（地图上的一个节点）根据差异值进行着色：节点越蓝，则相应学科中男性越活跃；节点越蓝，则表明男性在相应学科中越活跃；节点越橙色，表示女性在相应学科中越活跃。

还对国际和国家层面的女性和男性合作模式进行了调查。在这个项目中，一个国家的女（男）作者的国际合作率计算为女（男）作者与另一个国家的其他人合作完成的论文数量除以该国至少有一篇论文的数量署名中的女（男）作者。同样，一个国家的女性（男性）作者的国家合作率的计算方法是，女性（男性）作者与同一国家的其他人合作完成的论文数量除以该国家至少有一名女性的论文数量（男）署名作者。这里采用条形图来按国家显示女性和男性的国际和国内合作情况。（在线互动版本：http://dx.doi.org/10.1038/504211a，其中国家按女性国家合作率降序排列。）

构建了热图来显示不同作者类别的出版物影响力的差异。需要注意的是，引用量是在开放的引用窗口下计算的，并以同一专业当年发表的论文的平均引用率进行归一化。这里的热图是每个国家在不同出版物类别中的影响的可视化，即按出版物类别划分的国家矩阵，每个交叉单元格显示引用计数。引用计数使用红-白-绿不同的调色板进行颜色编码，每种颜色对应于最小-中值-最大引用数。即颜色越红，引用次数越少；颜色越绿，收到的引用越多。

1.莫德，高频*科学计量学***35**，177-191（1996）。
2. Moed, HF、De Bruin, RE 和 van Leeuwen, TH。N。*科学计量学***33**，381-422（1995）。
3. 舒伯特，A. 和布劳恩，T.*科学计量学***9**，281-291（1986）。
4. Larivière, V.、Archambault, É.、Gingras, Y. 和 Vignola-Gagné, É.。*J. Am. 苏克。信息。科学。技术。***57**，997–1004（2006）。
5.阿尔尚博，É。Vignola-Gagné, É.、Côté, G.、Larivière, V. 和 Gingras, Y. *科学计量学***68**，329–342（2006）。