



基于多主体建模的科研人员跨国流动研究：以人工智能领域为例

分享人

陈文杰 清华大学经济管理学院 charleschen01@foxmail.com
闵超 南京大学信息管理学院 mc@nju.edu.cn

目录

- 01 研究背景**
- 02 数据与模型**
- 03 模型分析**
- 04 总结与展望**



知识经济时代，**科研人员**作为科技创新的核心要素和科学知识的重要载体，具备**高度的流动性**，也在不断地突破机构、地域和国家的限制，更加自由地流动。



科研人员之于经济增长的作用日趋显著，作为一种战略资源受到**各国高度重视**，各种政策出台，吸引人才流入（美国- 改革H-1B签证；英国- 全球人才签证；我国-“人才是第一资源”）



研究并掌握科研人员跨国流动趋势、剖析流动机制与流动动因，有利于国家调整优化人才管理制度和推行精准引才措施，对于聚集全球范围内的创新要素、提升人力资源配置效率具有重要的现实意义



国内外研究现状

学术界现有关于科研人员跨国流动的**主要研究方法**:

- (1) 基于**履历分析法**和**文献计量法**, 从ORCID、Scopus等学术数据库收集科研人员的CV资料或科研产出数据, 去描述性统计分析科研人员的流动模式;
- (2) 借鉴**计量经济学方法**, 引入经济和制度相关的外部数据, 实证科研人员跨国流动的影响因素;
- (3) 通过问卷调查、专家访谈等**定性方法**探索人才流动的相关议题。

存在的**研究空白**:

- 其一, 大多基于已有数据描述分析科研人员跨国流动的现状和模式, 较为缺少针对未来人才流动进行**定量预测**的研究成果;
- 其二, 当下对科研人员流动的研究大多是静态的描述, 事实上人员流动模式的演变是阶段性, 甚至是连续性的, **欠缺动态视角的观察**;
- 其三, 过往研究多聚焦于群体画像的描绘和聚合统计, 但学术流动本质上是科研人才的**个体行为**, 现有研究缺乏对于流动主体自身决策机制以及主体同主体间互动关系的研究分析。



ABM相关研究

Agent-Based Modeling (多主体建模)

➤ ABM则将元胞自动机的思想引入了社会科学研究领域，在ABM的视角下，每个**主体**都具有自主决策能力、信息处理能力和行为规则，复杂系统所呈现的**宏观现象**来源于相互独立的个体之间在非线性的互动关系下涌现出的行为模式或目标准则。 (eg. Shelling 种族隔离模型)

Agent-Based Modeling 与人口流动

➤ 研究城乡人口流动和跨国移民：Fu和Hao-中国城乡人口流动模型；Kniveton-气候变化下的人口迁移模型...
➤ 探讨人才流动相关课题：Vaccario-科学家跨城市流动模型；Furkan-人才流失模型；Biondo-人才回流模型；Amorim-Lopes-医生迁移模型...

ABM方法为研究人才流动提供了新的解决方案，然而相关研究还较少，已有的模型较抽象，论证了方法的可行性，但是在实用性上仍有不足；其发轫于国外，在国内还属于研究空白，



本研究的工作

研究概述

- 着眼于科研人员个体流动，综合计划行为理论和理性选择理论，提出科研人员三阶段流动决策机制，利用**多主体建模方法**搭建科研人员流动模型（Global Researchers Flow Model, GRFM），以**人工智能领域科研人员**为主要研究对象，引入实证数据进行模型校准。

研究问题

- 全球人工智能领域科研人员的个体属性和流动特征是怎样的？
- 科研人员流动受到哪些因素的影响及其跨国流动决策机制？
- 科研人员的个体决策行为如何涌现出复杂的跨国流动现象？如何利用ABM方法在模型世界复刻其全球流动？

研究创新点

- 理论视角创新：宏观层面的流动研究 -> 个体流动的心理行为与决策机制
- 研究方法创新：科学计量分析方法获取真实世界的跨国流动网络 & 设计并搭建ABM模型模拟科学家的全球流动

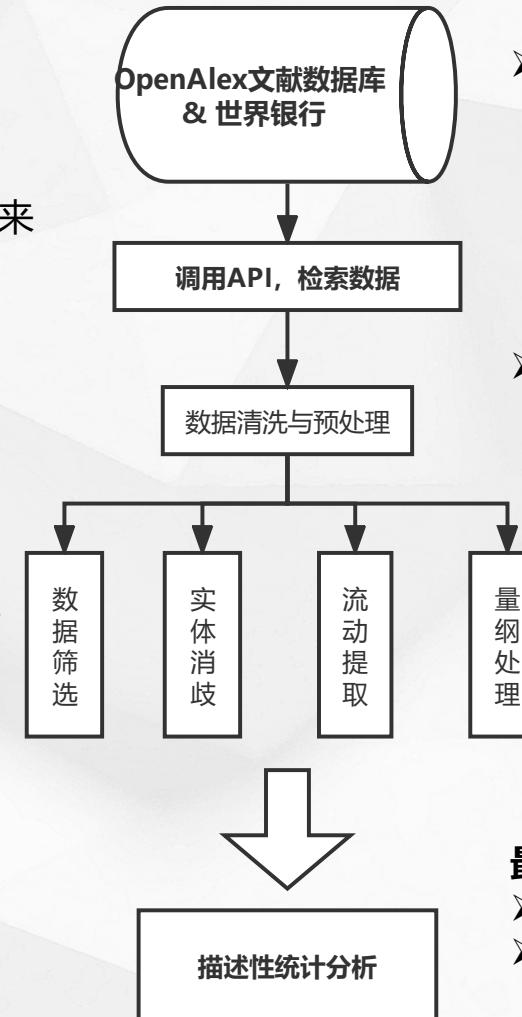


数据

- 根据CCF划定的人工智能领域A类期刊和会议来筛选人工智能领域科研人员作为研究样本

- 基于“自底向上”的归并思想进行**作者实体消歧**

主要考虑了作者机构、合作者、年份冲突等因素，识别出歧义作者6616位，对应32638篇文献。



- OpenAlex继承了MAG的全量数据，是一个由文献、作者、机构等五类实体构成的异构有向图文献数据库

- 将时间限定在2000~2020年，初步检索到39013位作者，及其对应的718094条文献发布记录

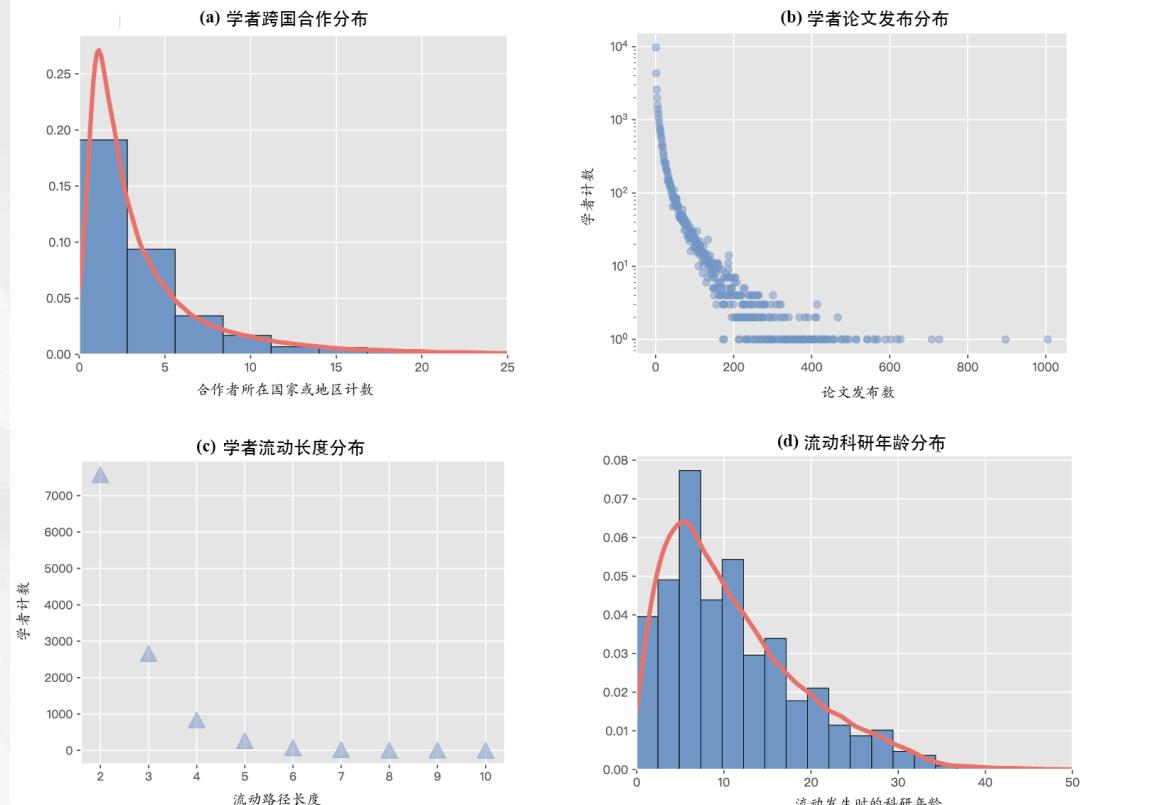
	作者ID	流动年份	流动起始地	流动目的地
1	A2250875712	2013	IN	GB
2	A2043954528	2009	US	KR
3	A2418460153	2017	CN	US
		...		

最终数据集包括：

- 39013位具有唯一标识的作者和685456条文献记录
- 54554条流动记录，涉及11443位AI学者，人均流动4.77次，整体流动率约29.3%

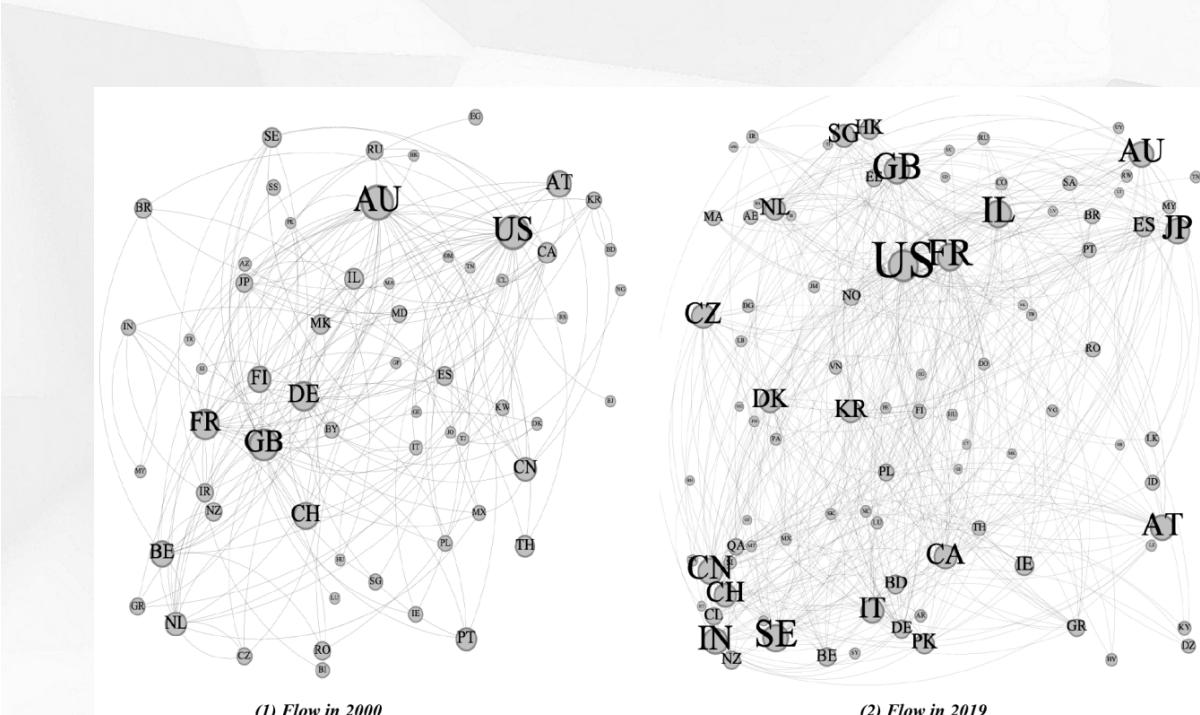


数据探索性分析



人工智能领域科研人员的个体特征和流动特征

- 学术产出符合幂律分布；整体高合作倾向；流动年龄偏态分布...



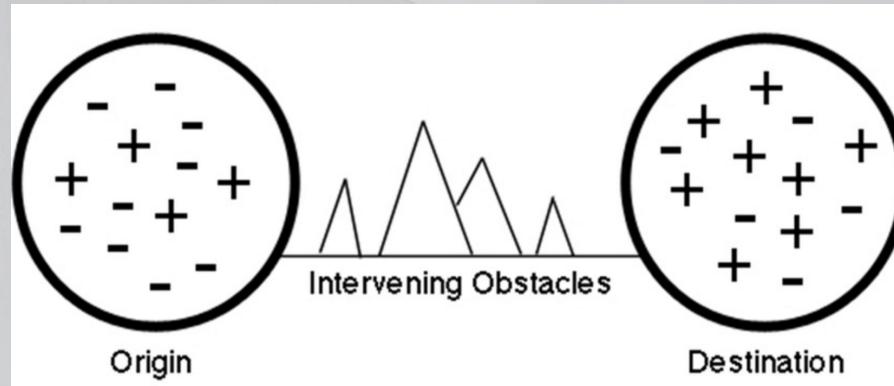
2000和2019年科研人员流动网络



理论基础

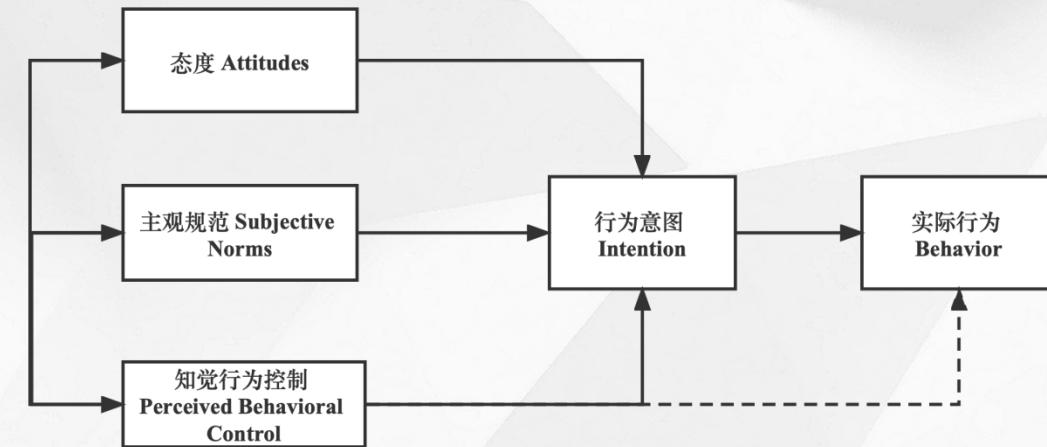
理性选择理论

- 迁移被视为一种理性行为，人们是在个体利益最大化的基础上，通过权衡成本和收益来进行决策，理论的核心是预期效用模型，用于**解释移民的选择性**
- Ravenstein-人口流动七定律；Lee Push-Pull theory



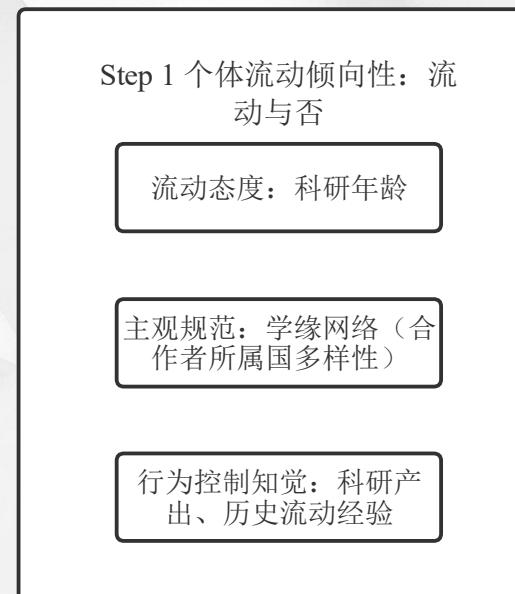
计划行为理论

- 态度、主观规范与行为控制知觉三个变项共同决定个人的行为意图。
- 计划行为理论可以被用来描述个人做出迁移决策的心理推理机制 (eg. Grothmann 个体响应气候变动的心理感知过程)





科研人员流动概念模型



Step2 国家相对吸引力：流向偏好

国内生产总值
失业率
科研实力
地理距离
流动惯性

确定流动目的地
偏好清单

Flow Preference List
Target Country1
Target Country2
Target Country3
Target Country4
...

Step3 流动匹配模型：
最终去向

Country's List:
Researcher1[accepted]
Researcher2[accepted]
Researcher3[accepted]
...
ResearcherN[rejected]

➤ 最后，搭建流动匹配模型来确定科研人员是否流动成功以及最终流向的国家或地区。

- 首先，按照计划行为理论分析科研人员的流动倾向性，每个个体都会根据自身对于流动行为的态度、受到的主观规范和行为控制知觉来考虑流动与否，超过一定阈值的个体将会进行流动。

- 其次，基于推拉理论来判定不同国家对于个体的相对吸引力，决定流动的个体都会得到一份流动目的地偏好清单，旨在回答具备一定流动倾向性的科研人员会选择流向何处的问题。



全球科学家跨国流动ABM模型 (GRFM)

遵循 ODD (Overview, Design concepts, and Details) 协议设计科研人员流动模型 (GRFM) 的架构 -> 描述清晰，易于复现，增强模型科学性和可靠性。

1. 模型实体，实体属性和时空尺度

- 两类Agent: Country & Researcher
- 将 $t = 0$ 视作2000年，模型校准阶段设定时间上限为2020，空间涉及到的142个国家或地区。

表1：科研人员流动模型的主体属性特征

Agent类型	属性名称	属性说明
Researcher	$current_Country(t)_i$	Researcher _i 于第t年在模型世界所处的国家或地区
	$age(t)_i$	Researcher _i 在第t年时的科研年龄
	$strength(t)_i$	Researcher _i 的科研产出，利用该学者累计发布的论文数量衡量
	$co_workers_diversity(t)_i$	截至第t年，Researcher _i 合作者所属国别多样性
	$flow_history(t)_i$	截至第t年，Researcher _i 的历史流动记录
	$flow_CD(t)_i$	Researcher _i 当前的流动冷却期
Location	$flow_intention(t)_i$	Researcher _i 的个体流动倾向
	$Country_j$	模拟世界中Country _j 所处的位置，是由其首都经度组成的二元组，与现实地理位置保持一致，用于后续迁移距离的计算
	$N(t)_j$	Country _j 在第t年所拥有的Researcher数量。
	$GDP(t)_j$	国内生产总值GDP，用于描述Country _j 的经济实力
	$Strength(t)_j$	Country _j 的科研实力，统计方法为该Country _j 中所有Researcher的strength汇总并求取均值
	$Volume(t)_j$	Country _j 的Researcher容量上限，设定为实证数据中同年的学者数
Country	$Unemployment(t)_j$	失业率，用于描述社会治安状况
	$Centrality_i$	Country _j 在全球科研人员流动网络中的重要程度，基于社会网络分析的特征向量中心性指标进行计算



全球科学家跨国流动ABM模型 (GRFM)

2. 模型运作流程

每一时间步t的模型运作流程概览如下：

- a. 所有Researcher的科研年龄*age*加1；
- b. 部分Researcher退出模型世界，一定数量的Researcher诞生；
- c. 更新所有Researcher的科研产出、合作者所属国多样性，据此计算其个体流动倾向性，并结合流动冷却机制与流动冷却期判定该Researcher是否会做出流动决策。
- d. 根据实证数据更新所有Country的*GDP*和*Unemployment*属性，同时根据模型内数据更新Country的平均科研实力属性；
- e. 计算Researcher的流动目的地偏好清单，Country对Researcher的科研实力进行评估并结合自身当前的容量决定是否允许迁入，最终完成流动匹配；
- f. 根据匹配结果，更新参与流动的Researcher的所在国家、流动冷却期属性，Country的学者数量N、科研实力及中心性。



全球科学家跨国流动ABM模型 (GRFM)

3. 子模型设计

(1) 科研人员的诞生和退出

- 假设科研人员在每个时间步都有一定概率 p_e 退出的模型世界
- 借用Logistic增长模型来模拟科研人员的增长

$$N(t)_{logistic} = \frac{KN_0e^{rt}}{K+N_0(e^{rt}-1)}$$

(2) 科研产出与合作者多样性的更新

- 采用泊松分布作为科研人员在第t时间步新增科研产出的概率分布，其期望值是国家或地区的每年人均新增发文量

$$strength(t+1)_i = strength(t)_i + Poisson(\lambda_{loc}) \quad (4-5)$$

- 合作者多样性受到其所在Country的中心性Centrality的影响

$$\begin{aligned} co_worker_diversity(t+1)_i &= co_worker_diversity(t)_i + \\ &Centrality_{loc} \cdot \eta \end{aligned} \quad (4-6)$$



全球科学家跨国流动ABM模型 (GRFM)

3. 子模型设计

(3) 流动决策与流动匹配

➤ 阶段一：个体流动倾向性计算

训练随机森林模型 H

INPUT: [流动态度、合作者所属国多样性、科研产出和历史流动次数]

OUTPUT: 流动倾向性 flow_intention

➤ 阶段二：个人流向目的地偏好计算

综合考虑流动惯性、社会经济实力、科研实力、地理距离的影响

$$\begin{aligned} \text{flow_preference}_j = & \gamma * \Theta(\text{Country}_j, \text{flow_history}_i) + \\ & [\text{Strength}_j, (\text{GDP}_j - \text{Unemployment}_j), \text{Geo_Distance}_{j, \text{loc}}] \cdot \vec{\alpha} \quad (4-8) \end{aligned}$$

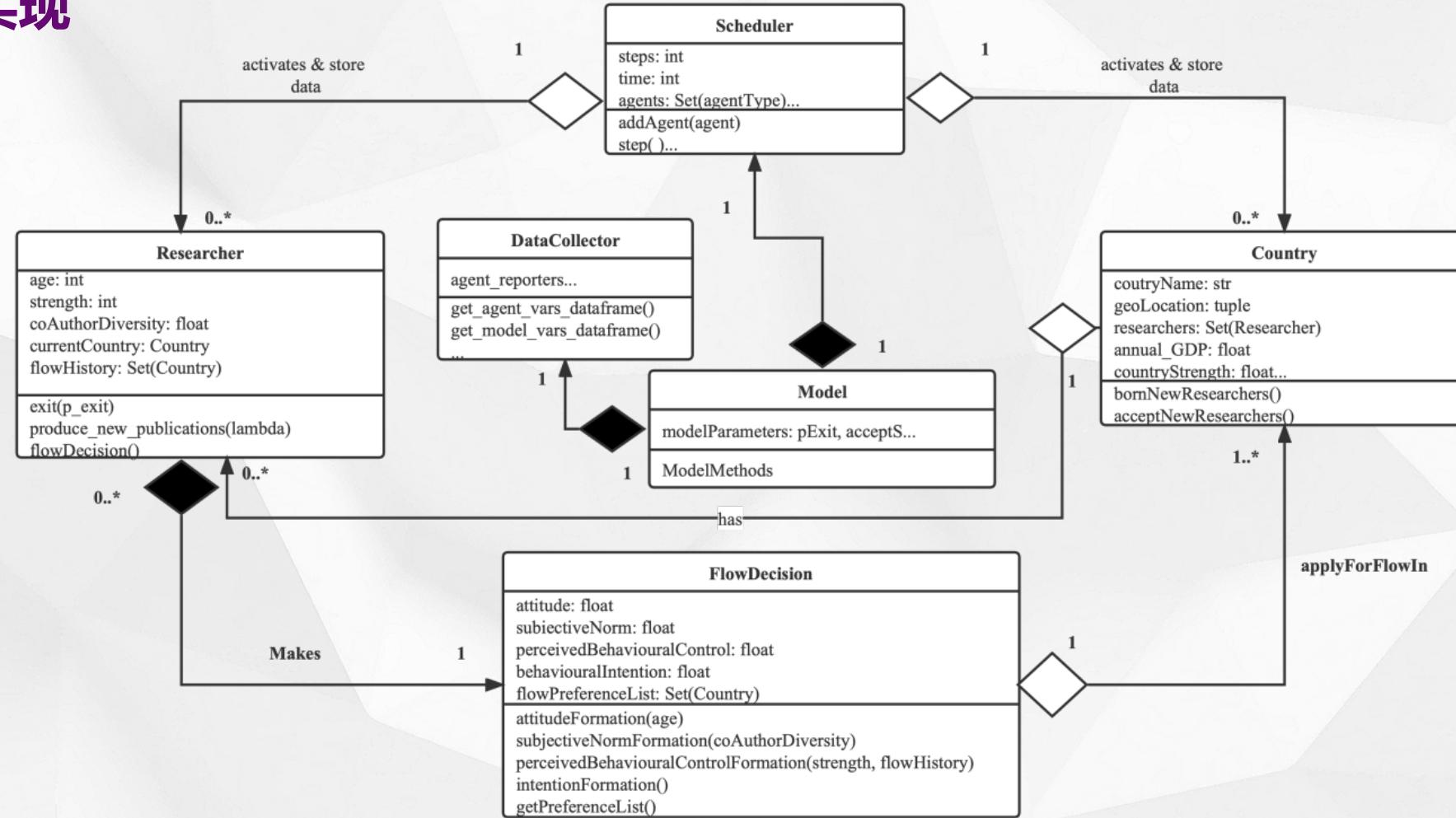
➤ 阶段三：流动匹配

优先接纳科研产出较多的学者，其次以一定概率尚处于发育期的学者

$$p(j, i) = \begin{cases} 1, & \text{if } \text{strength}(t)_i \geq \text{Strength}(t)_j \\ \left(\frac{\text{strength}(t)_i}{\text{Strength}(t)_j} \right)^{\text{accept_s}}, & \text{otherwise} \end{cases} \quad (4-9)$$



模型实现



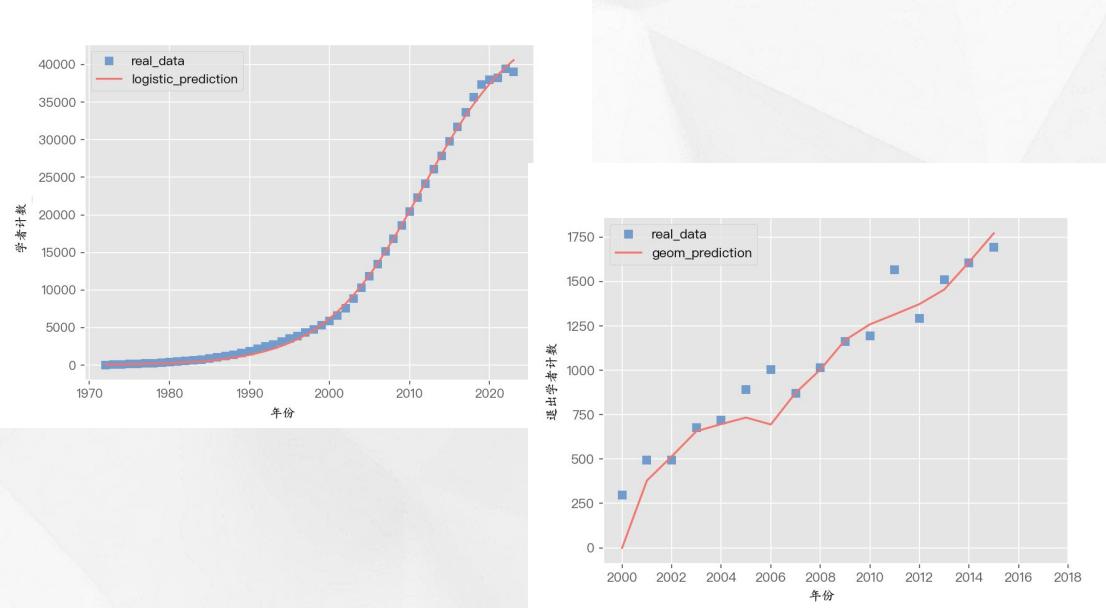
所用工具：python-Mesa



参数拟合与模型校准

1. 基于实证数据的参数估计

(1) 科研人员的诞生和退出



(2) 流动倾向性计算模型

- 对比各类分类模型的适用场景和实际效果后，最终决定选用随机森林模型
- 类别不平衡影响Recall -> 重采样技术，模型精确度虽然有所降低，但是整体在召回率上得到较大提升，满足本研究预测流动发生事件的需求

	未做特殊处理			重采样		
	LR	DC	RF	LR	DC	RF
Accuracy	94.84%	94.04%	94.43%	69.58%	73.72%	71.38%
Precision	34.29%	23.21%	23.92%	10.04%	13.06%	19.73%
Recall	1.79%	7.59%	4.45%	62.74%	73.89%	77.22%
F1 score	0.034	0.114	0.075	0.173	0.221	0.314

容量K = 46893.51; 初始值P₀ = 75; 生长率r = 0.1627

$$p_e^{opt} = 0.05$$



参数拟合与模型校准

2. 基于仿真实验的模型校准

- 难以通过实证数据观测，在系统仿真的过程中对部分参数进行校准

参数	参数解释	最优取值	参数空间
accept_s	体现Country对于科研实力较弱的Researcher的接受程度	0.5	{0, 0.5, 1, 2.5, 5, 7.5, 10}
γ	流动惯性因子	0.4	
α_1	科研水平偏好因子	0.025	
α_2	社会经济偏好因子	0.3	{0.0 , 0.05, 0.1, 0.25, 0.5, 0.75, 1.0}
α_3	地理距离偏好因子	0.05	

- 模型优化指标：Common Part of Commuters (CPC)

$$CPC = \frac{2 \sum_{i,j} \min(y^m(C_i, C_j), y^r(C_i, C_j))}{\sum_{i,j} y^m(C_i, C_j) + \sum_{i,j} y^r(C_i, C_j)} \quad (5-1)$$



模型评价

		来源国																		来源国																	
		US	CN	GB	DE	FR	JP	CA	IT	IN	NL	CH	AU	IL	ES	KR	US	CN	GB	FR	DE	IN	JP	IT	CA	AU	NL	SG	ES	IL	CH						
国 家 名 称	US	2,512	2,019	1,877	1,340	972	883	819	523	545	545	536	481	354	326	US	2,367	1,739	679	1,086	559	796	386	450	223	409	213	208	742	367							
	CN	2,211		680	246	357	131	248	100	133	180	181	177	166	129	81	CN	2,082		317	558	136	29	296	46	152	223	37	316	35	22	29					
	GB	1,365	589		246	366	111	137	81	123	130	122	135	123	87	24	GB	1,814	322		326	326	55	55	116	203	145	24	136	154	105						
	DE	2,059	446	277		181	84	89	73	56	76	90	75	64	63	24	FR	629	99	271		388	79	52	158	220	29	21	15	54	38	194					
	FR	1,104	449	465	171		115	85	73	67	43	34	34	44	35	18	DE	1,626	139	308	199		64	81	79	83	54	78	19	57	63	192					
	JP	1,221	156	180	101	74		44	29	32	23	30	26	19	17	11	IN	445	22	48	77	47		28	85	22	13	30	14	6	2	10					
	CA	979	149	112	109	86	96		64	16	19	12	11	21	10	21	JP	799	299	86	59	79	10		9	35	21	6	17	15	1	21					
	IT	847	148	136	92	77	62	42		27	26	21	20	27	16	19	IT	351	41	119	148	66	35	7		17	23	44	3	65	4	62					
	IN	632	128	45	74	40	54	23	49		13	9	6	3	2	13	CA	604	160	97	234	95	29	32	15		18	26	12	24	48	24					
	NL	367	77	43	52	28	43	29	30	12		10	9	9	2	5	AU	204	243	190	27	55	31	25	29	21		24	56	13	4	29					
	CH	299	68	44	78	32	31	15	27	13	6		5	6	4	11	NL	394	30	137	19	86	7	7	50	23	24		10	15	7	89					
	AU	329	62	33	43	11	39	12	26	16	3	3		1		4	SG	202	321	22	16	18	25	28		11	52	31		3	2	8					
	IL	204	48	30	43	22	20	15	29	2	4	7	1		3	2	ES	194	14	101	65	42	7	14	70	18	9	13	4		2	28					
	KR	124	24	11	16	6	15	9	11	15	2	1	2	1		1	IL	729	38	108	46	65	6	2	3	43	8	7	3	5		20					
	ES	119	14	14	15	13	4	5	5		2	1					CH	338	35	48	136	228	15	39	73	25	25	79	8	33	14						

(1) 模型仿真结果

(2) 真实世界数据



鲁棒性分析

1. 极端条件分析

参数	参数取值	取值说明	结果分析
accept_s	0	准入门槛较低	涌入到头部若干国家
accept_s	10	准入门槛较高	中上游国家迎来流入
流动惯性因子 γ	0	不考虑历史流动经验	流入头部国家且不再回流
流动惯性因子 γ	1	完全根据历史流动经历	未见显著变动

2. 局部敏感性分析

- 检查模型对参数值微小变化的敏感性
- 对参数值的微小变化不太敏感，模型具备一定的健壮性

	参数基准值	CPC+	CPC-	CPC基准值
accept_s	0.5	0.5110	0.5215	
γ	0.4	0.4914	0.4914	
α_1	0.25	0.5107	0.5228	
α_2	0.35	0.5211	0.5319	
α_3	0.05	0.5227	0.5301	0.5229



1. 对研究问题的回应

- 科研人员的流动是一种复杂的社会现象，还应该透过现象看到本质，即关注科研人员个体的流动决策机制，从而对流动做出更好的诠释
- **科研人员个体跨国流动的三阶段概念模型：**①流动与否？②流向何方？③流动结果？
- 利用Python语言在Mesa多主体建模库的基础上搭建了完整的**全球科学家跨国流动ABM模型（GRFM）**，证明了方法的可行性和分析框架设计的合理性

2. 研究展望

- 引入更多的外部数据集，例如Scopus、Aminer、ORCID等，进行多方验证。
- 不同科研领域因其学科属性不同，科研人员的流动模式可能会存在较大的差异，扩展到其他学科领域
- 针对一些小概率外部事件的影响缺少响应机制（例如新冠疫情）
- 部分操作指标也有待后续的设计优化



3. 管理启示

- **深耕科研人员流动治理，开展国际合作拓展智库交流。**本文以人工智能领域为例，通过学术论文数据识别领域人才，关注其个体属性以及历史流动记录，发现科研人员的跨国流动已经成为一股热潮。政府和相关机构应重视跨国科研人员流动，引导科研人员的跨国流动朝着良性健康的方向发展，实现国际社会的互惠共赢。
- **建立全球科研人才网络，精准引才实现国家战略协同。**当前人类已进入信息时代，各国政府及相关人才机构可以基于科学、技术和产业等数据搭建科研人员数据仓库和全球性科研人才网络，立足国家发展战略，从海量人才中找到最合适的人才，从而实现对海外人才“雷达式扫描”和“精准式引进”。
- **完善科研人员进出政策，厚植制度优势吸引人才流入。**各国政府应坚持以人为本的价值导向，优化全球科研人员的流入流出政策以实现人才自由流动，此外还可以建立创新型国际人才社区，打造人才“流入-聚集-进一步流入”的良性循环。
- **推出人才分层评价体系，科学锚定海外目标人才群体。**在本文提出的科研人员三阶段流动模型中，科研人员的跨国流动最终是个体和国家之间的双向选择。各国政府构建人才评估指标体系，既兼顾中高端人才的引进，也考虑处于职业发展初期的高潜力科研人员的引入与培植，做到精准引才下的优质育才。



谢谢大家！
恳请批评指正

分享人
陈文杰 清华经管 charleschen01@foxmail.com
闵超 南大信管 mc@nju.edu.cn