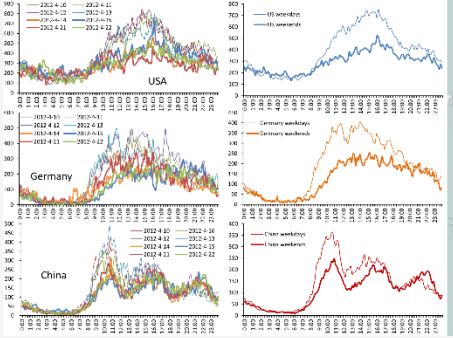
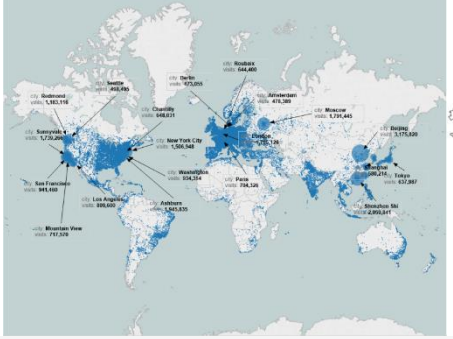
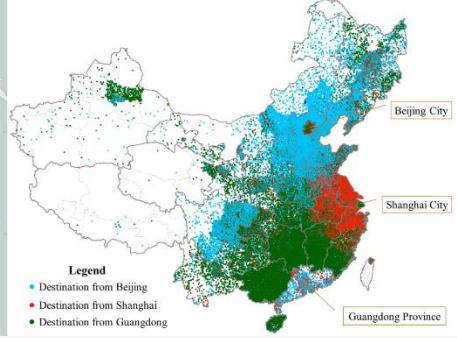
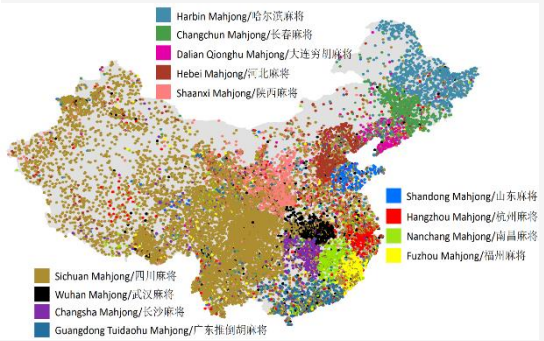




大数据时代的情报学研究

王贤文

大连理工大学

人文与社会科学学部

2016.06　深圳

目　录

01

PART ONE

情报学与大数据

Information Science and Big Data

1.情报学与大数据

- 情报（information）本身就蕴含着大量的数据
  - 各类型文献信息资源及其描述信息(如期刊论文、图书、专利、标准等题录、引文数据、全文本数据)
  - 科学数据(如观测数据、实验数据、基因图谱等)
  - 科技管理数据(如用户数据、统计数据、项目数据等)
  - 这些数据经过科学发展的累积过程,目前已经成为庞大的数据体系。

1.情报学与大数据

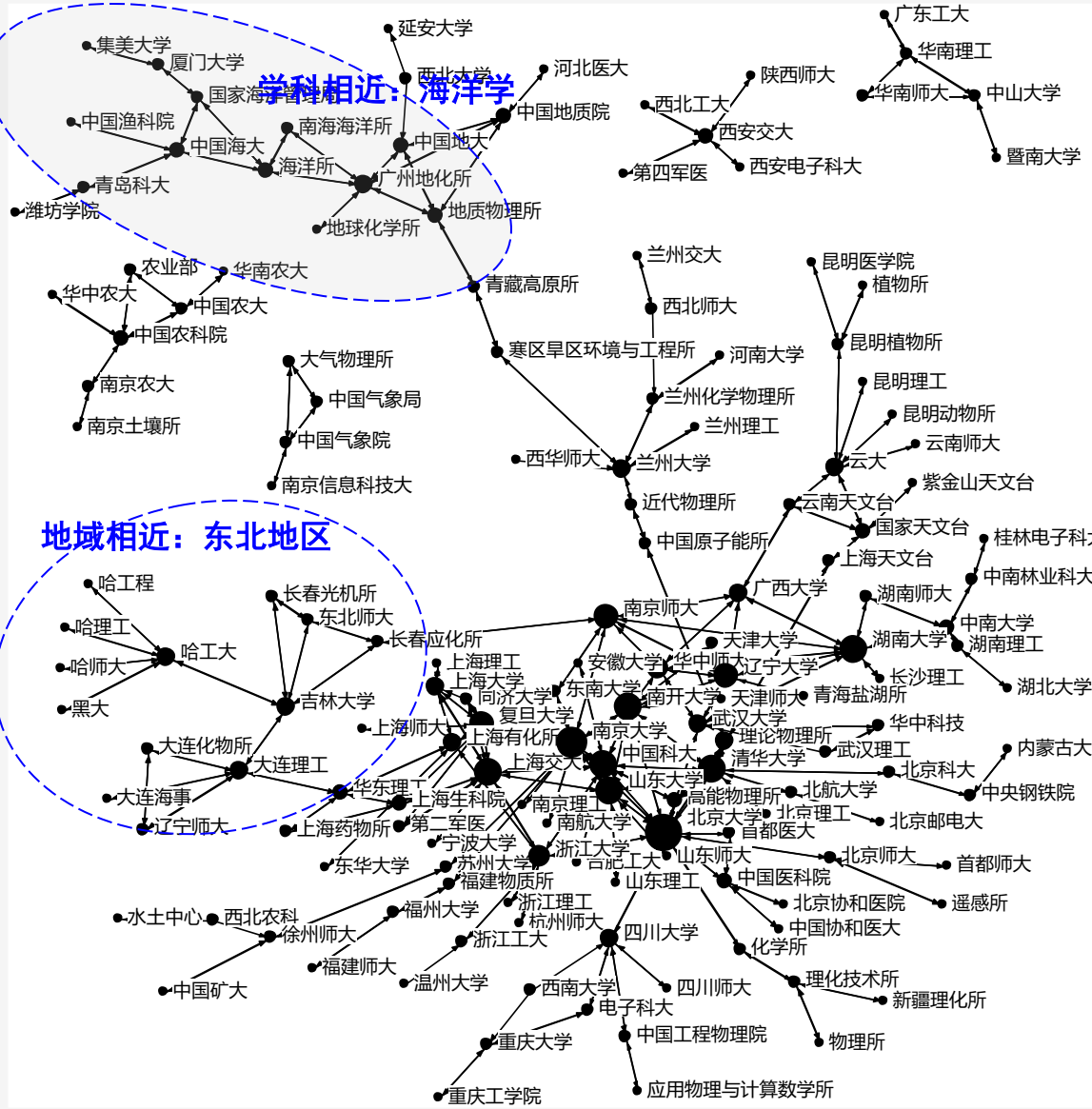
- 以Web of Knowledge平台为例
  - Web of Science数据库（1900-至今）：6200余万篇科学文献，数十亿条参考文献信息
  - Derwent Innovations Index数据库（1963-至今）：3000余万篇专利文献，数十亿条专利引文信息
  - Zoological Record数据库（1864-至今）：420余万篇生物学分类文献

http://www.cssn.cn/kxk/qbwxx/201509/t20150904_2146388.shtml

1.情报学与大数据

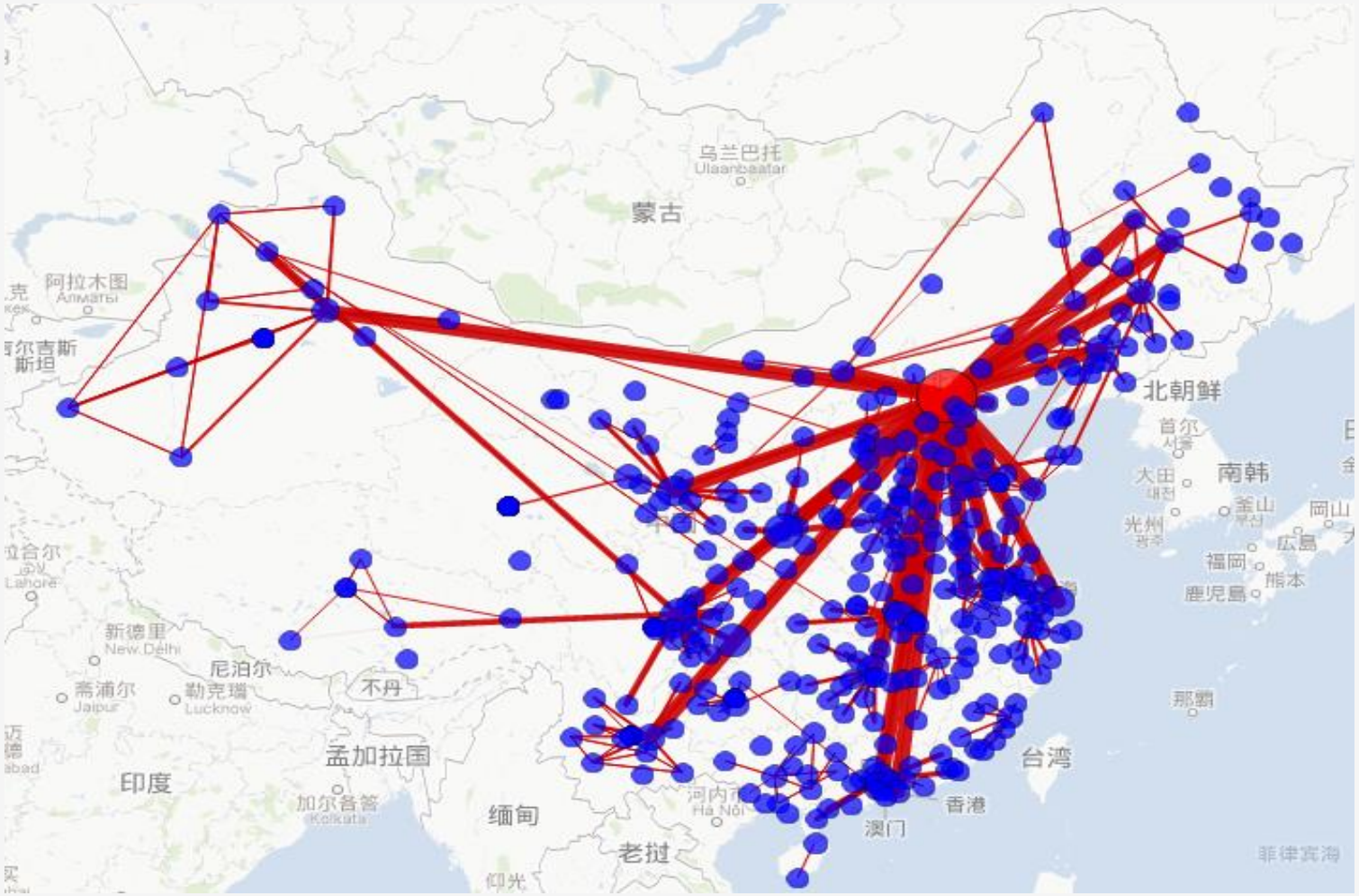
基于SSCI论文的合著分析，研究中国研究机构的合作网络

- **王贤文**, 丁堃, 朱晓宇. 中国主要科研机构的科学合作网络分析［J］. 科学学研究, 2010, 28（12）:1806-1812.



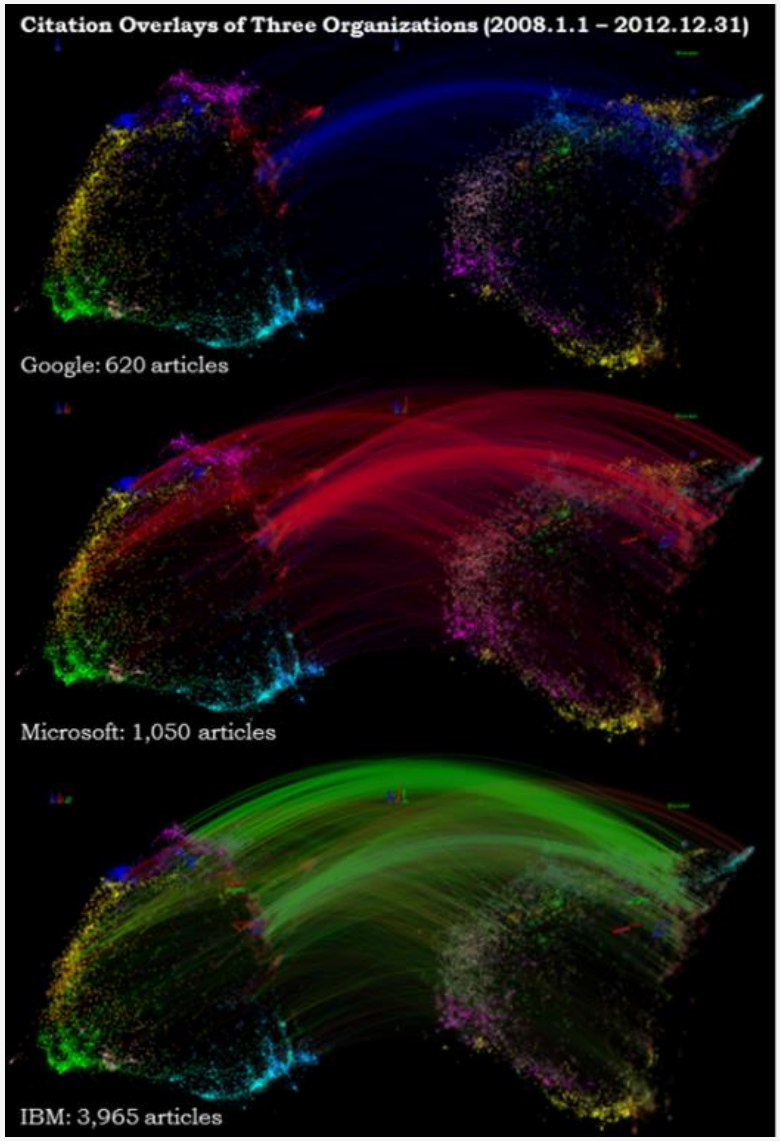
机构合作网络

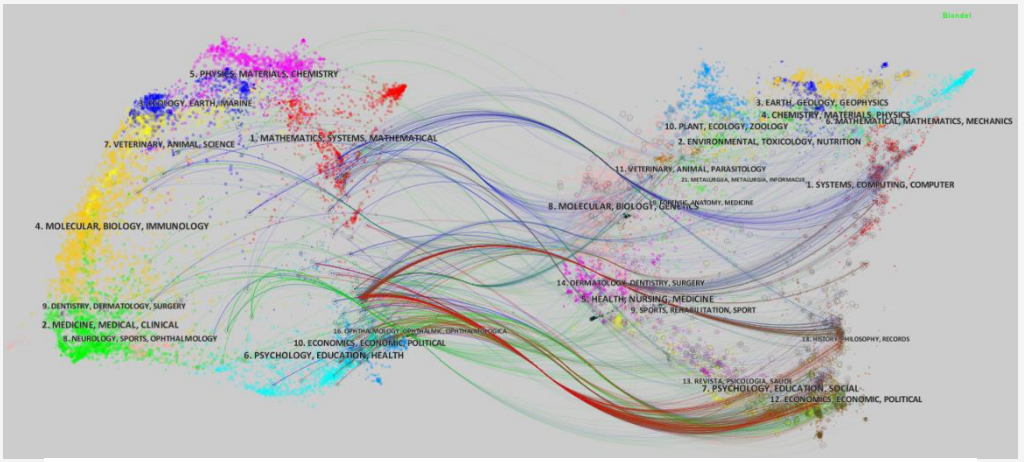
1.情报学与大数据

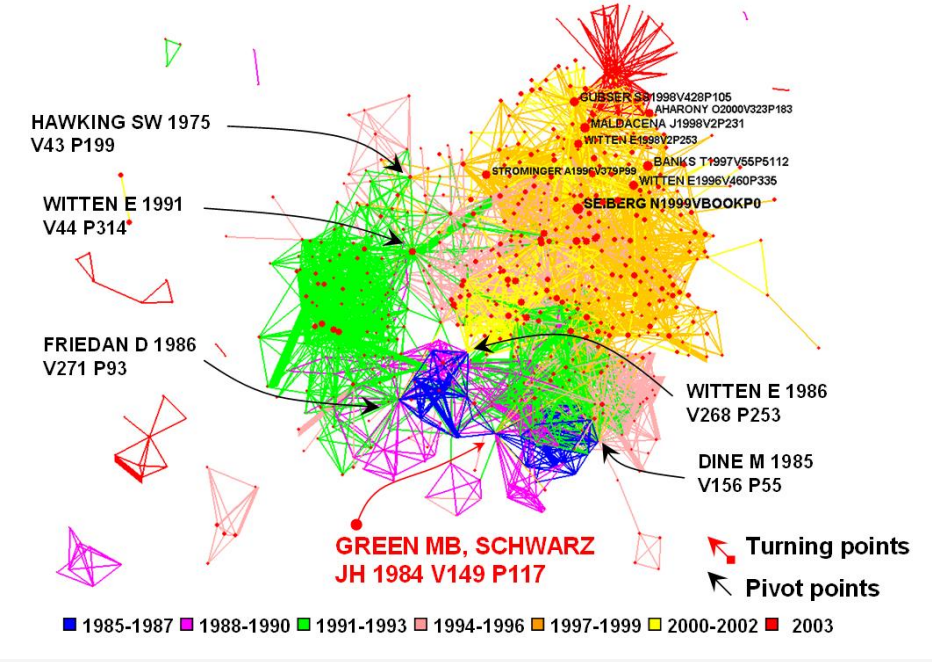


城市合作网络

1.情报学与大数据







利用CiteSpace进行文献引用分析

2.第一篇SSCI论文的发表

基于数千万条专利的共被引分析，研究企业之间的竞争与联系

- **王贤文 等. 基于专利共被引的企业技术发展与技术竞争分析:以世界500强中的工业企业为例[J]. 科研管理,2010,04:127-138.**

世界500强企业之间的专利共被引网络

02

PART TWO

互联网大数据与情报学

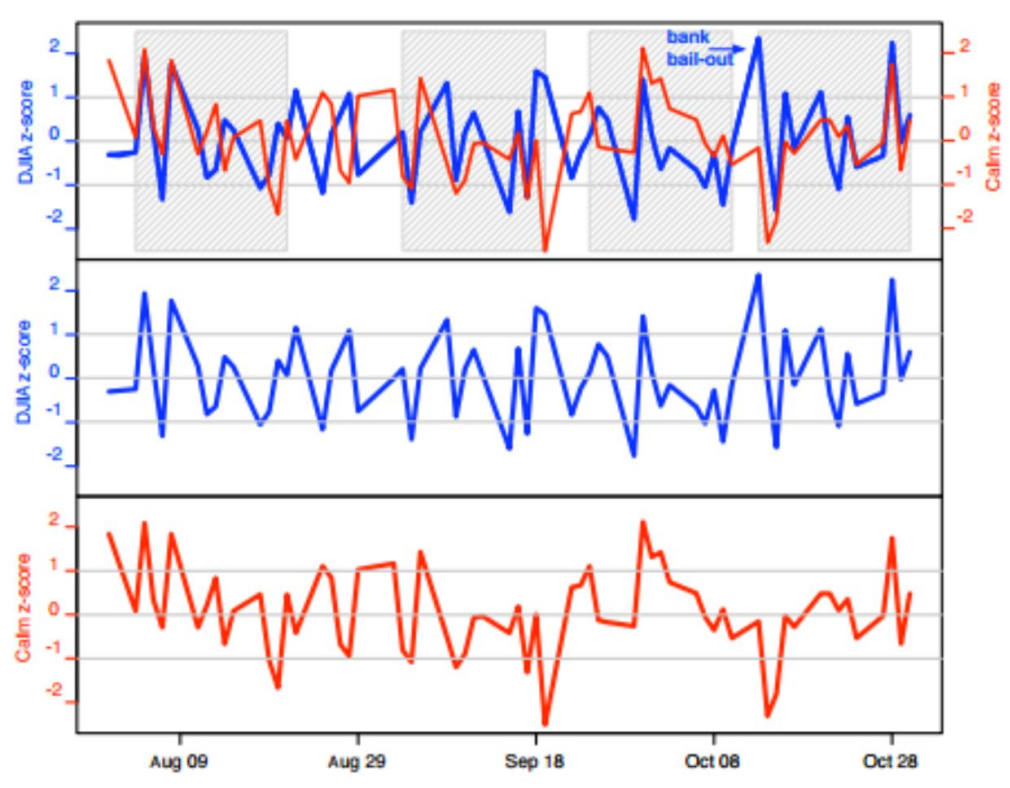
Internet Big Data and Information Studies

2.互联网大数据与情报学

- 互联网大数据与科学研究
  - 大数据来源：Google、Facebook、Twitter.....
  - TS=Facebook检索，得到4320篇SCI与SSCI论文
  - TS=Twitter，2911篇

2.互联网大数据与情报学

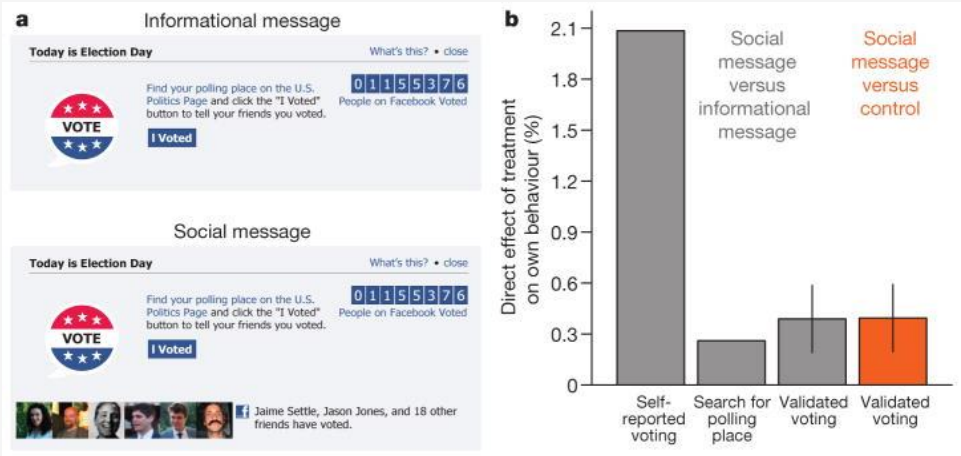
- 互联网大数据与科学研究：研究案例
  - Twitter预测股市

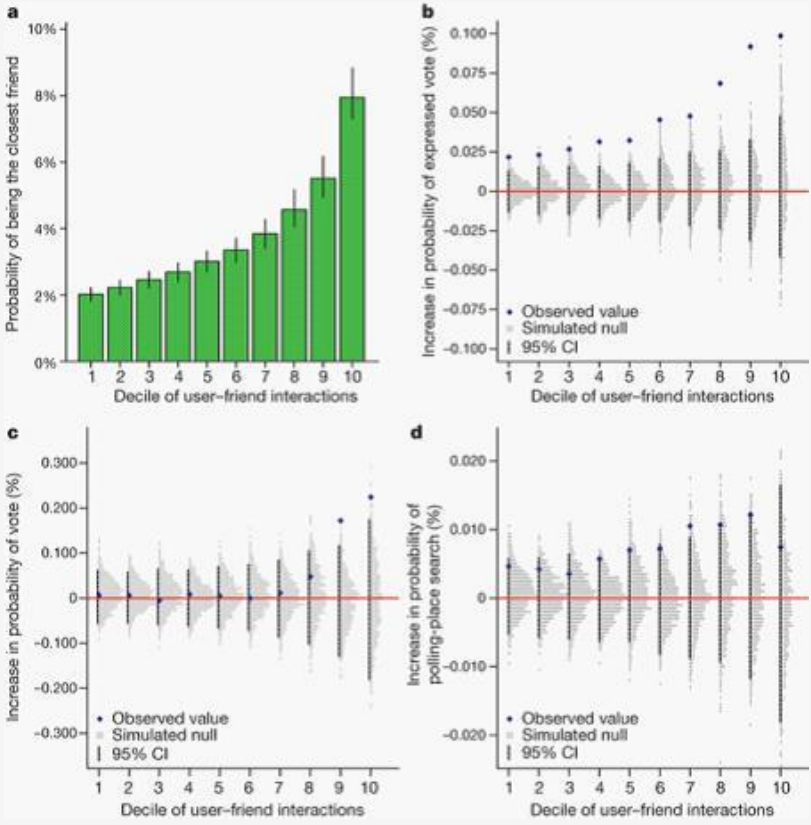


Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1-8.

2.互联网大数据与情报学

- 互联网大数据与科学研究：研究案例
  - Facebook影响选举结果





Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 295-298.

PART THREE

Altmetrics大数据

Altmetrics Big Data

3.Altmetrics大数据

- Altmetrics基本概念
  - altmetrics is the creation and study of new metrics based on the Social Web for analyzing, and informing scholarship.
  - Altmetrics是对根植于社交网络的系列新指标的构建和研究，其目的在于分析和理解学术研究。

3.Altmetrics大数据

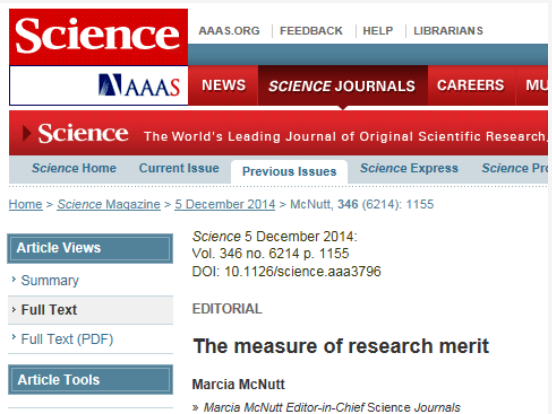
- Altmetrics测度类别
  - 网络采集：喜欢、书签、保存、读者数等
  - 网络讨论：评论、博文、媒体报道等
  - 社交媒体：转发、分享、推荐、点赞等

3.Altmetrics大数据

- Altmetrics产生背景
  - 出版业的变革
    - 电子出版对纸质期刊的取代，甚至有完全电子版的期刊
    - 人们的阅读习惯也随之数字化转变
  - 社交网络的爆发式增长
    - 大众社交：Facebook、Twitter
    - 学术社交：Mendeley、Researchgate
    - 越来越多的人，包括科学家自身参与到社交网络

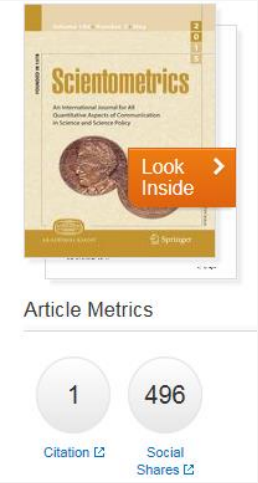
3.Altmetrics大数据

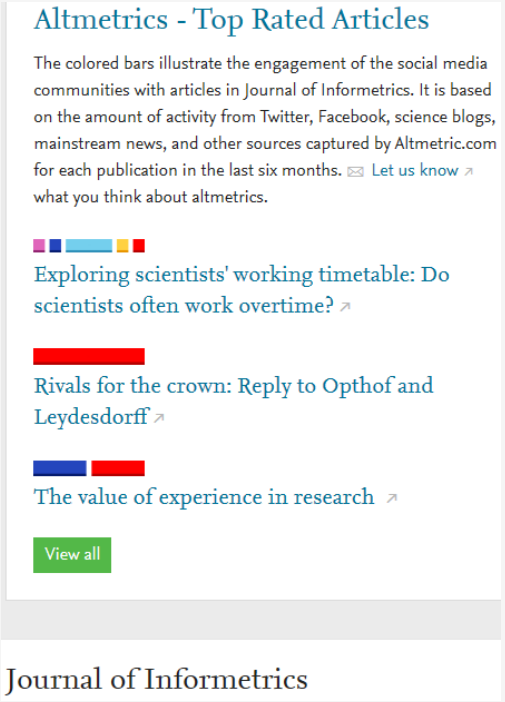
- Altmetrics产生背景
  - 2012年，北卡大学教堂山分校信息与图书馆学系的在读博士生Jason Priem提出altmetrics概念，迅速风靡学术圈。
  - SCI/SSCI检索altmetric*，已经发表135篇论文。（截至2016年6月6日）
  - Nature发表14篇，Science发表3篇



3.Altmetrics大数据

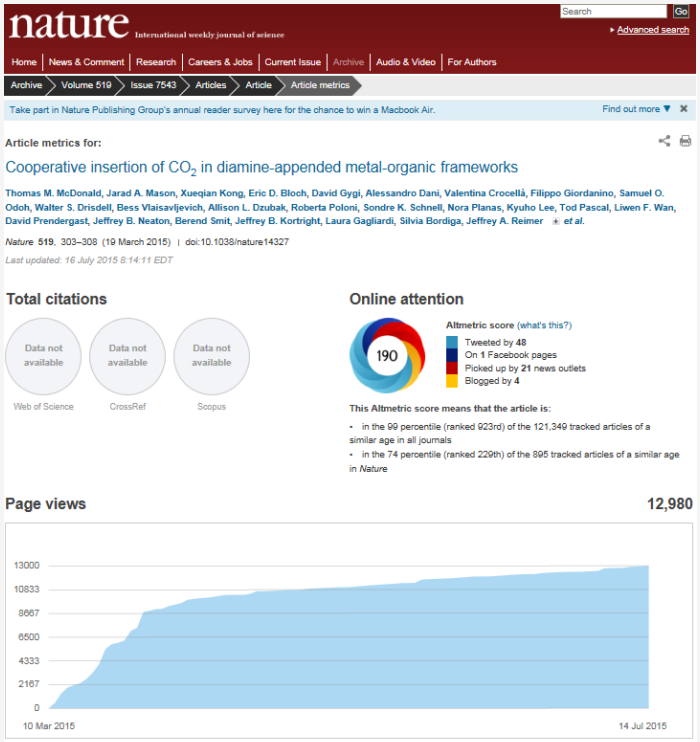
- Altmetrics应用
  - 采用altmetrics指标的学术出版商
    - Elsevier
    - Springer
    - Wiley

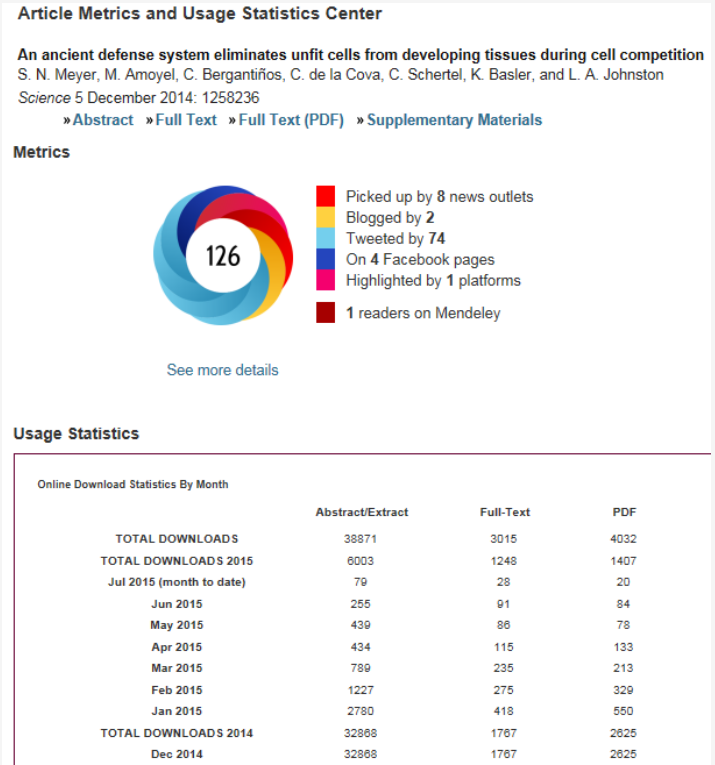




3.Altmetrics大数据

- Altmetrics应用
  - 采用altmetrics指标的学术出版商
    - Nature
    - Science





3.Altmetrics大数据

- Altmetrics工具与方法
  - Impactstory

www.impactstory.org

  - Plumanalytics

www.plumanalytics.com

  - Altmetric

www.altmetric.com

3.Altmetrics大数据

- Altmetrics工具与方法
  - altmetric.com

    - 成立于2001年
    - 分析学术论文在社交媒体、报纸和杂志上产生的影响。
    - 通过监测论文在网络上的各种影响力，将各种影响力指标赋予不同权重，换算成一个综合得分。

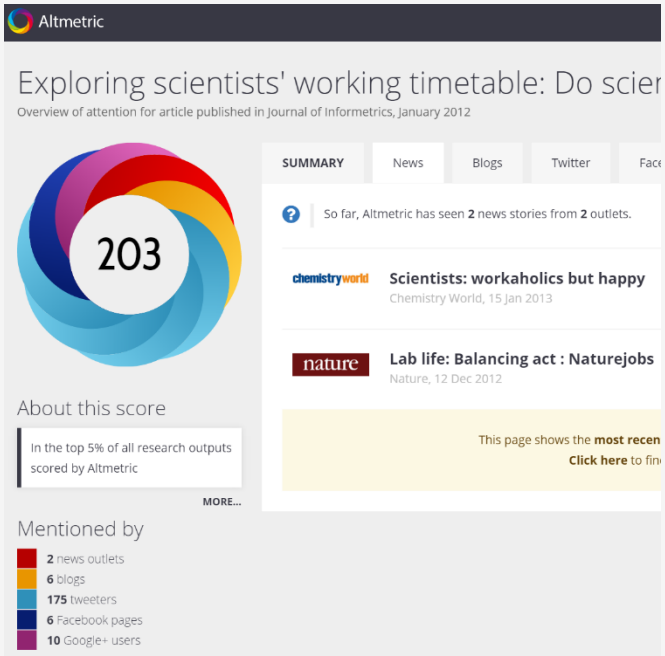
3.Altmetrics大数据

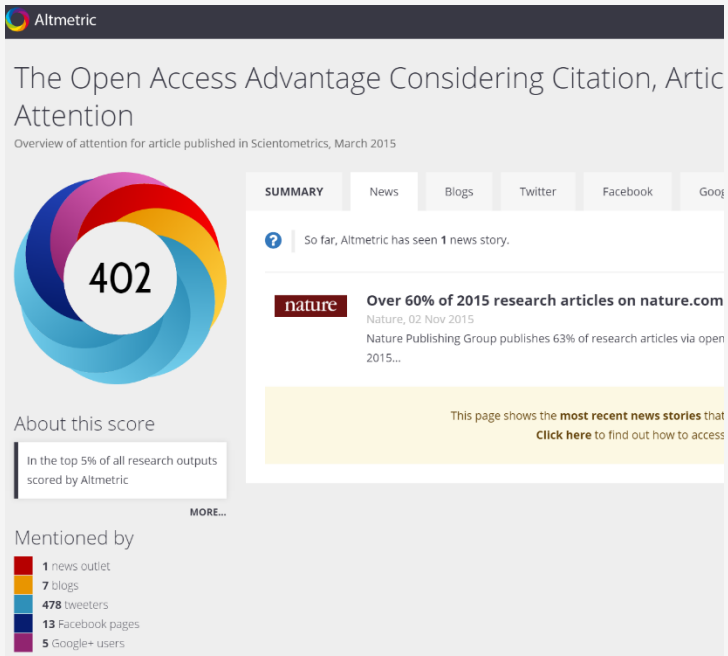
- Altmetrics工具与方法
  - Altmetric综合得分计算

Metrics	Score
News	8
Blogs	5
Twitter	1
Facebook	0.25
Sina Weibo	1
Wikipedia	3
Policy Documents (per source)	3
Q&A	0.25
F1000/Publons/Pubpeer	1
YouTube	0.25
Reddit/Pinterest	0.25
LinkedIn	0.5

3.Altmetrics大数据

- Altmetrics工具与方法
  - 查询一篇论文的altmetric得分
    - http://www.altmetric.com/details.php?doi=10.1016/j.joi.2012.07.003
    - http://www.altmetric.com/details.php?doi=10.1007/s11192-015-1547-0





PART FOUR

科学论文使用大数据

Usage Big Data of Scholarly Articles

4.科学论文使用大数据

- 使用数据定义
  - 论文被点击、下载、阅读、保存等的信息
  - 一篇论文可能只会被引用数十次，但是绝大多数论文都会被下载数百上千次。使用数据是比引用数据大得多的大数据。

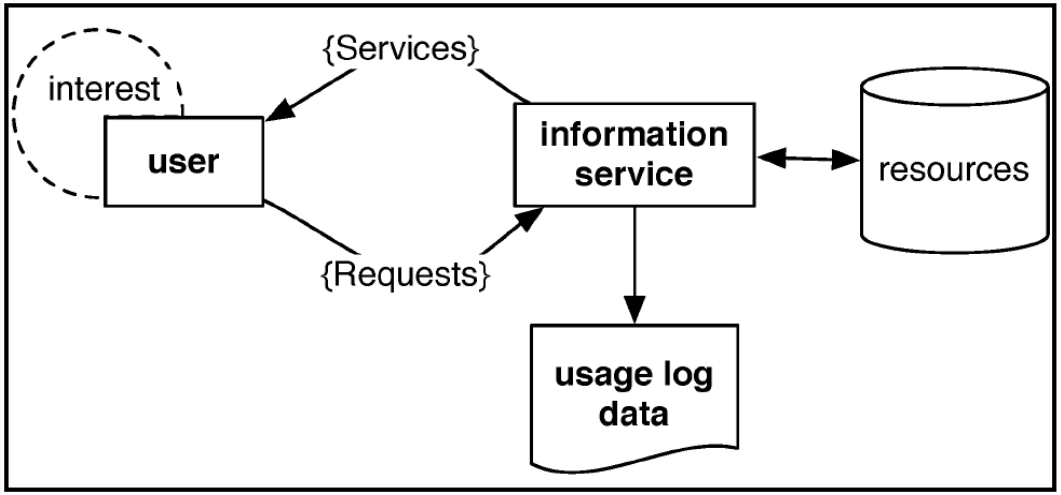


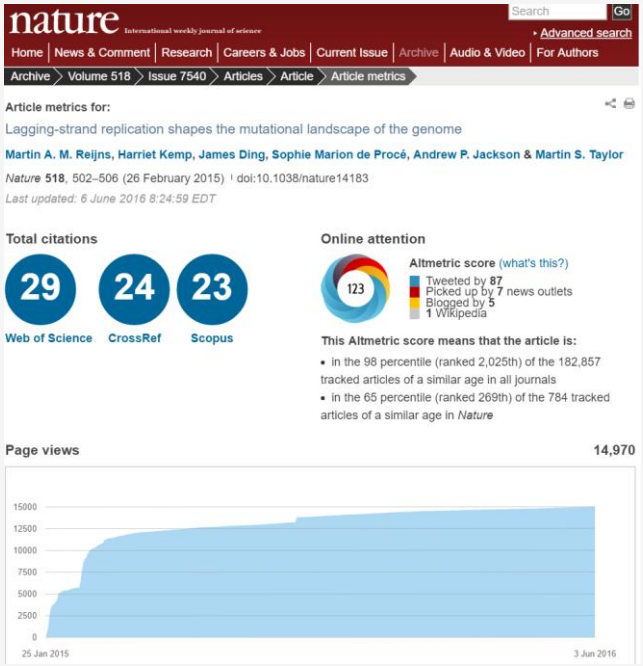
Figure 1.1 Service request model underlying definition of usage

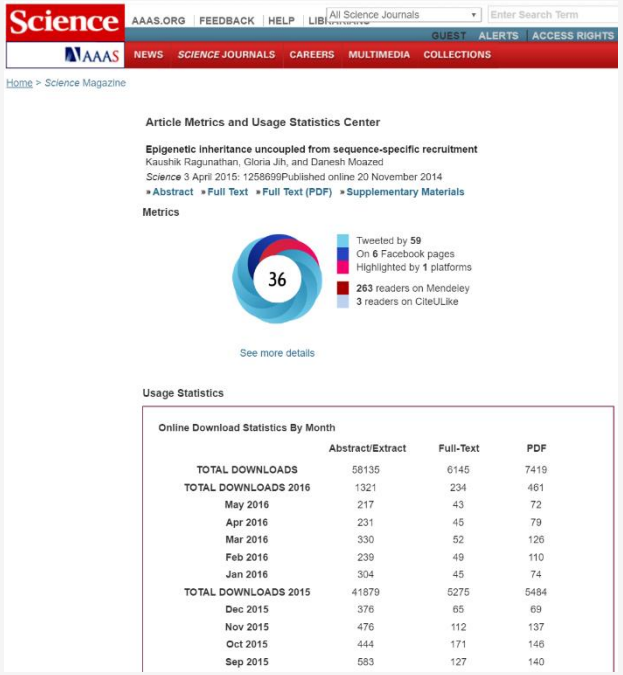
4.科学论文使用大数据

- 使用数据形式
  - 数据维度：访问者的ip地址、访问时间、停留时间、访问渠道、来源地区、访问内容...
  - 数据类型：html浏览、pdf下载
  - 用户画像：男性/女性、年龄、职业、学历...

4.科学论文使用大数据

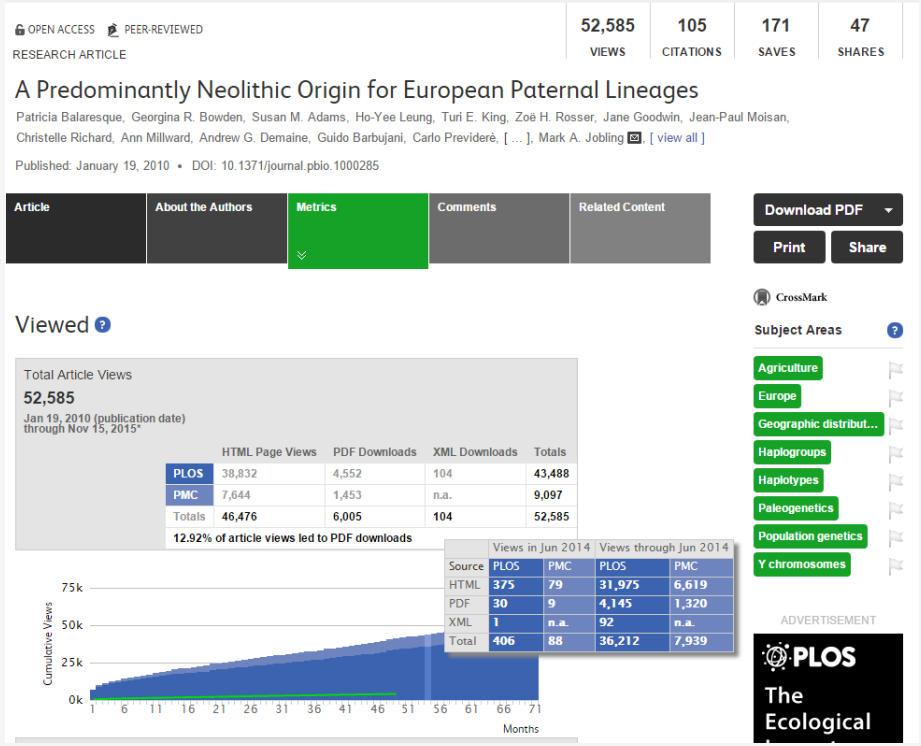
- 使用数据的收集来源
  - Nature
    - 每一篇研究性论文在每一天被浏览和下载的次数
    - http://www.nature.com/nature/journal/v518/n7540/full/nature14183.html
  - Science
    - 每一篇论文在每个月的摘要和全文浏览、pdf下载次数
    - http://www.sciencemag.org/articleusage?gca=sci;348/6230/1258699





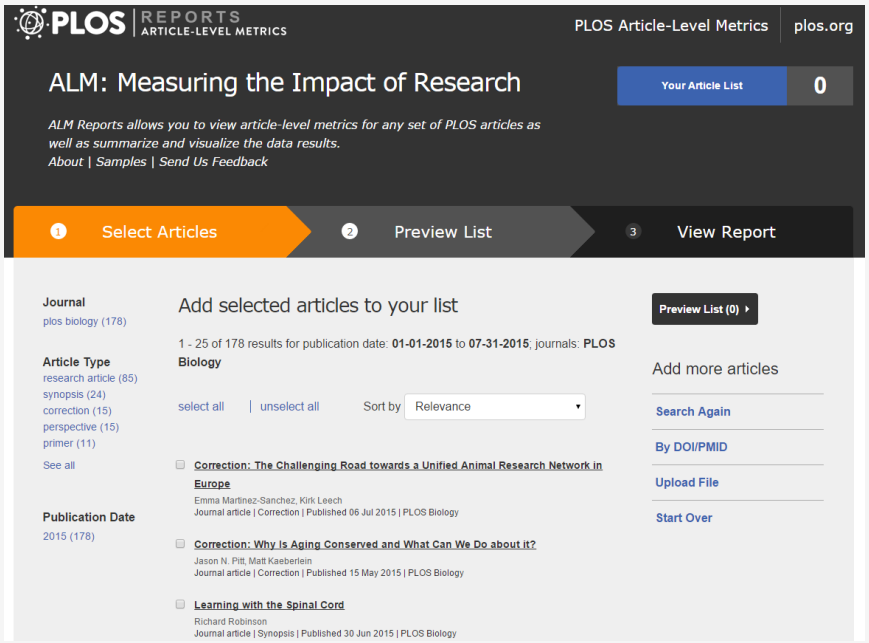
4.科学论文使用大数据

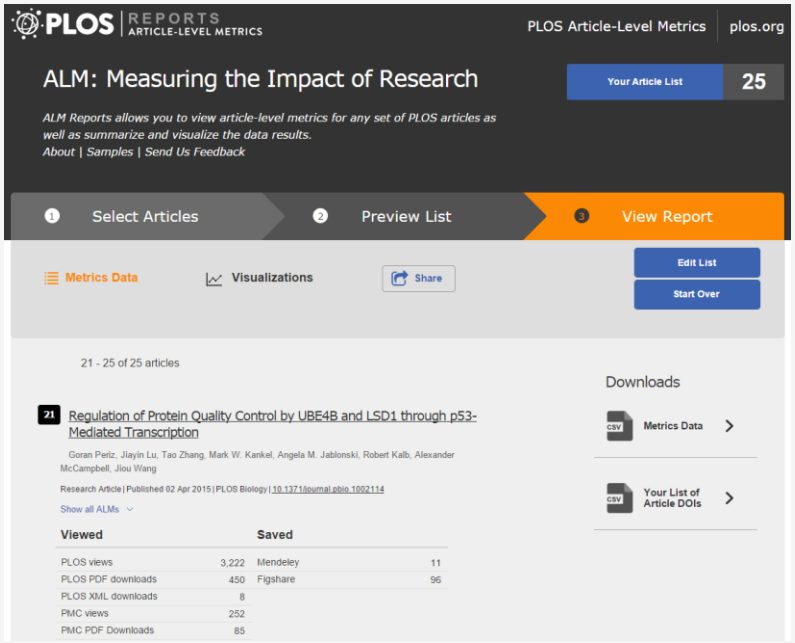
- 使用数据的收集来源
  - PLOS
    - 每一篇论文的metics数据
    - http://www.plosbiology.org/article/Metrics/info:doi/10.1371/journal.pbio.1000285



4.科学论文使用大数据

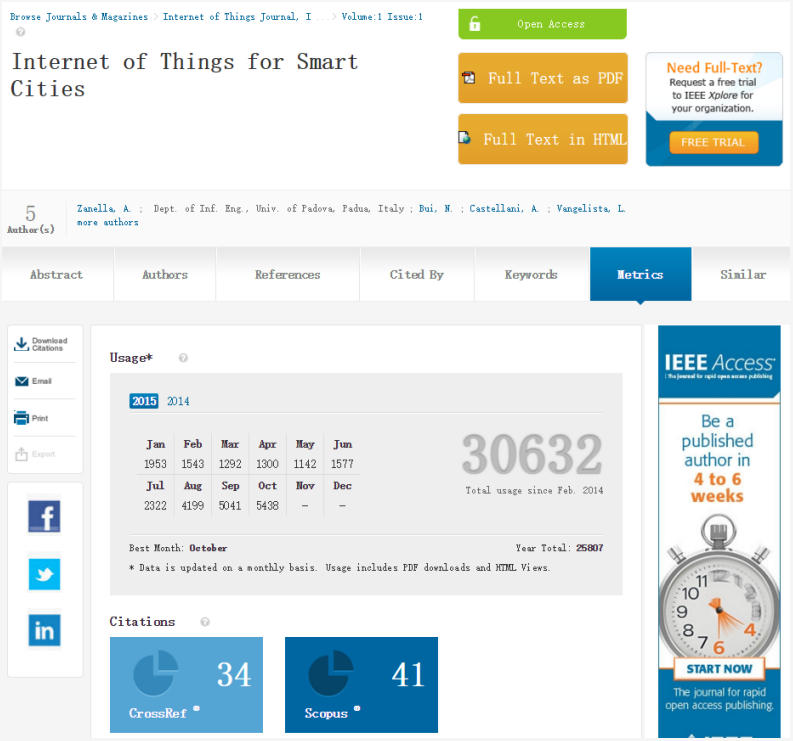
- 使用数据的收集来源
  - PLOS
    - 批量下载PLOS出版论文的发文数据、使用数据、altmetrics数据
    - http://almreports.plos.org//





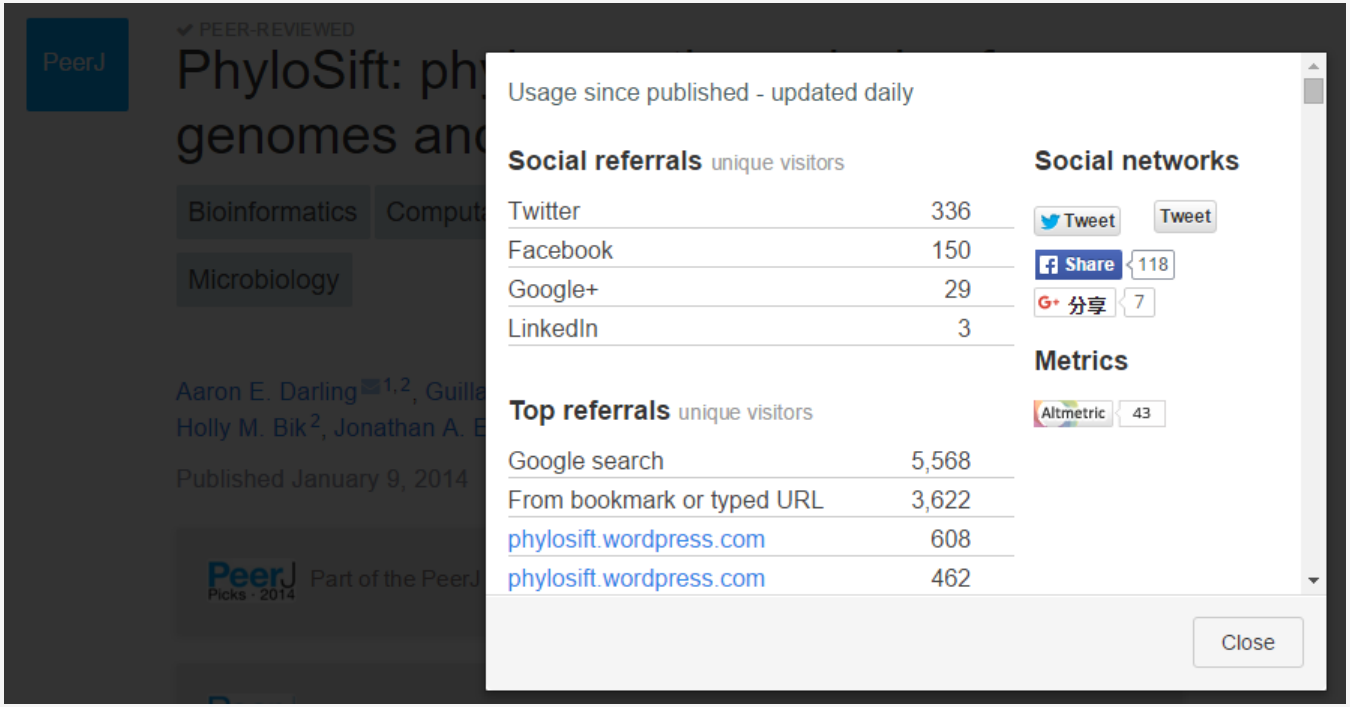
4.科学论文使用大数据

- 使用数据的收集来源
  - IEEE
    - http://ieeexplore.ieee.org/xpl/abstractMetrics.jsp?arnumber=6740844



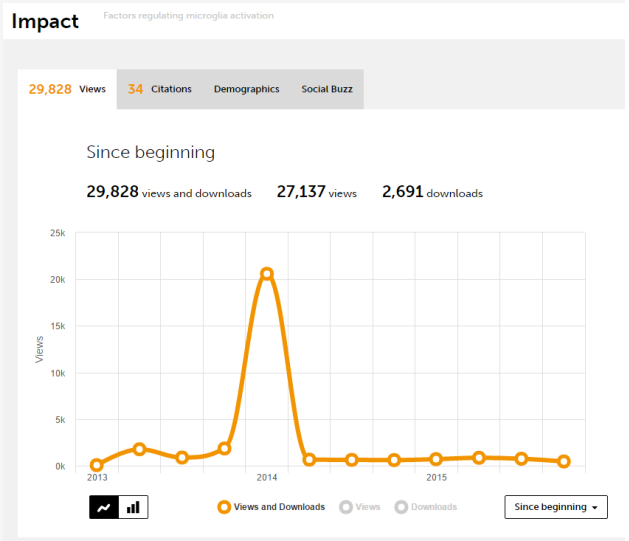
4.科学论文使用大数据

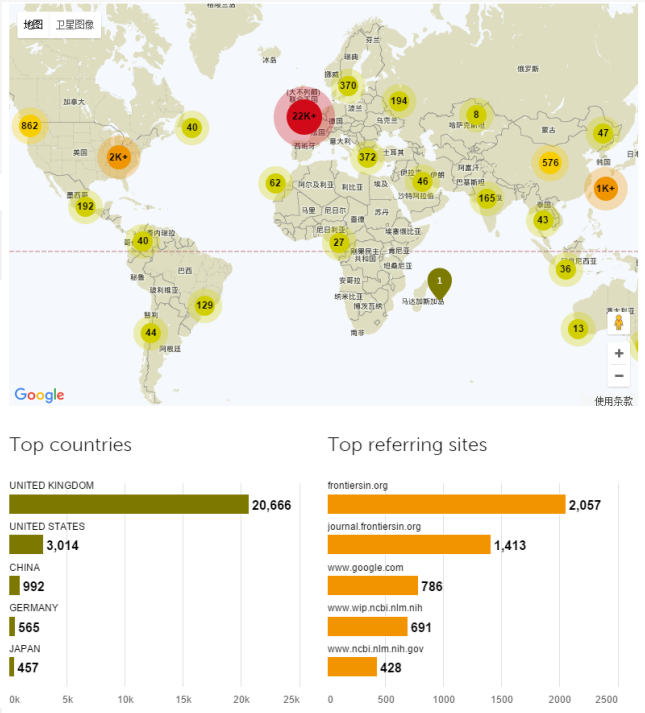
- 使用数据的收集来源
  - PeerJ
    - 可以查看每一篇Peerj论文的metrics数据，包括访问者数、访问次数、下载次数、访问来源
    - https://peerj.com/articles/243/



4.科学论文使用大数据

- 使用数据的收集来源
  - Frontiers
    - 每一篇Frontiers出版论文的metrics数据，包括访问次数、下载次数、访问来源
    - http://journal.frontiersin.org/article/10.3389/fncel.2013.00044/abstract#impact





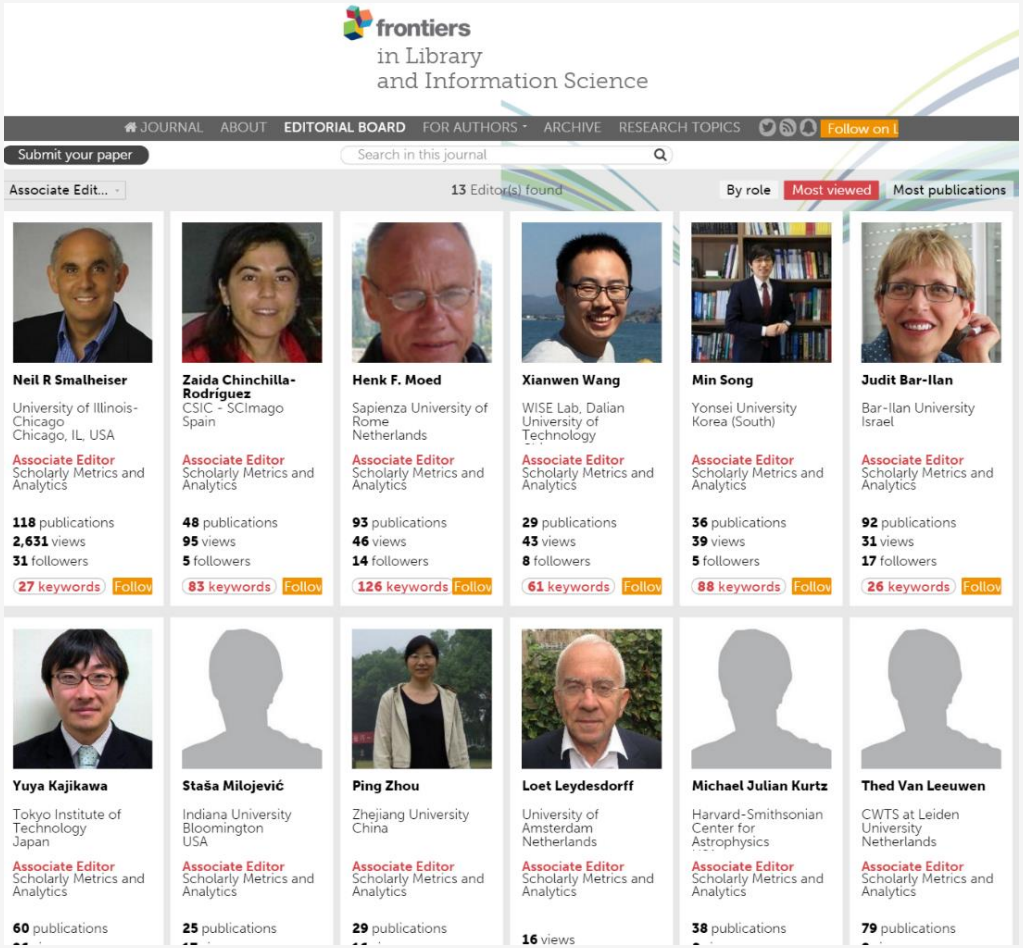
4.科学论文使用大数据

- Frontiers in Scholarly Metrics and Analytics

主编

Chaomei Chen（Drexel University，USA）

副主编：



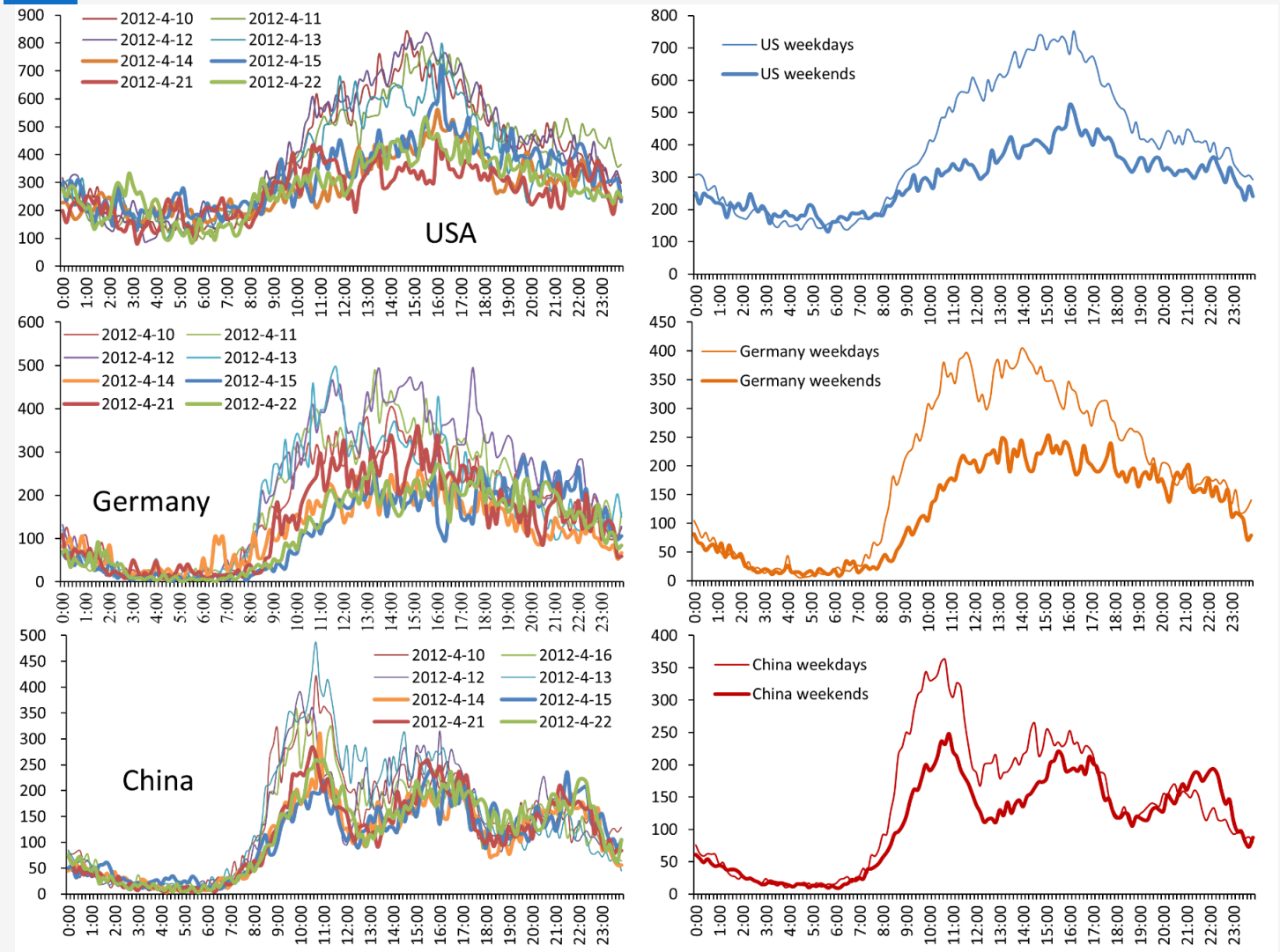
4.科学论文使用大数据：探索科学家的工作时间表

- **以往的研究**
  - 研究方法
    - 案例研究、问卷调查
  - 局限
    - 样本有限
- **我们的研究**
  - 新的数据形式
    - 以科学家下载论文的行为来反映科学家的工作状态
    - http://realtime.springer.com/map
  - 研究中的技术难题
    - 数据收集 – 数据变化太快
  - 最终
    - 为期两周的监测，我们收集到将近200万条数据

3.科学论文使用大数据： 探索科学家的工作时间表

科学论文实时下载地图

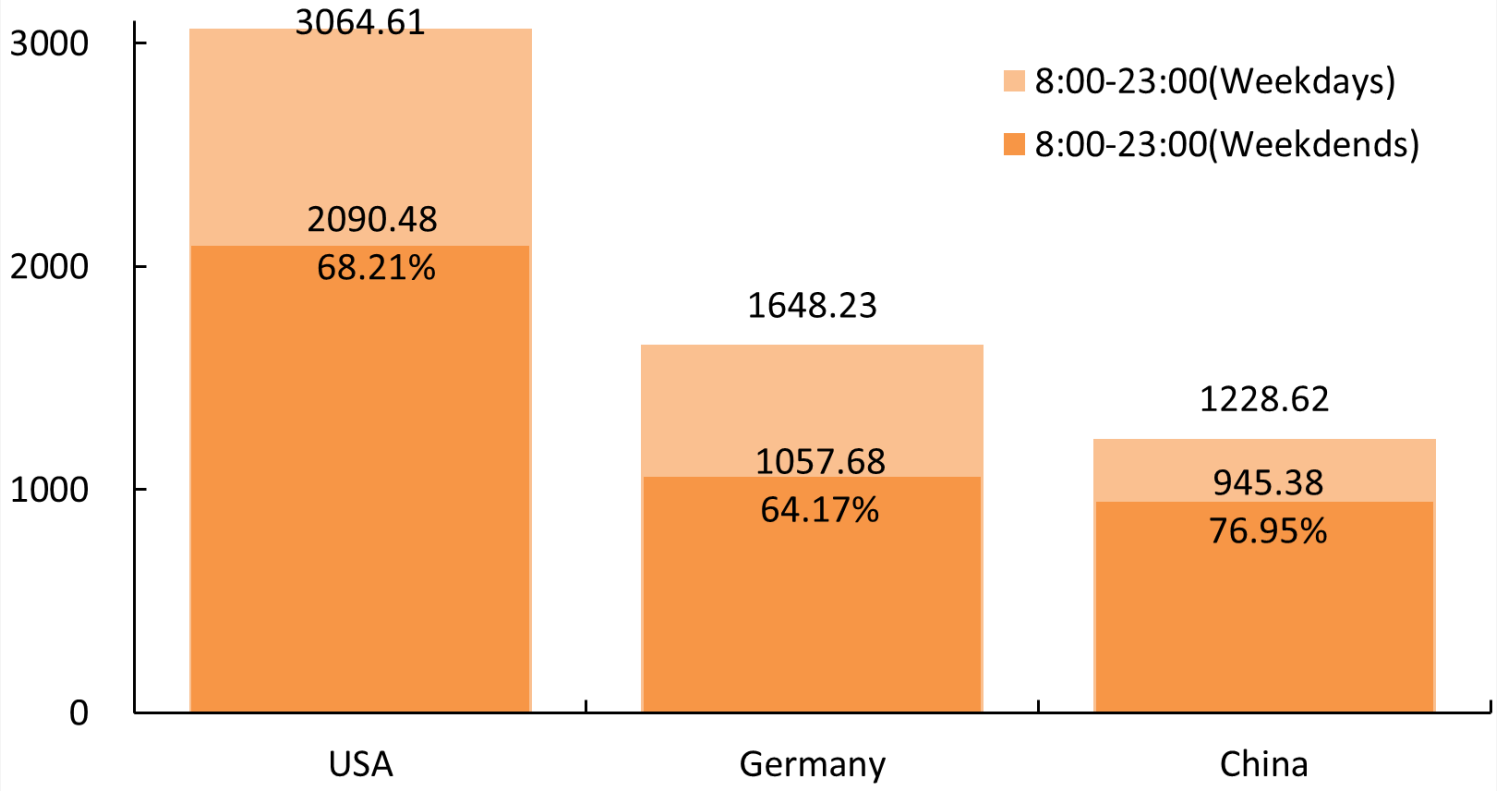
3.科学论文使用大数据： 探索科学家的工作时间表



科学家的工作时间曲线

Xianwen Wang et al. **Exploring Scientists' Working Timetable: Do Scientists Often Work Overtime?** [J] *Journal of Informetrics*. 2012, 6(4): 655-660.

3.科学论文使用大数据：探索科学家的工作时间表



周末论文下载量与平时的对比

Xianwen Wang et al. **Exploring Scientists' Working Timetable: Do Scientists Often Work Overtime?** [J] *Journal of Informetrics*. 2012, 6(4): 655-660.

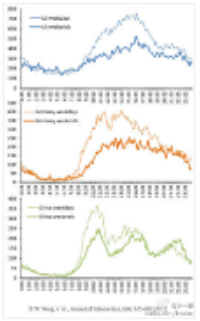
3.科学论文使用大数据：探索科学家的工作时间表

- **结论与启示**
  - 科学家的工作时间虽然自由，但不像人们想象的那样轻松
  - 放弃了休息、娱乐、锻炼、和家人陪伴的时间。
  - 晚上熬夜、周末加班成为工作常态
  - 科研工作者需要把握工作与生活的平衡。
  - 科研不是百米短跑冲刺，而是一场漫长的**马拉松**

3.科学论文使用大数据：探索科学家的工作时间表

- 新浪微博：转发5500余次
  - @李开复: 很有意思： “1）科学家基本上没有周末。 2）科学家基本上不分上下班。 3）中国的吃饭时间管得挺牢，美国还喜欢夜战。”

“发现个神paper：科学家的工作时间。大连理工的王贤文等人利用论文下载时间来反映美国德国中国的科研工作者的工作时间，得出挺有趣的结论： 1）科学家基本上没有周末。 2）科学家基本上不分上下班。 3）中国的吃饭时间管得挺牢，美国还喜欢夜战。”原来：美国是全日型，德国日用型，中国三段式



收藏 | 5596 | 575 | 15

3.科学论文使用大数据：探索科学家的工作时间表

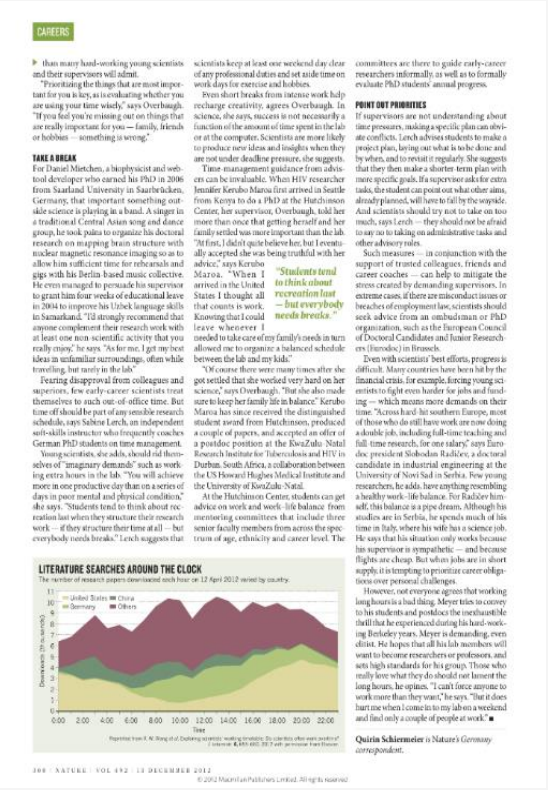
2012年12月12日，***Nature***以两个版面的形式发表了对我们一项研究的长篇采访报道。


CAREERS

Balancing act

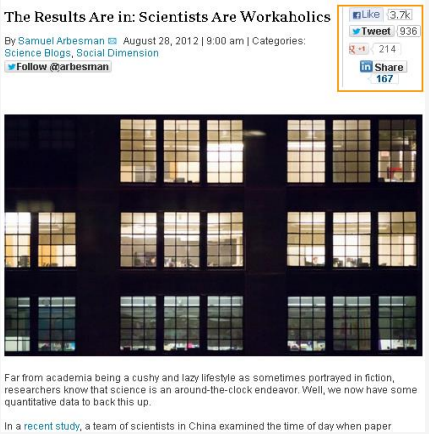
Many researchers struggle to take time off from the lab. But scientists should try to improve their work–life balance.

BY QUIRIN SCHIERMEIER


CAREERS

LITERATURE SEARCHES AROUND THE CLOCK

3.科学论文使用大数据：探索科学家的工作时间表


The Results Are in: Scientists Are Workaholics

By Samuel Arbesman | August 28, 2012 | 9:00 am | Categories: Science Blogs, Social Dimension

Follow @arbesman

Like 3.7k Tweet 936 +1 214 Share 167

Far from academia being a cushy and lazy lifestyle as sometimes portrayed in fiction, researchers know that science is an around-the-clock endeavor. Well, we now have some quantitative data to back this up.

In a recent study, a team of scientists in China examined the time of day when paper

美国《连线》wired


RSC Advancing the Chemical Sciences

Publishing Platform | ChemSpider | MyR

chemistryworld

ARCHIVE

HOME | NEWS | OPINIONS | FEATURES | REGULARS | JOBS

RESEARCH | BUSINESS | NEWS AND ANALYSIS

Scientists: workaholics but happy

6 September 2012 Ned Stafford

Like 935 Tweet 156 +1 11

A new study has found what many already know – that when the sun goes down researchers don't go to bed, but continue working on into the wee small hours of the night. And many of them also carry on working through the weekend too.

'In the science community, achievements always correlate with hard work,' according to a paper written by a team of Chinese researchers, who have been following the working habits of scientists around the world. 'Motivated by the pressure of competition from all over the world, most scientists today are living and breathing their work.'

In the Chinese study, researchers at Dalian University of Technology used a 'novel method' to monitor researcher work habits. They analysed

Scientists work long hours, but enjoy their jobs

英国皇家化学学会《化学世界》
Chemistry World


Arbeitszeit von Forschern

Freizeit? Wissenschaftler haben Freizeit?

06.09.2012 • Deutsche Forscher zwischen Amerikas Nachtarbeitern und Chinas Kantinenanbetern: Eine Studie zeigt, dass in wissenschaftlichen Instituten auch jenseits klassischer Tarifarbeitszeiten hoher Druck herrscht.

Von SIBYLLE ANDERL

Artikel | Bilder (2) | Lesermeinungen (1)

© REUTERS

Die wissenschaftliche Arbeitsmoral ist hoch - aber es gibt nationale Unterschiede

德国法兰克福汇报


FORSCHUNG AKTUELL

26.09.2012

Wenn Forscher zu lange forschen

Neue Studie deutet auf viele Überstunden und Wochenendarbeit hin

Von Michael Böddeker

Bei Medizinern sind lange Arbeitszeiten samt Schichtdienst, Bereitschaft und Überstunden fast die Regel, aber wie sieht es bei forschenden Wissenschaftlern an Hochschulen und in Forschungseinrichtungen aus? Das ist gar nicht so einfach herauszufinden, denn die Forscher können ihre Arbeitszeiten an der Uni oft flexibel gestalten. Informatiker aus China haben mit einer neuen Messmethode erste Ergebnisse erhalten.

Wissenschaftler mach offenbar viele Überstunden. (Bild: Stock.XCHNG/Jan Frin

德国公共广播电台


教科新闻

光明日报

北京正负电子对撞机上发现新的共振结构

奇特强子可能存在

辽宁精细化工协同创新中心启动

首批全民终身学习体育段位证书颁发

天津理工大学研究成果表明：

科学家都是"工作狂"

光明日报


中国科学报

基金

一个考察科研人员生存状态的独特视角发现：

熬夜加班已成科研常态

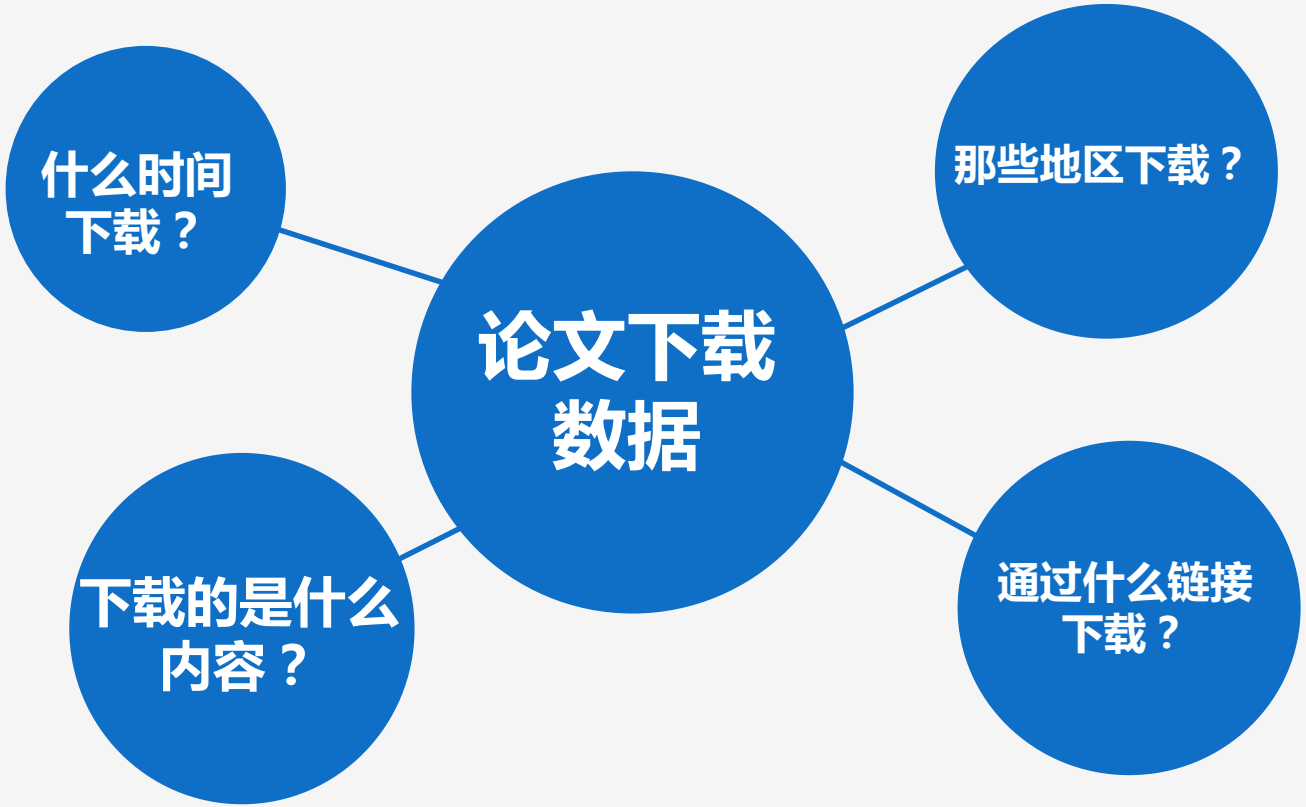
USA

Germany

China

中国科学报

4. 科学论文使用大数据



4.科学论文使用大数据：实时探测研究热点和前沿

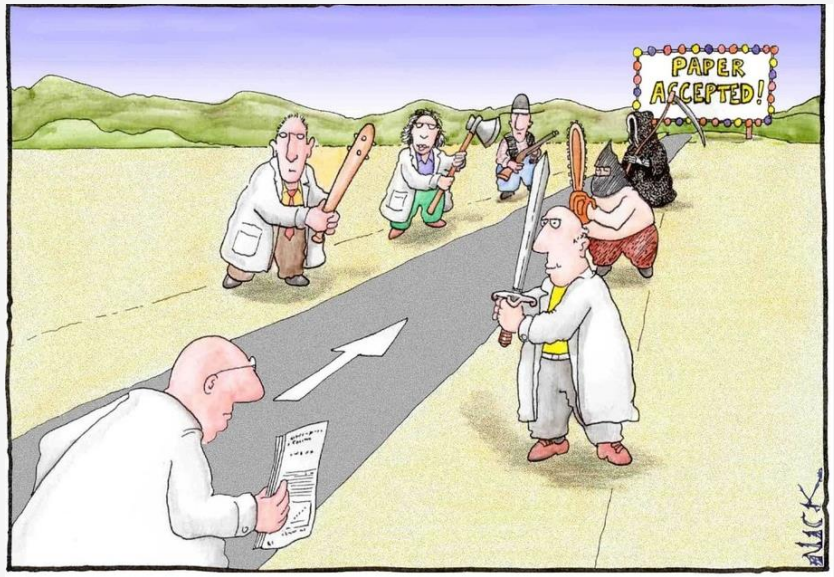
- **论文下载数据：下载的是什么内容？**
  - 实时追踪科研新趋势

- **科学研究的竞争是来自于全世界的竞争**
- **对于科研工作者来说，准确及时地知道领域同行所正在从事的研究主题，有助于**
  - 把握最新的研究动向
  - 走在领域的国际科学前沿
  - 在全球科学技术竞争中占据先发优势

4. 科学论文使用大数据：实时探测研究热点和前沿

- **以往的研究**
  - 通过对已发表文献进行回溯研究，总结研究趋势。
  - 存在较大的时间滞后。

- **为什么会产生时间滞后**
  - 新产生idea→查阅文献→实验→论文写作→投稿→外审→修改→录用→发表→被人下载→被人引用→引用的论文发表



4. 科学论文使用大数据：实时探测研究热点和前沿

- **新的思路**
  - 根据科学文献的实时下载次数进行判断，如果某一主题的正在被持续地频繁下载，说明这类主题正在得到大量关注，极有可能是目前的研究热点。
  - 通过获取研究者正在下载、阅读和使用的科学文献信息，则可以反过来判断科学家目前正在从事的研究主题。

4. 科学论文使用大数据：实时探测研究热点和前沿

- **研究热点**
  - 根据科学文献的实时下载次数进行判断，如果某一主题正在被持续地频繁下载，说明这类主题正在得到大量关注，极有可能是目前的研究热点。

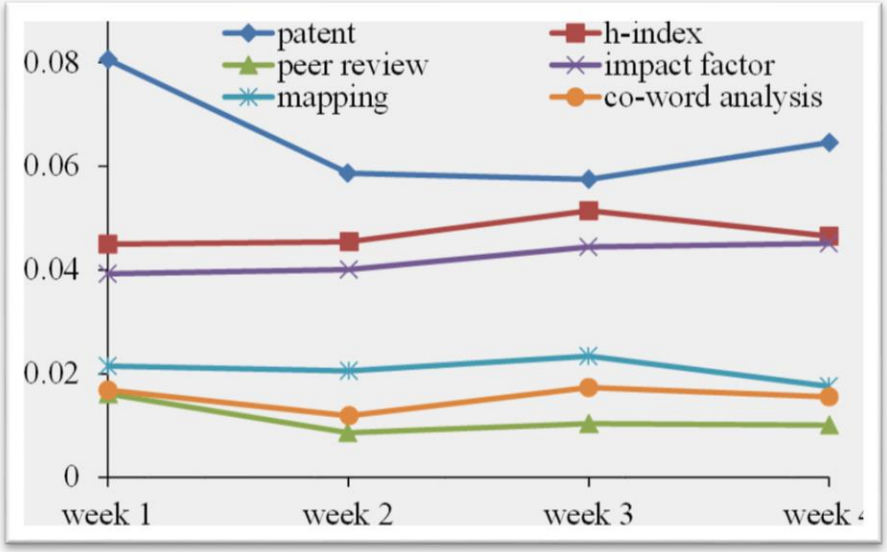
$$Ratio1 = \frac{downloads\ of\ the\ keyword}{total\ \ downloads}$$

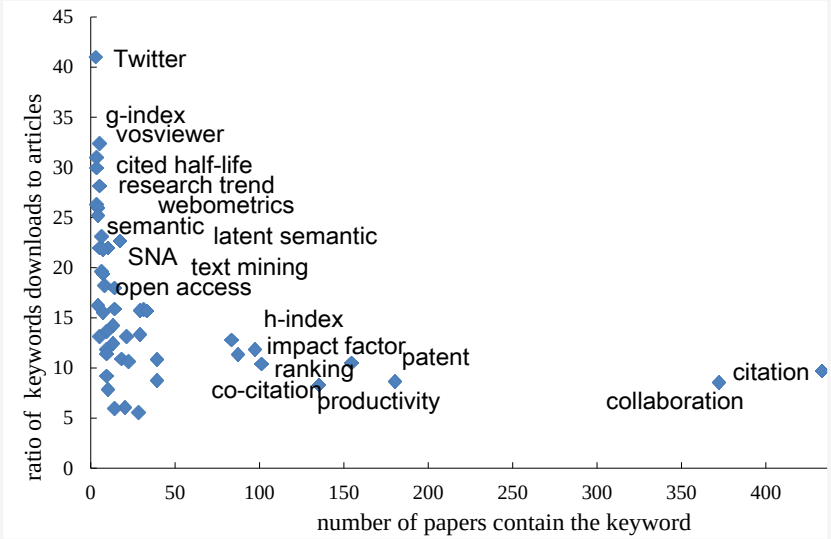
4.科学论文使用大数据：实时探测研究热点和前沿

- **研究前沿**
  - 将主题的被下载次数除以该主题已经发表的文献数量，再对这类主题论文的发表时间进行比对，如果除商为一个较高的数值，并且文献的平均发表时间比较新，那么该主题有可能会成为研究前沿。
  - 换句话说，某类主题已发表的论文数量虽然不多，但是发表时间比较新，说明这个主题的思想是近年刚提出的。并且如果这个主题的论文最近一段时间被频繁下载（尽管下载的次数不是所有主题中最多的），那么说明这个新思想正在引起大家的强烈关注，很有可能成为领域的研究前沿。

4. 科学论文使用大数据：实时探测研究热点和前沿

- 研究结果

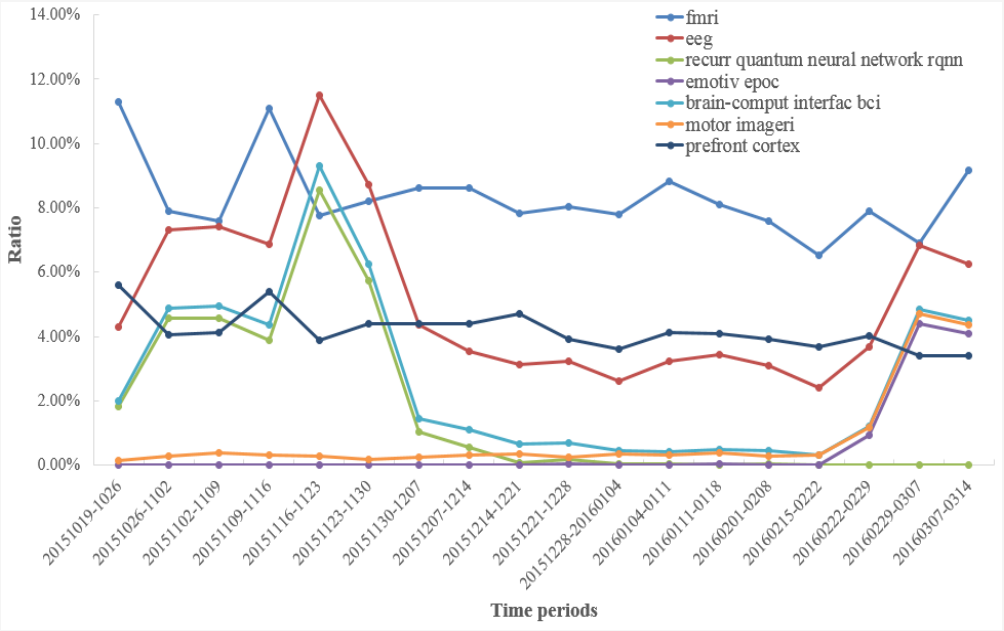


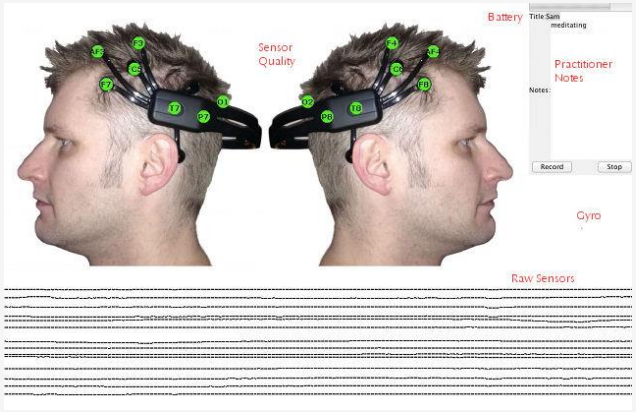


Xianwen Wang et al. Tracing scientists' research trends realtimely[J] *Scientometrics*, 2013, 95 (2): 717-729.

4. 科学论文使用大数据：实时探测研究热点和前沿

- 研究结果





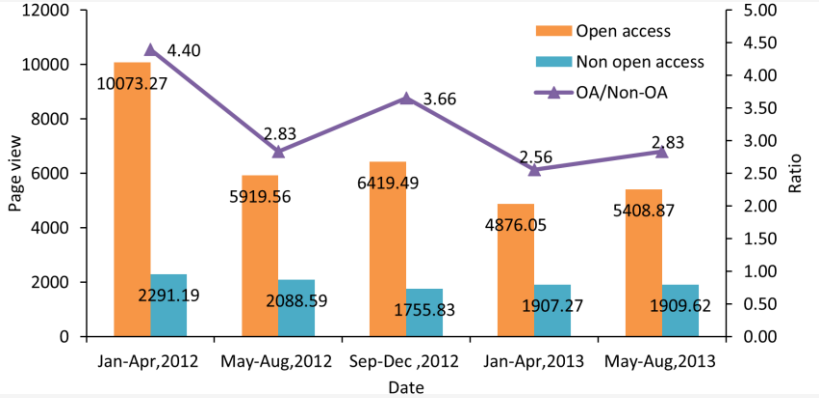
Emotiv epoc/
意念控制器

Detecting and Tracking Real-time Hot Topics Using WoS Usage Count: A Study on Computational NeuroscienceTracing[C] *STI 2016*, 2016, submitted

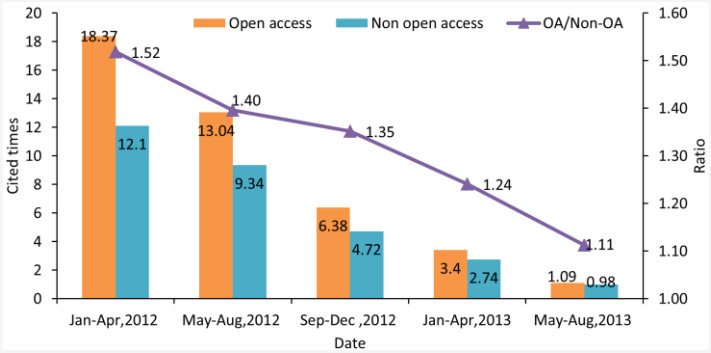
4. 科学论文使用大数据：开放获取优势

基于下载、社媒讨论、引用三重指标的论文开放获取优势

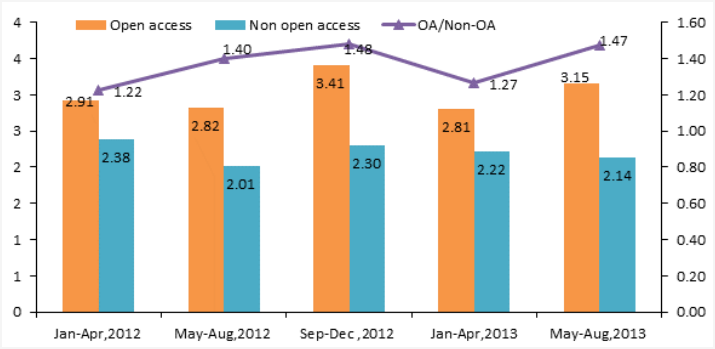
- **研究问题**
  - 论文Open Access能否带来更高的关注度和影响力？



OA与非OA下载次数比较



OA与非OA被引次数比较



OA与非OA社交媒体数据比较

4. 科学论文使用大数据：开放获取优势

基于下载、社媒讨论、引用三重指标的论文开放获取优势

Article
Scientometrics
May 2015, Volume 103, Issue 2, pp 555-564

First online: 12 March 2015

The open access advantage considering citation, article usage and social media attention

Xianwen Wang ✉ , Chen Liu, Wenli Mao, Zhichao Fang

Article Metrics

Citations 1

Social Mentions 586

Co-published with

NATUREJOBS | NATUREJOBS BLOG

Open research: Open up to open access

20 Oct 2015 | 12:00 GMT | Posted by Julie Gould | Category: NJCE15, Open access

Six myths about open access were addressed in an open research workshop at the 2015 *Naturejobs* Career Expo in London.

Guest contributor Gaia Donati

How open-minded do you feel about open access publishing?

IMAGE CREDIT: GAIA DONATI

The Open Research workshop at the 2015 London *Naturejobs* Career Expo, led by Mithu Lucraft (head of Open Research Marketing at NPG) and Ros Pyne (Research and Development manager of the Open Research Group at Springer Nature, who manage the Open Research portal), explored several myths about open access publishing, now a well-established alternative route to disseminating scientific results.

Myth 1: Open access benefits readers, but not authors

Open access is great for readers, but the advantage for researchers may seem less obvious at first. A study of open access and subscription-only PNAS articles found that earlier, more frequent citations characterize the former category when compared with the latter. A more recent study of the citations for papers published in *Nature Communications* (before it became fully open access) seems to confirm these findings and extends the observations to downloads and social-media interest, with open access articles experiencing higher downloads. Interestingly, these also appear to be sustained over a longer period of time – "attention lasts longer," said Lucraft. In this way, open access – together with similar initiatives such as open data – may well be a primary route to accelerate and facilitate science while ensuring reproducibility.

about npg

Search

About NPG home > NPG press room > Press release archive

Site content
- About NPG homepage
- Company information
- NPG in the community
- NPG press room
  - Press releases
  - Contact us
- Work @ NPG
- Contact NPG

NPG Journals

by Subject Area

Chemistry
- Chemistry
- Drug discovery
- Biotechnology
- Materials
- Methods & Protocols

Clinical Practice & Research
- Cancer
- Cardiovascular medicine
- Endocrinology
- Gastroenterology & Hepatology
- Methods & Protocols
- Pathology & Pathobiology
- Urology

Earth & Environment
- Earth sciences
- Evolution & Ecology

Life sciences
- Biotechnology

Press release archive

OVER 60% OF 2015 RESEARCH ARTICLES ON NATURE.COM ARE OPEN ACCESS

Nature Publishing Group publishes 63% of research articles via open access models; 96% of authors choose CC BY

20 October 2015

Contact: Amy Bourke-Waite
Senior Communications Manager
Nature Publishing Group/Palgrave Macmillan
T: +44 (0)20 7843 4603 | M: +44 (0) 7703717212
a.bourke-waite@nature.com

Open access is thriving at Nature Publishing Group (NPG). Sixty three per cent of original research articles published to date on nature.com in 2015 are open access, nearly 10,000 papers. Ten years ago, NPG introduced its first fully open access journal. Today, NPG publishes over 80 journals with an open access option.

In January 2015, NPG introduced Creative Commons Attribution license (CC BY) as the default open access license option on its 20+ fully owned open access journals. The percentage of authors choosing CC BY across all of NPG's open access journals has risen dramatically - from 26% in 2014 to 96% in September 2015. Other licenses are still available on demand.

This week is global Open Access Week, and also marks one year since NPG, now part of Springer Nature, announced that *Nature Communications* would become its flagship open access journal.

Sam Burridge, Managing Director, Open Research at Springer Nature said: "We believe we're the first of the longstanding science publishers to reach the landmark of over 60% open access content. By switching *Nature Communications* to full open access one year ago, we demonstrated our willingness to take a bold step and innovate in the open research space, creating a home for the highest quality open research. And we're encouraging our authors to choose more permissive licenses too.

"By combining our portfolio with BioMed Central, Springer Open and Springer Plus, Springer Nature is the largest publisher of open access articles. But we want to lead on offering outstanding service in open access to authors, not just on scale. We also want to lead the research community in innovation, which is why we are prioritising 'open research' - including open data. Our goal is to release the enormous positive power that open approaches can have in facilitating collaborative and interdisciplinary research to solve today's global challenges."

Nature Communications has gone from strength to strength in the last year. It is now the leading open access journal in the multidisciplinary science field,* and number three in its *Journals Citation Report* category after *Nature* and *Science*. Research has also shown that open access articles published in *Nature Communications* are more highly viewed and cited**. Submissions to *Nature Communications* have increased from 1600 per month in 2014 to 2000 per month in 2015.

4. 科学论文使用大数据：封面论文影响力更高？

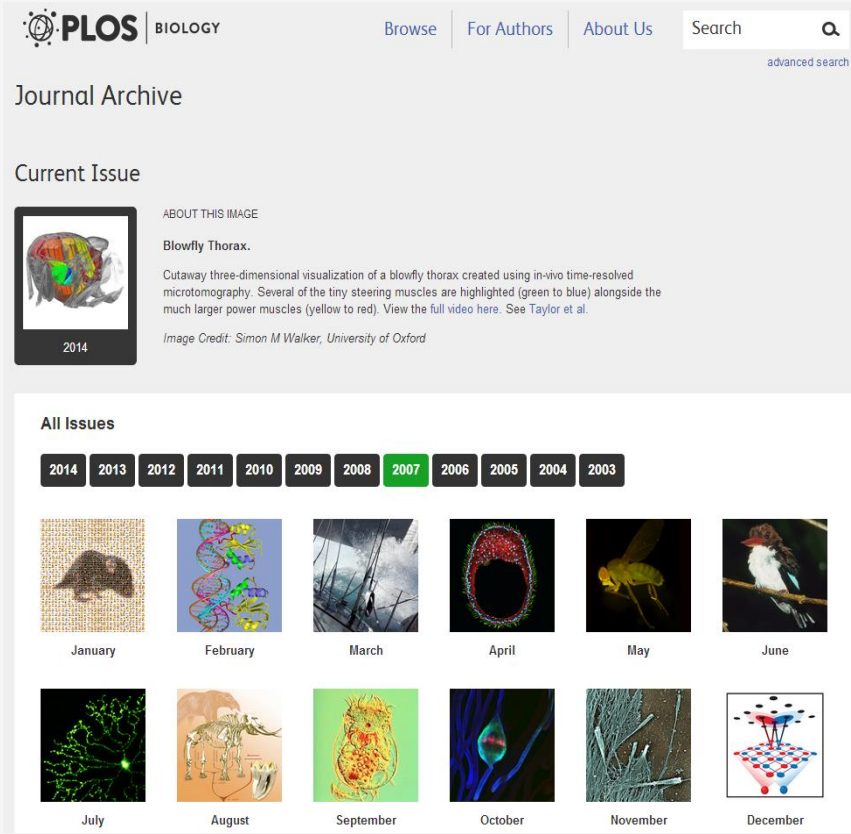
封面论文是否具有更高的影响力？

- **问题的提出**
  - 研究者的论文如果能以封面论文的形式发表，被当成了一种极大的荣耀。许多研究机构和作者甚至会为了庆祝发表一篇封面论文，专门刊出新闻报道。
  - 报道的内容往往暗示读者，其发表的论文被选中作为封面图片，被证明是该期刊同期论文中最好的一篇。
  - 有些基金评审机构会对发表封面论文的申请者有所偏向，有的大学和研究机构会对封面论文作者进行荣誉和物质奖励。
  - 那么一篇论文在被选作封面论文之后，由于封面图片的高显示度，它是否会比同期其他论文获得更高的关注度（下载次数）和影响力（被引次数）。

4. 科学论文使用大数据：封面论文影响力更高？

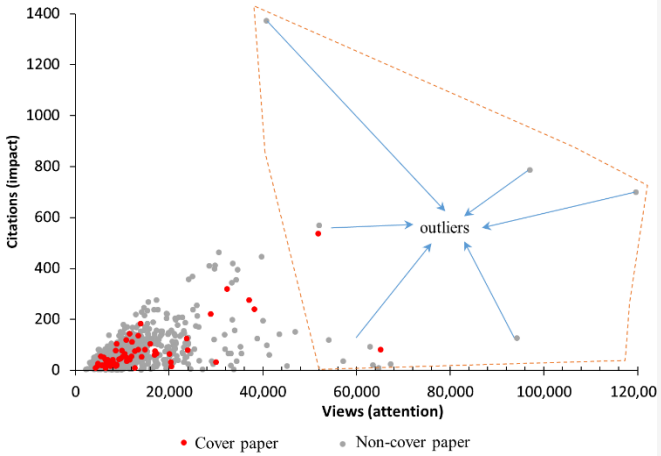
封面论文是否具有更高的影响力

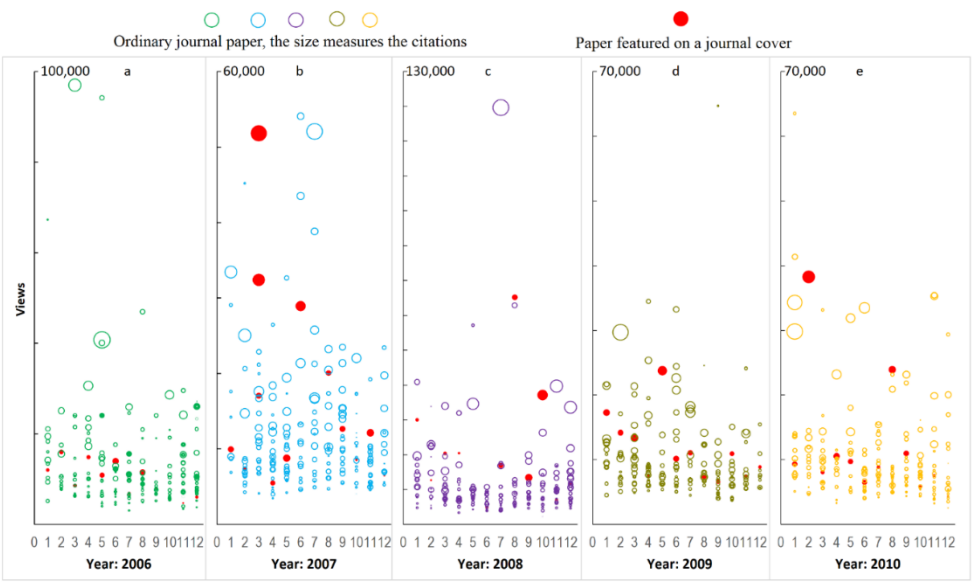
- **研究设计**
  - 选择*PLOS Biology*作为研究对象，收集了2006-2010年一共60期、1025篇研究性论文的文献计量数据，包括论文被下载次数和被引用次数，进行实证分析，包括对比分析和ANOVA分析。



4.科学论文使用大数据：封面论文影响力更高？

封面论文是否具有更高的影响力？答案：NO



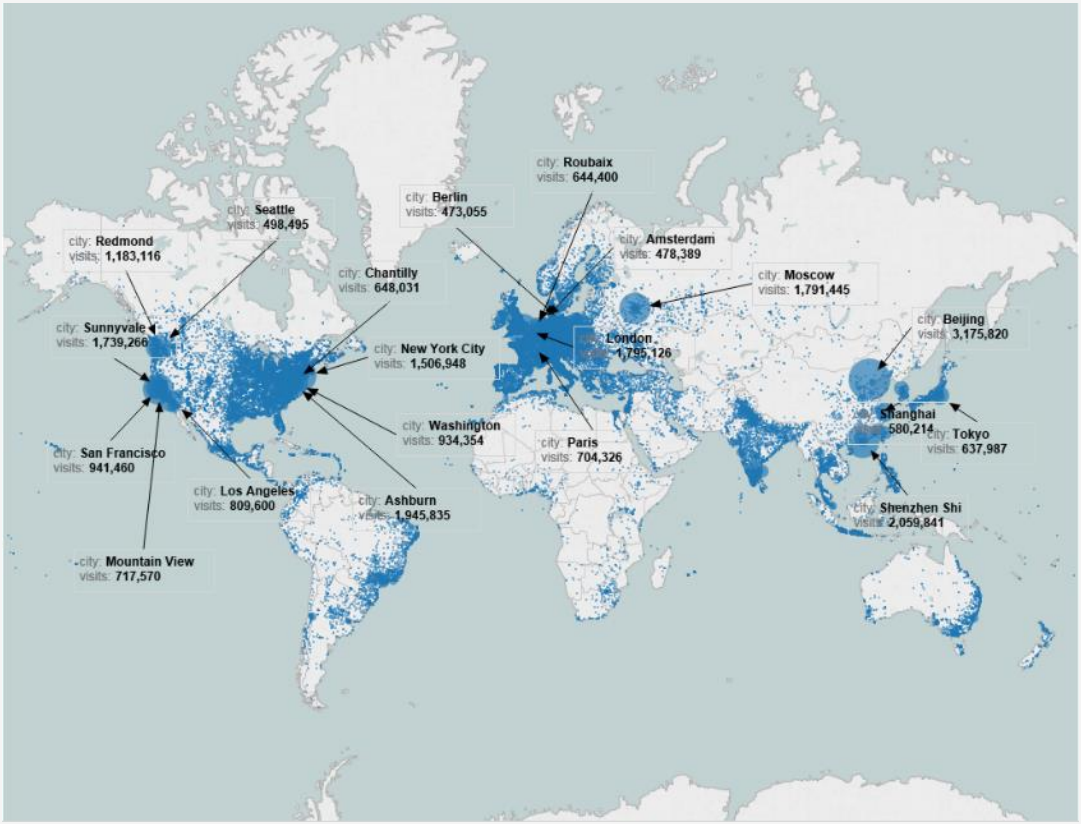


	Mean (SD) 封面论文	Mean (SD) 非封面论文	F	Significance
all_views	14378.37 (11430.899)	11249.54 (9395.644)	6.095	p = .014
all_views adjusted	**12856.21 (7934.183)**	**10600.18 (6611.137)**	**6.213**	**p = .013**
all_citations	74.58 (88.544)	60.99 (79.865)	1.614	p = .204
all_citations adjusted	**66.59 (65.931)**	**59.05 (71.991)**	**.604**	**p = .437**
2006_views	9918.58 (3828.352)	13290.46 (12266.014)	0.897	p = .345
2006_views adjusted	**9918.58 (3828.352)**	**12011.80 (7386.990)**	**.942**	**p = .333**
2006_citations	75.33 (51.286)	98.63 (131.176)	0.373	p = .542
2006_citations adjusted	**75.33 (51.286)**	**94.90 (121.193)**	**.308**	**p = .580**
2007_views	17052.85 (13443.393)	11649.97 (7904.959)	5.175	p = .024
2007_views adjusted	**14154.67 (8833.739)**	**11245.8341 (6789.985)**	**2.009**	**p = .158**
2007_citations	128.69 (148.764)	69.08 (67.070)	7.949	p = .005***
2007_citations adjusted	**94.92 (89.242)**	**66.42 (57.526)**	**2.595**	**p = .109**
2008_views	21792.27 (17255.556)	11387.37 (11225.557)	8.405	p = .004***
2008_views adjusted	**17450.40 (10021.298)**	**10338.62 (6509.176)**	**10.710**	**p = .001*****
2008_citations	64.45 (78.31)	62.05 (75.656)	0.01	p = .919
2008_citations adjusted	**63.00 (82.392)**	**58.81 (60.981)**	**.043**	**p = .835**
2009_views	11542.75 (5028.989)	10054.77 (6757.889)	0.56	p = .455
2009_views adjusted	**11542.75 (5028.989)**	**9754.45 (5415.260)**	**1.237**	**p = .267**
2009_citations	46.25 (30.666)	44.36 (38.631)	0.028	p = .868
2009_citations adjusted	**46.25 (30.665)**	**44.55 (38.643)**	**.022**	**p = .882**
2010_views	11980.33 (9638.092)	10027.81 (7600.706)	0.725	p = .396
2010_views adjusted	**11980.33 (9638.092)**	**9761.57 (6608.291)**	**1.206**	**p = .273**
2010_citations	52.83 (61.707)	34.33 (43.816)	1.922	p = .167
2010_citations adjusted	**52.83 (61.707)**	**34.41 (43.912)**	**1.897**	**p = .170**

4.研究主线：科学论文的下载分析

- 基于全球论文下载的地域数据，测度全球研发活动强度的地区分布

- Measuring Global Research Activities Using Geographic Data of Scholarly Article Visits，in progress



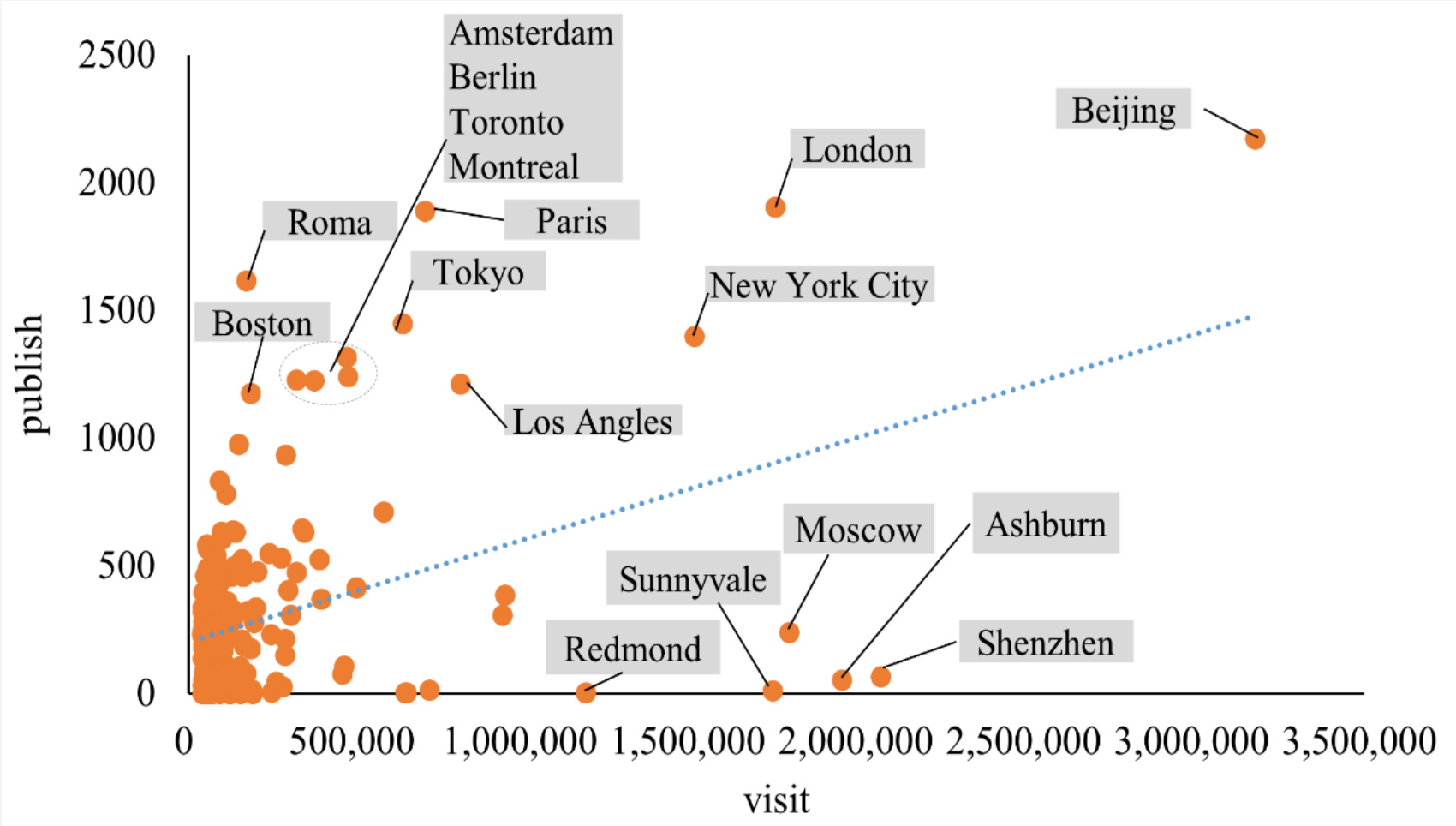
4. 科学论文使用大数据：全球R&D活动测度

基于全球论文下载的地域数据，测度全球研发活动强度的地区分布

排序	城市	国家	论文访问量	论文发表量
1	北京	中国	3,175,820	2,171
2	**深圳**	**中国**	**2,059,841**	**65**
3	Ashburn	美国	1,945,835	53
4	莫斯科	俄罗斯	1,791,445	237
5	伦敦	英国	1,795,126	1,904
6	Sunnyvale	美国	1,739,266	11
7	纽约	美国	1,506,948	1,398
8	Redmond	美国	1,183,116	1
9	旧金山	美国	941,460	384
10	华盛顿	美国	934,354	307

4. 科学论文使用大数据：全球R&D活动测度

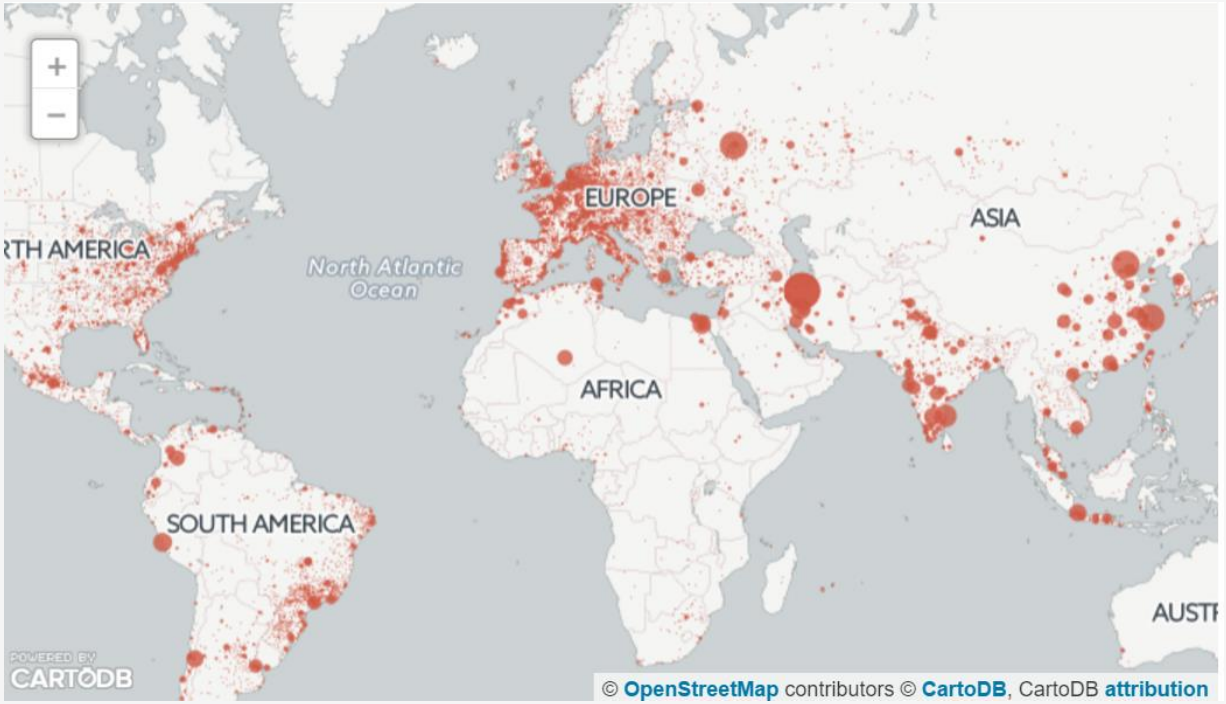
基于全球论文下载的地域数据，测度全球研发活动强度的地区分布



4. 科学论文使用大数据：全球R&D活动测度

- **基于全球论文下载的地域数据，测度全球研发活动强度的地区分布**

- **Science文章：**Bohannon, J. (2016). Who's downloading pirated papers? Everyone. Science, 352(6285). http://www.sciencemag.org/news/2016/04/whos-downloading-pirated-papers-everyone

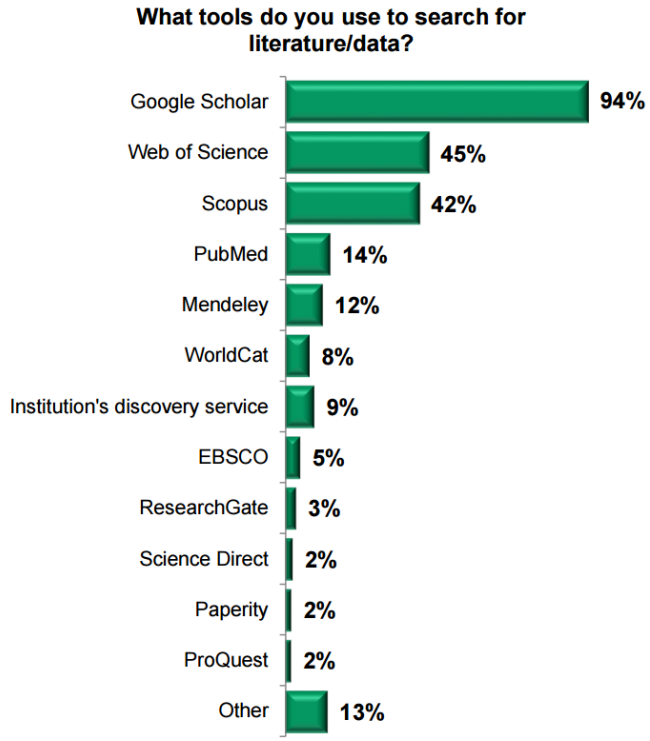


4. 科学论文使用大数据：追踪科学论文的访问足迹

Emerald对1034名研究者的问卷调查结果

Discovery

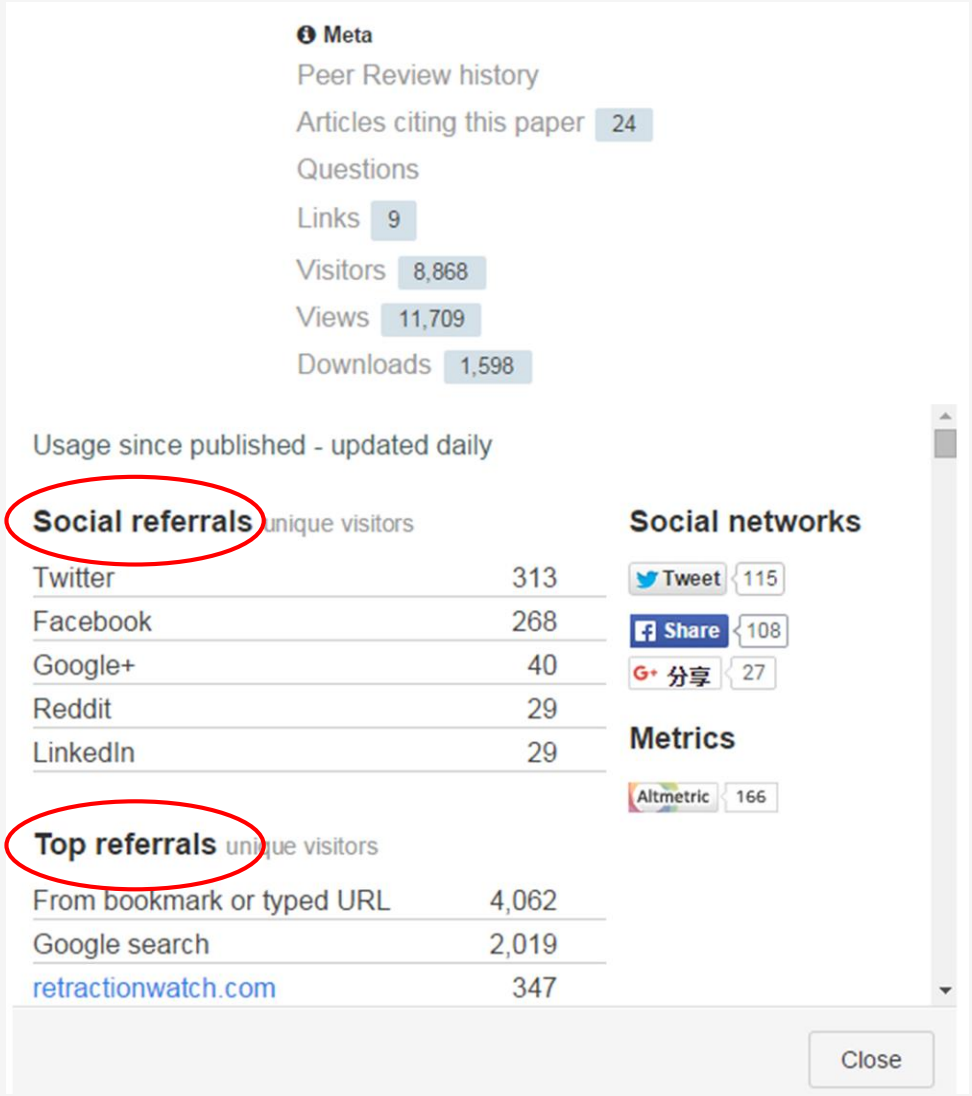
Google Scholar is dominant as the key content search tool for over 90% of Emerald researchers



Base [all authors who use these tools] = 1034

- Almost every respondent used Google Scholar to search for scholarly content.
- Web of Science and Scopus lead a long tail of miscellaneous search tools and databases referred to by scholars.
- Whilst Google was the most frequently used tool by authors in all subject groups, authors in Medicine, Life and Physical Sciences, and Law were more likely to use WorldCat and PubMed.
- Librarians also reported higher levels of use of PubMed (48% of Librarians surveyed) and WorldCat (34%) than those in other roles.

4. 科学论文使用大数据：追踪科学论文的访问足迹



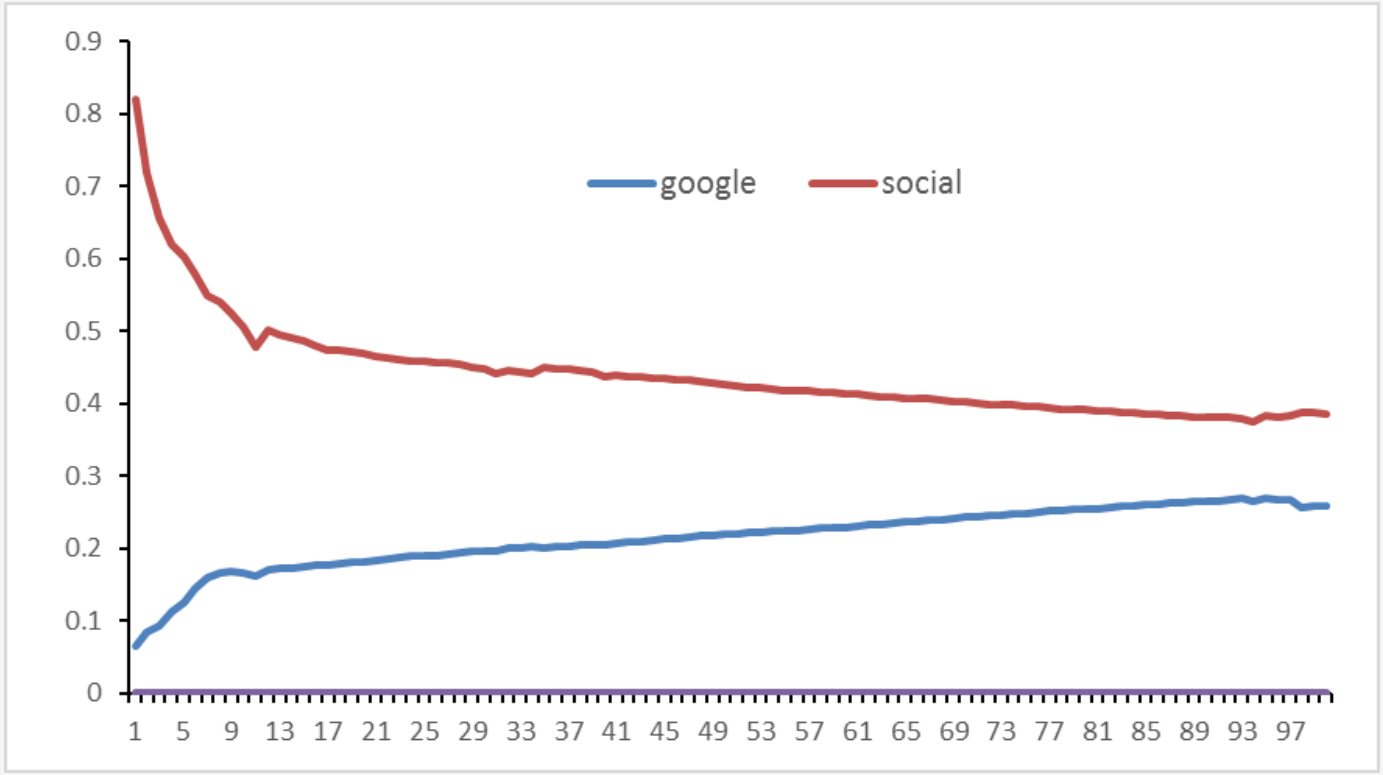
PeerJ

- 始于2013年2月
- **开放获取**的同行评议期刊
- **生物学和医学领域**
- 2015年影响因子2.112
- *PeerJ*提供了每篇论文详细的**使用数据计量**，其中便包括总**论文指引链接数据**（Top Referrals）和**社交媒体指引数据**（Social Referrals）。

4. 科学论文使用大数据：追踪科学论文的访问足迹

排序	指引链接	访问量	比例
1	From bookmark or typed URL	1,439,085	42.68%
2	Google	844,189	25.03%
3	Facebook	181,434	5.38%
4	Twitter	139,323	4.13%
5	Reddit	99,396	2.95%
6	www.ncbi.nlm.nih.gov	95,162	2.82%
7	From PeerJ Content Alert Emails	47,601	1.41%
8	Yahoo	14,043	0.42%
9	Web of Knowledge	12,901	0.38%
10	Bing	10,573	0.31%

4. 科学论文使用大数据：追踪科学论文的访问足迹



来自Google搜索与社交媒体的流量的时间变化趋势，二者刚好相反。

05

PART FIVE

地理位置大数据

LBS Big Data

5.地理位置大数据： 春运人口流动

- **春运**
  - 自20世纪80年代末以来，一场前所未有的劳动力从农村向城市转移的景况在中国出现。这种转移为中国经济的迅猛增长做出了巨大贡献。
  - 也形成了一年一度的春运，这是一场全球最大规模的人口季节性迁徙活动。
  - 正是由于这场浩荡迁徙的宏大规模和其中饱含的浓浓情思，使得春运每年都受到全国乃至全世界的关注。
- **人口流动**
  - 对于人口迁徙的研究，以往研究集中在识别主要的迁徙起始地与目的地，以及影响迁徙的社会经济因素等问题上。
  - 几乎过去所有的研究都采用了多样化的人口迁徙模型来对普查数据进行分析。

5.地理位置大数据： 春运人口流动

- **春运**
  - 由于普查数据的可用性、完整性和一致性的限制，要对人口迁徙的路径、模式以及时间与空间趋势等方面进行深入研究分析几乎是不可能办到的。
  - 智能手机为我们提供了一种收集人类地理位置信息的新途径，那就是空间定位服务（LBS）。
  - 实现全国范围内人类行为的模拟和数据处理在过去看来简直是天方夜谭，但“大数据”将其变为了现实。

5.地理位置大数据： 春运人口流动

- **春运**
  - 由于普查数据的可用性、完整性和一致性的限制，要对人口迁徙的路径、模式以及时间与空间趋势等方面进行深入研究分析几乎是不可能办到的。
  - 智能手机为我们提供了一种收集人类地理位置信息的新途径，那就是空间定位服务（LBS）。
  - 实现全国范围内人类行为的模拟和数据处理在过去看来简直是天方夜谭，但“大数据”将其变为了现实。

5.地理位置大数据： 春运人口流动

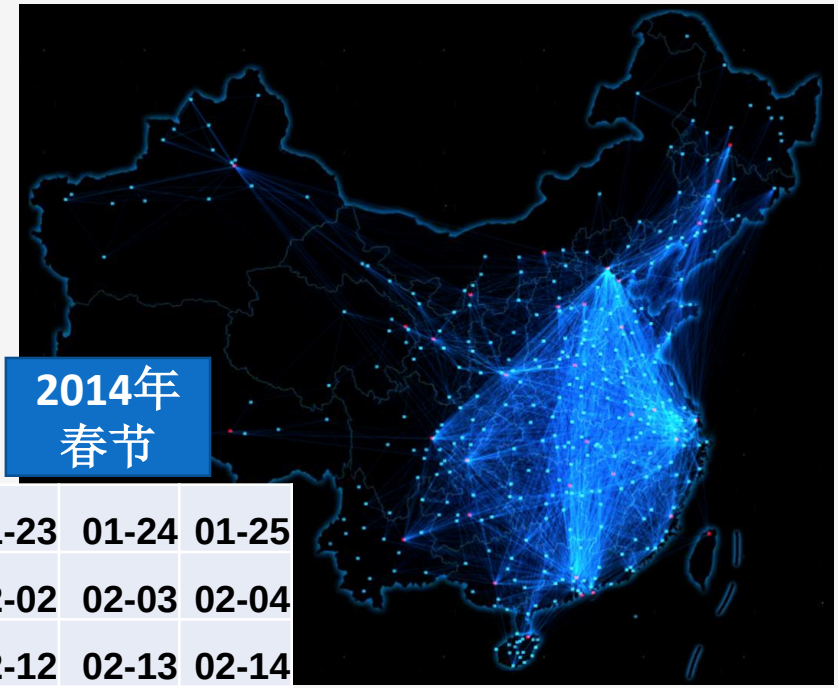
- **研究问题**
  - 我国春运期间的大规模人口流动的时空规律？
  - 这些规律又反映了关于地区间发展状况的怎样的事实？

5.地理位置大数据： 春运人口流动

- **数据来源**
  - 百度公司

    qianxi.baidu.com
  - 数据收集的时间窗口

    2014年1月16 - 2月18日

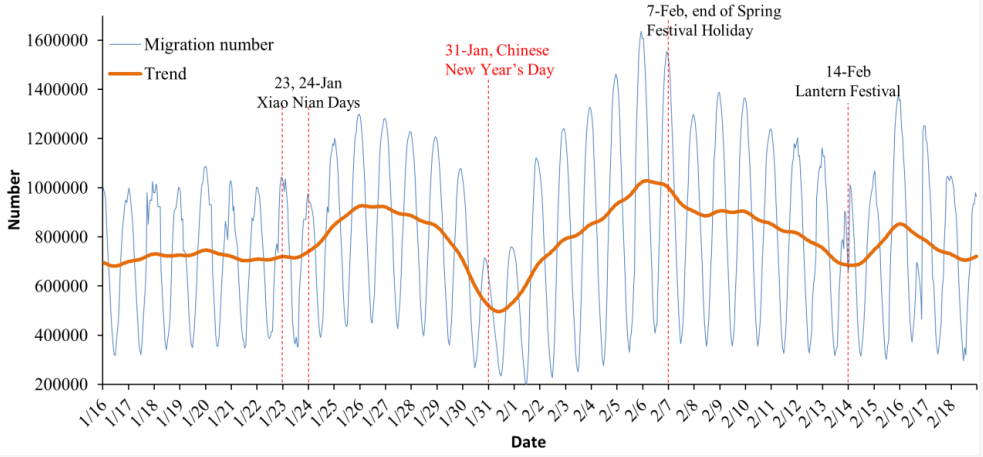


01-16	01-17	01-18	01-19	01-20	01-21	01-22	01-23	01-24	01-25
01-26	01-27	01-28	01-29	01-30	01-31	02-01	02-02	02-03	02-04
02-05	02-06	02-07	02-08	02-09	02-10	02-11	02-12	02-13	02-14
02-15	02-16	02-17	02-18						

5.地理位置大数据： 春运人口流动

- **结果分析**
  - 时间趋势
    - 由于每小时数据曲线波动较为剧烈，因此我们选用HP滤波得到的趋势曲线来进行分析。

- 1月16-1月23号期间趋势曲线较为平坦。
- 小年后，曲线快速爬升并在1月26号达到高峰，然后开始下降，在除夕到达最低点。
- 2月6号（春节长假最后一天）曲线达到年后高峰。
- 2月14日元宵节的低谷过后，2月16日又达到小高峰。

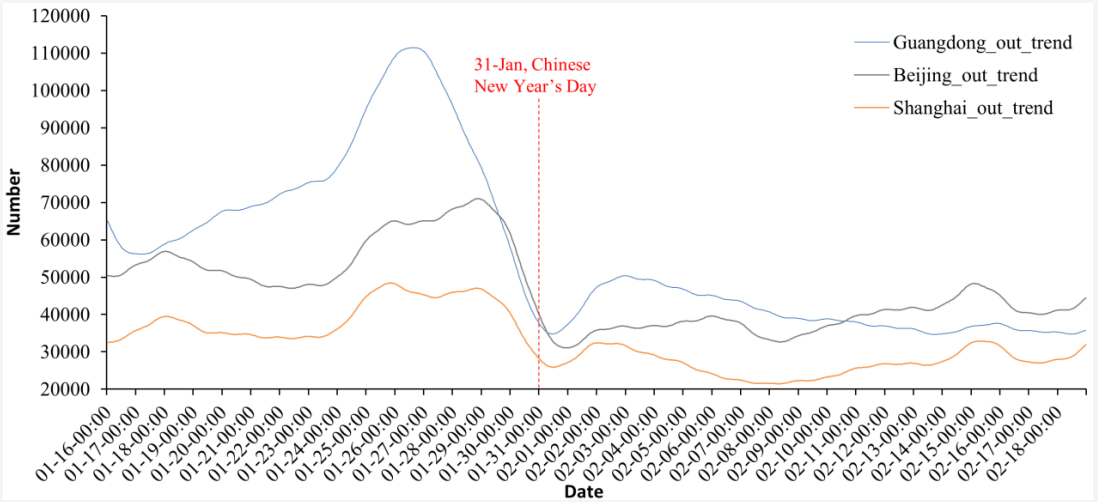


春运期间我国人口流动数量的时变趋势

5.地理位置大数据： 春运人口流动

- **结果分析**
  - 流入和流出
    - 广东省的人口迁徙曲线很早就开始攀升。从小年开始，广东的曲线开始进入另一个快速增长期并于1月26号达到顶峰。

- 北京和上海的曲线走势很相似。人口迁出的高峰在过年前两天左右。
- 以春节作为3条曲线的分界点，分界点左半部分显著高于右半部分。



广东、北京和上海的人口流出时变趋势

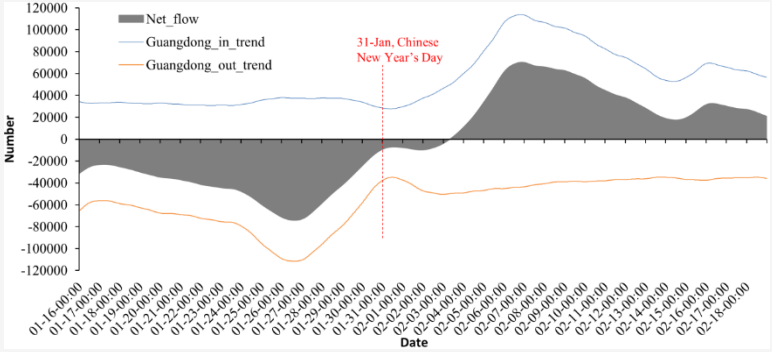
5.地理位置大数据： 春运人口流动

- **结果分析**
  - 流入和流出

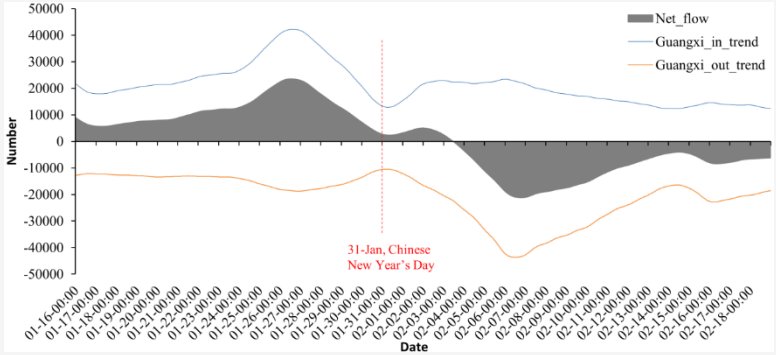
- 流入人口量设置为正值（蓝色曲线），流出人口量设置为负值（橙色曲线），净流动数据则由流入人口量减去流出人口量得到（灰色阴影区域）。

- 广东和广西的人口流入和流出的时间变化刚好相反，并且互为目的地的曲线刚好与上两图吻合。

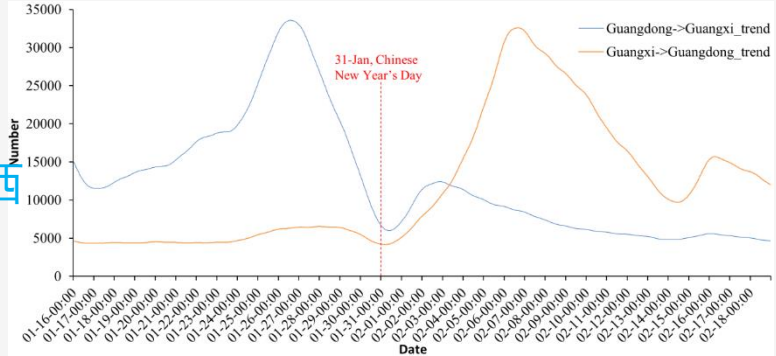
广东



广西



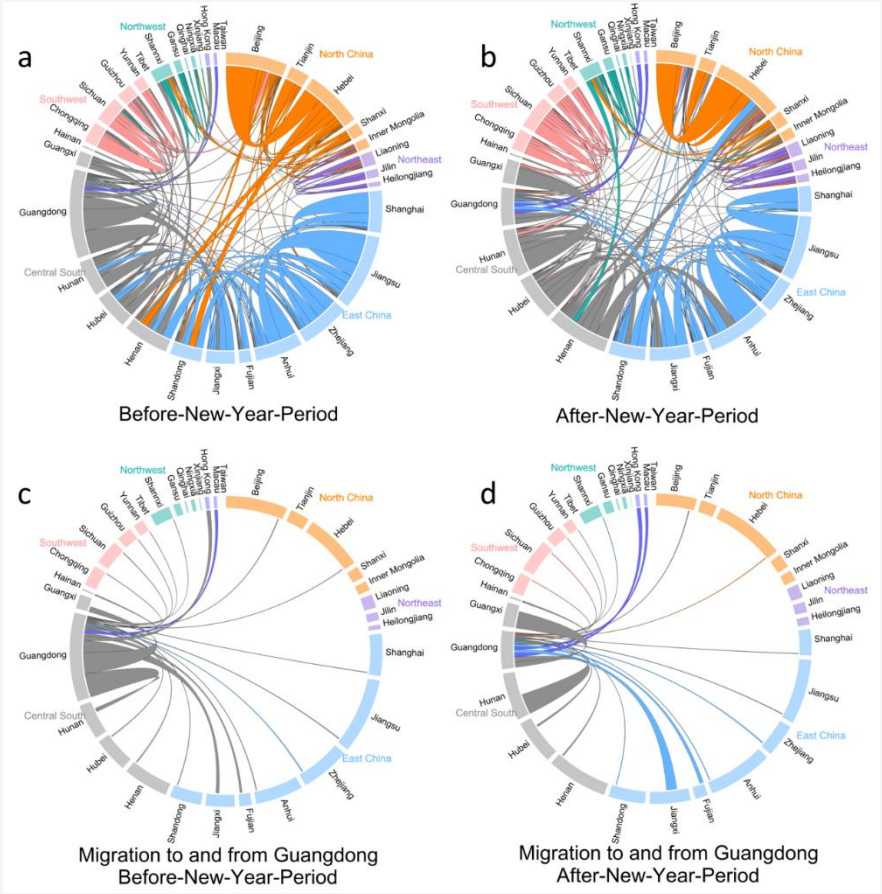
广东广西双向



5.地理位置大数据： 春运人口流动

• 迁徙网络分析

- 广东和广西的人口流入和流出的时间变化刚好相反，并且互为目的地的曲线刚好与上两图吻合。
- （a）表示春节前的人口迁徙网络，在此期间，人们返回自己的家乡。（b）是春节后时段的人口迁徙网络，人们在这段时期返回工作和学习的地区。
- 如果我们将鼠标移动到交互图中广东省的边缘，流入和流出广东的情况就会被突出显示，如（c）和（d）所示。

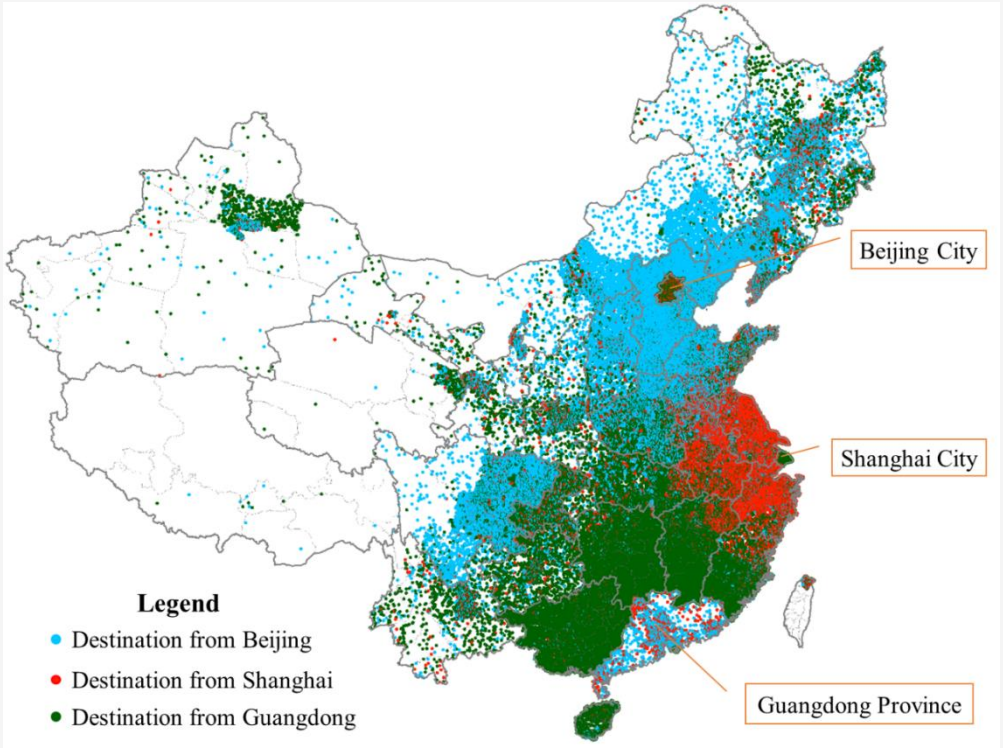


- **交互可视化网络地址**：http://xianwenwang.com/research/mig/

5.地理位置大数据： 春运人口流动

• 人口流向

- 绿色、红色和蓝色分别表示从广东、上海、北京迁出的人口所前往的目的地。
- 迁徙人口的起始地和目的地有明显的地域邻近特征。



春节前从北京、上海和广东流出人口的迁徙目的地

- **交互可视化网络地址**：http://xianwenwang.com/research/mig/

5.地理位置大数据： 春运人口流动

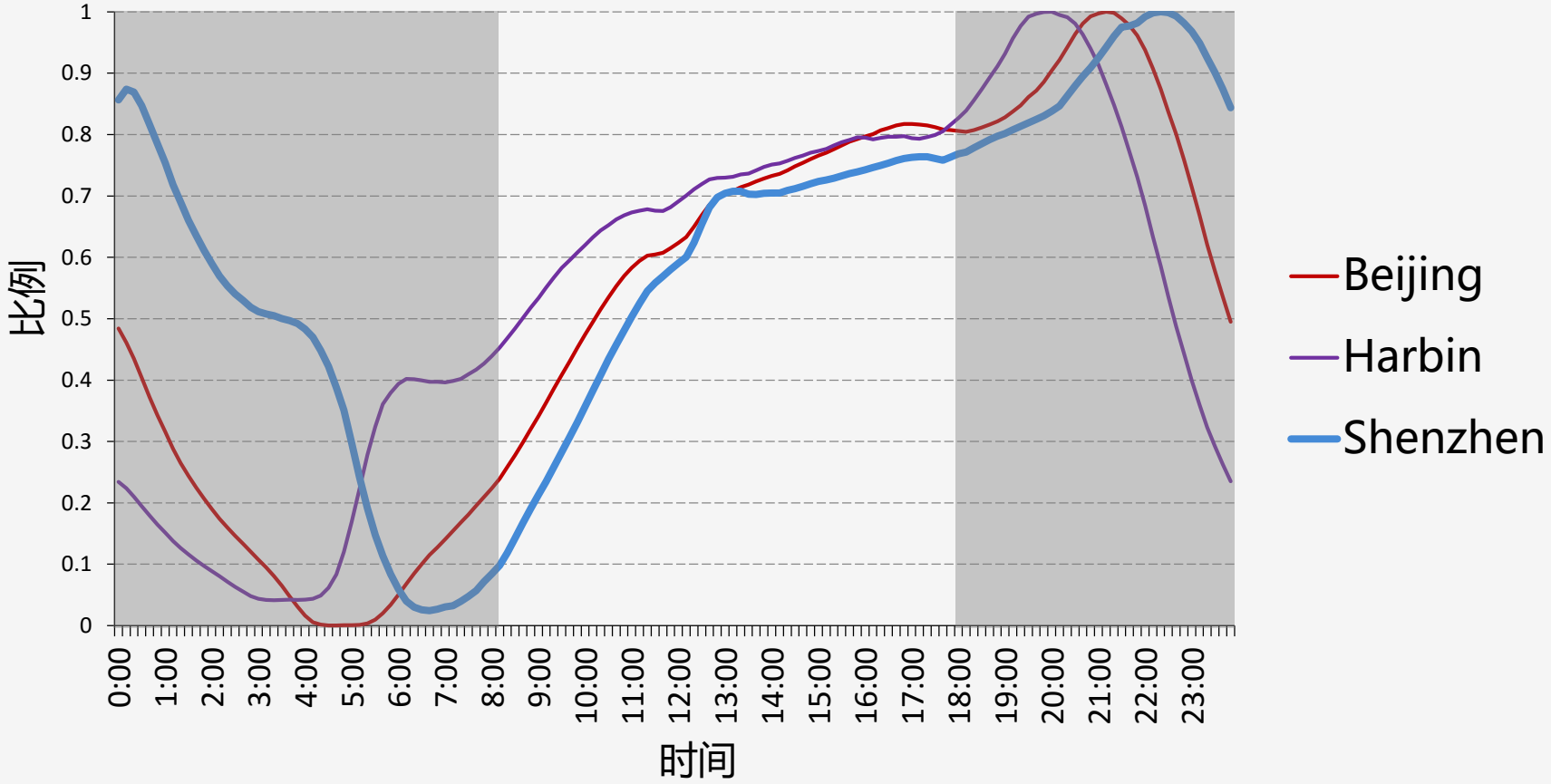
- **结论**
  - 时间角度
    - 以关键日期为分界点，人口迁徙表现出显著的规律性。
  - 空间的角度
    - 发达地区影响力的空间范围也通过人口迁徙的目的地表现出来，即使是中国最发达的地区，人口迁徙的起始地和目的地也具有明显的地理邻近特征。

5.地理位置大数据： QQ游戏数据

- 每时每刻每一个地方的游戏玩家数量

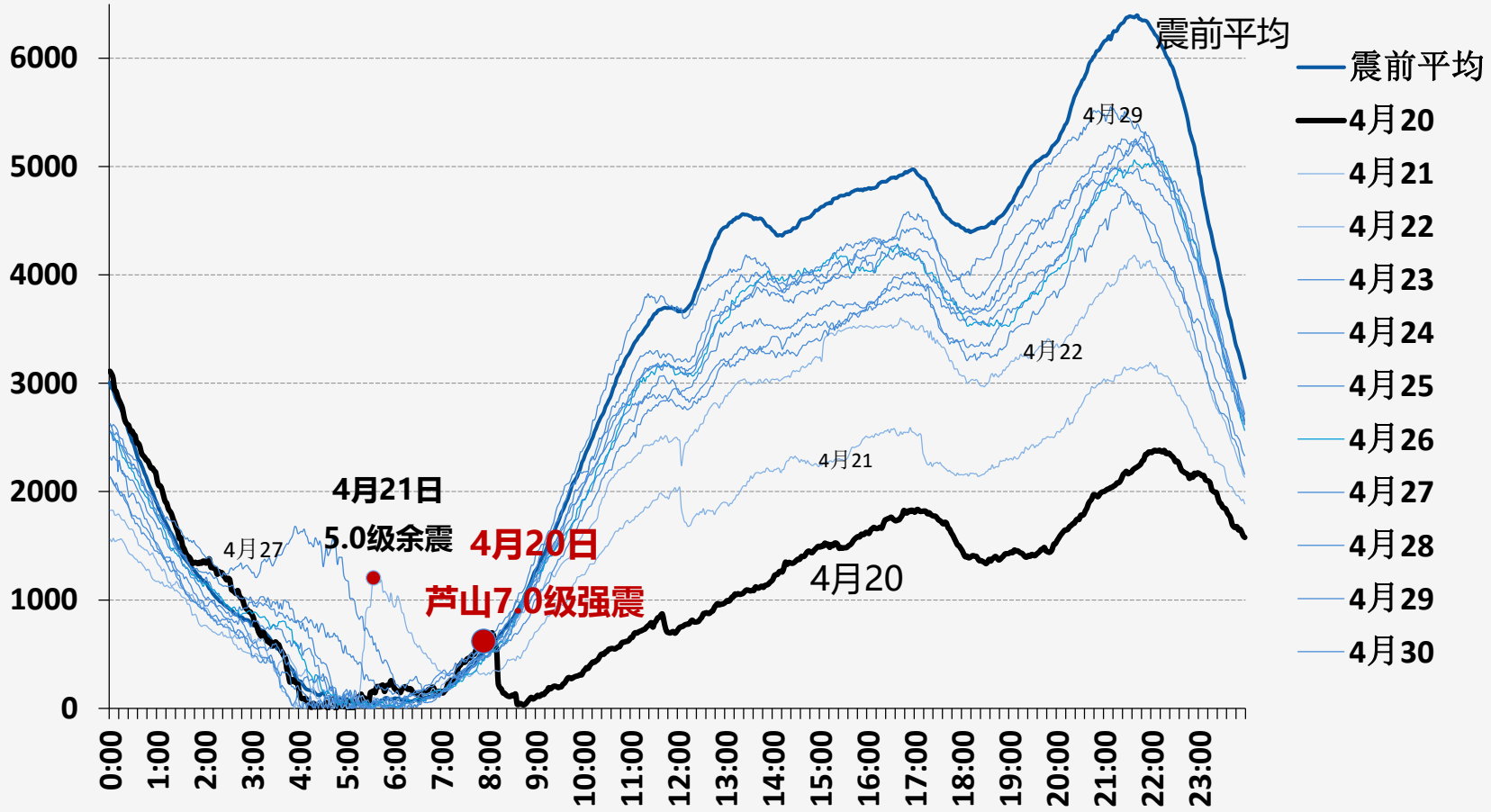
5.地理位置大数据： QQ游戏数据

- 每时每刻每一个地方的游戏玩家数量



5.地理位置大数据： 地震影响

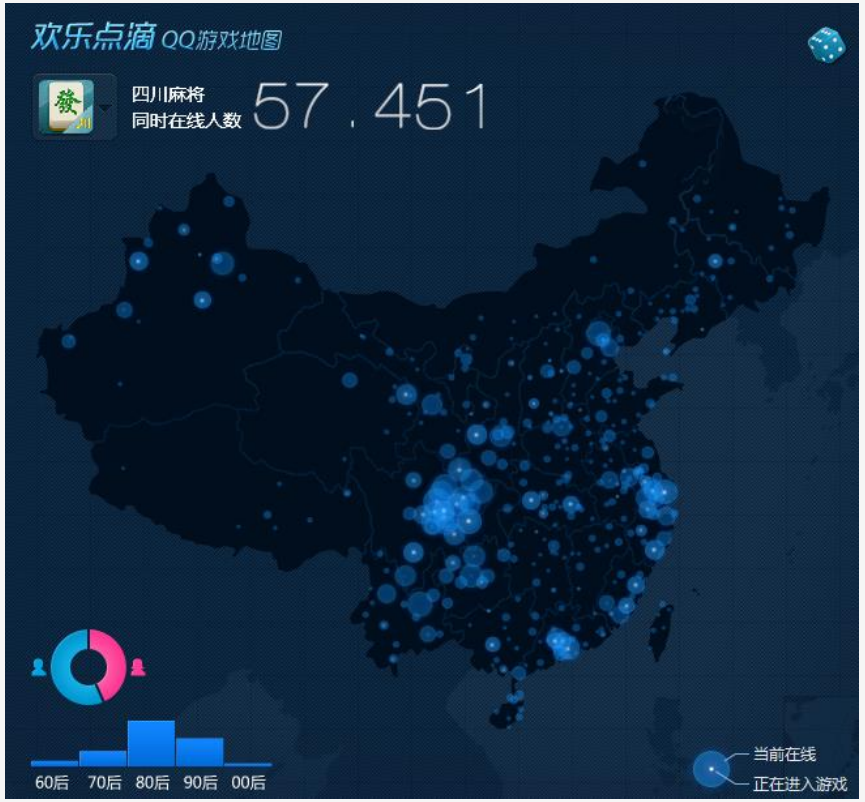
- 2013年4月20日四川省雅安地震



5.地理位置大数据： 中国地域文化

- **人们对于地域的划分**
  - 北方、南方
  - 东北、华南……

5.地理位置大数据： 中国地域文化

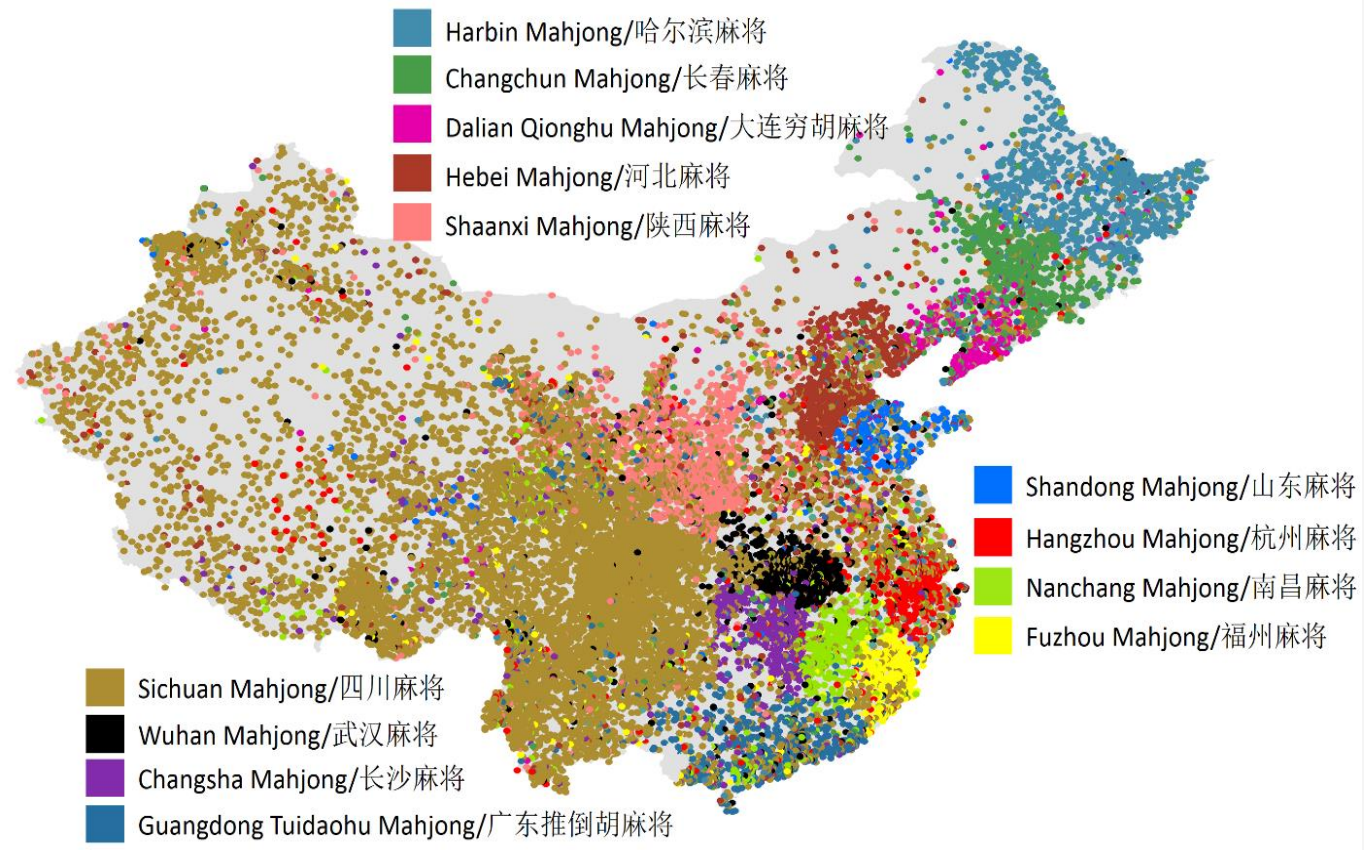


5.地理位置大数据： 中国地域文化

- **以地域特色的纸牌和棋牌游戏作为变量**
  - 我们观察到，不同地域的人喜欢玩不同的游戏，许多纸牌和棋牌游戏具有明显的地域分布特征。
  - 例如山东人喜欢玩山东麻将、保皇，安徽人喜欢玩惯蛋，四川、重庆、云南、甚至湖北西部的居民玩四川麻将等等。

5.地理位置大数据： 中国地域文化

- 麻将游戏的地域分布



5.地理位置大数据： 中国地域文化

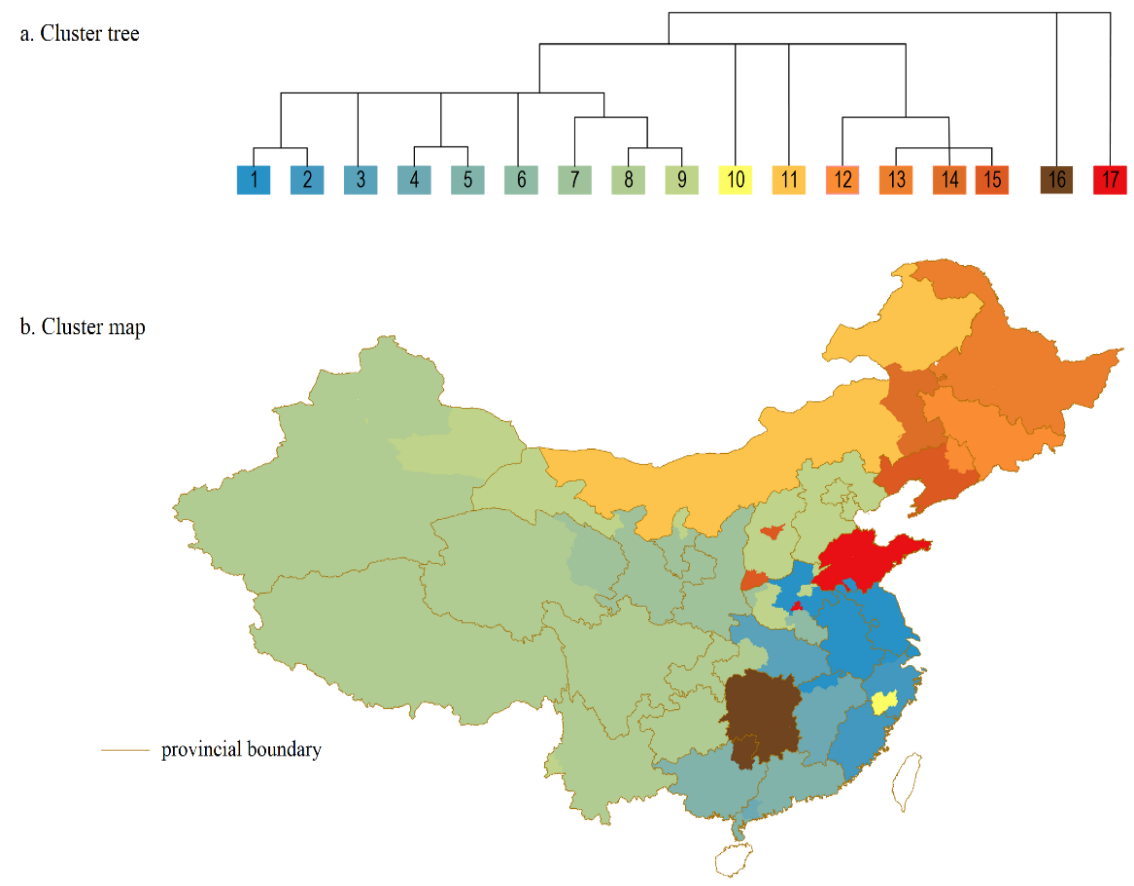
- **变量的选择与确定**

地区	游戏1	游戏2	游戏3
北京市	敲三家		
天津市	憋七	天津麻将	敲三家
辽宁省	刨幺	打滚子	四冲
山西省	太原立四麻将	撾龙	
内蒙古	打大A	赤峰对调	
广东省	**广东鸡平胡麻**	**广东推倒胡麻将**	
…	…	…	…

- 于是，我们选择了47种带有地域特色的游戏作为变量，采集了“什么地方的人在玩什么游戏”的数据。

5.地理位置大数据： 中国地域文化

- **聚类结果**
  - 对中国300多个地市级地区用这47个变量进行聚类分析，并将聚类结果映射到中国地图上。



感谢各位！

Thanks for Listening

王贤文
大连理工大学