BCL 2015年会

社交媒体地理：分析框架与应用

王 江 浩

(wangjh@lreis.ac.cn)

中国科学院地理科学与资源研究所
资源与环境信息系统国家重点实验室

2015-06-06 @ 北京交通大学

Contents

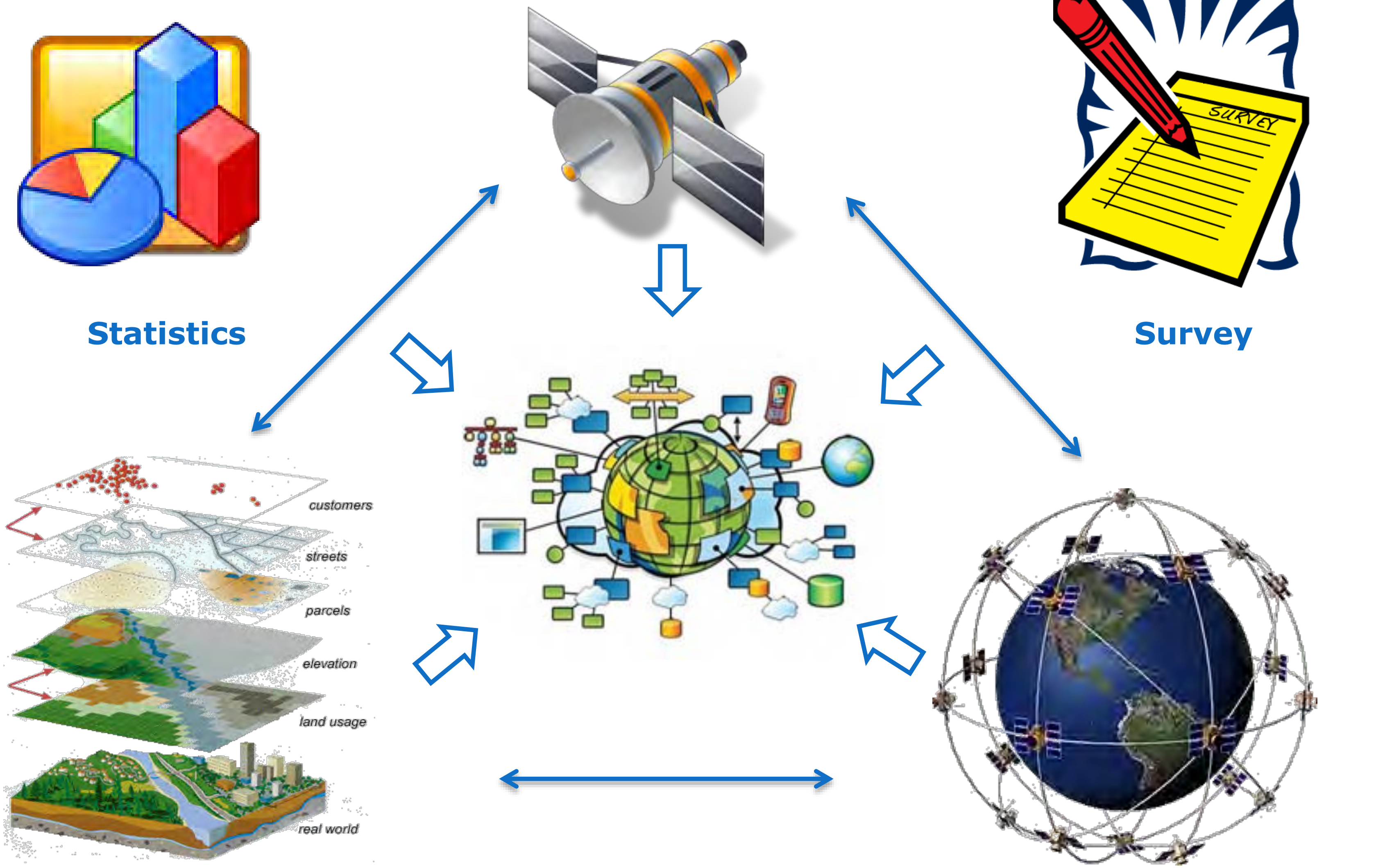
社交媒体地理的概念与分析框架

应用1：Tracking Human Mobility and Cultural Ties

应用2：Where are the Chinese?

1 社交媒体地理的概念与框架

传统空间数据获取分析方式



Remote Sensing (RS)

Statistics

Survey

Geographical Information System (GIS)

Global positioning System (GPS)

大数据时代的时空数据获取

□大数据与开放数据

□Data Driven

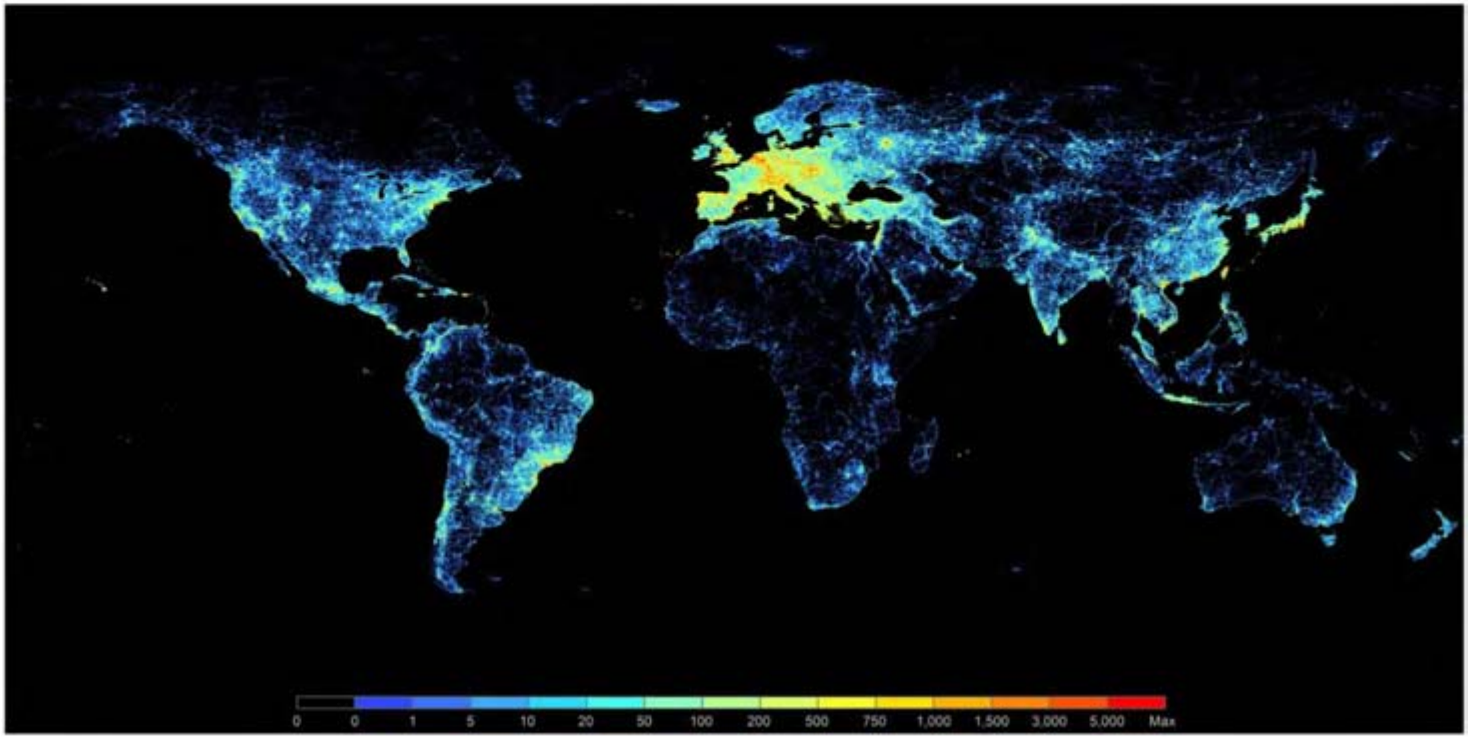
□VGI：Citizens as sensors: the world of volunteered geography.

—— Michael F. Goodchild，2007

- 随着计算机技术，GPS，移动终端技术发展，在互联网Web2.0驱动下， 传统地理信息由单向方式逐渐向交互双向协作方向发展。

社交媒体数据

□ 微博、微信、facebook，twitter，flickr，人人网，豆瓣、婚恋交友等



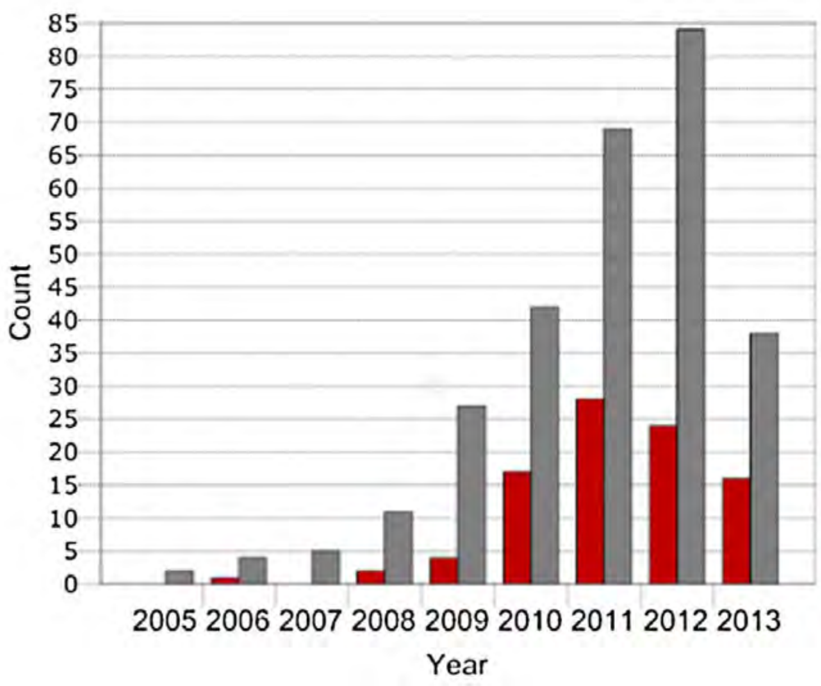
基于位置的社交媒体

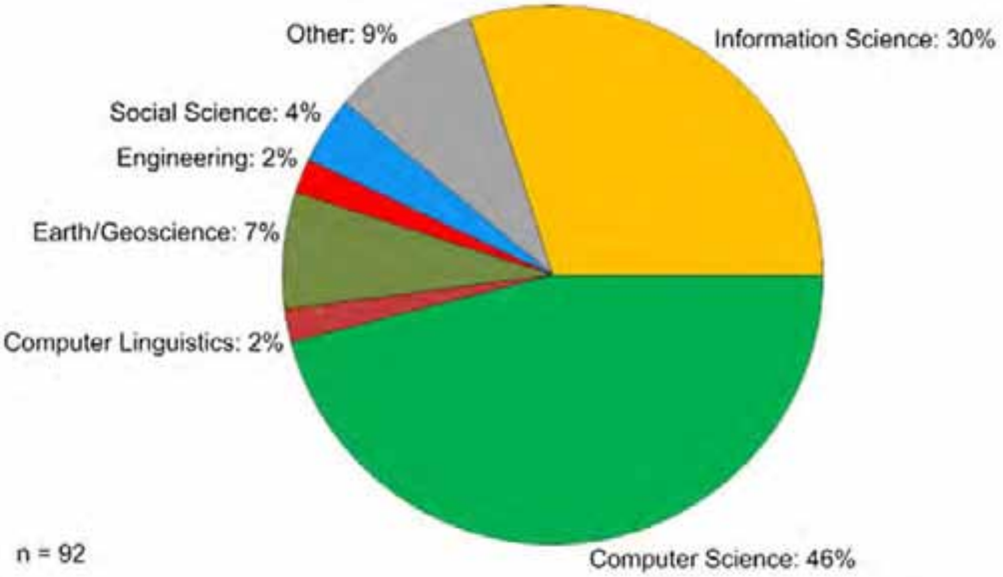
- 包含时空信息（**XYT**）和语义信息的三层结构

  (1) a social network (user layer);

  (2) a geographical network (location layer);

  (3) a semantic metadata network (content layer).

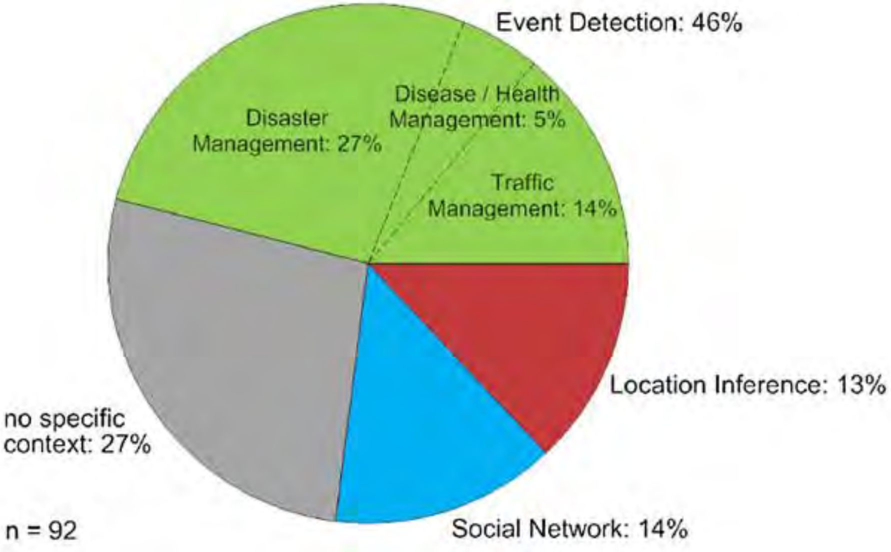




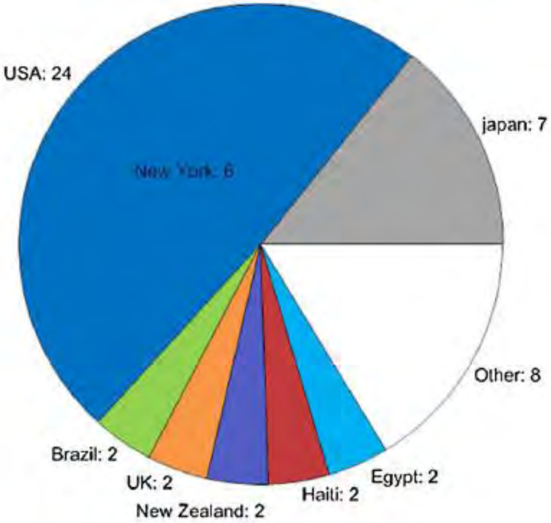
作者学科背景

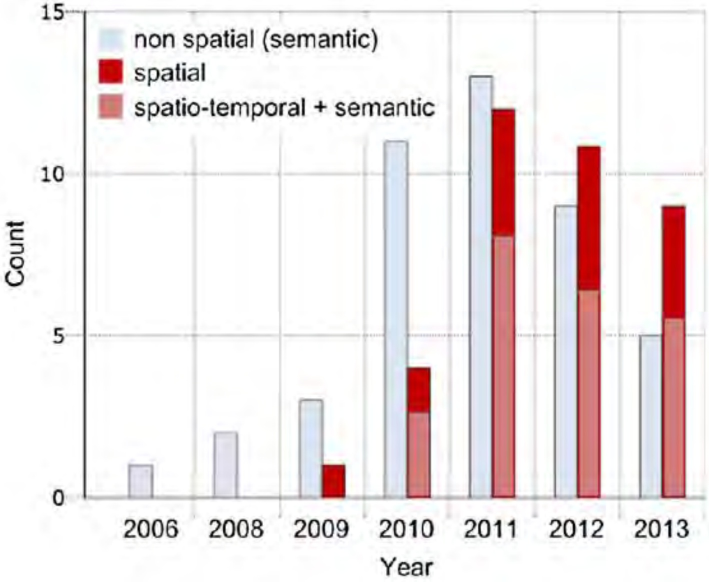
基于位置的**Twitter**研究（**Steiger, 2015, TGIS**）

基于位置的 Twitter 研究综述

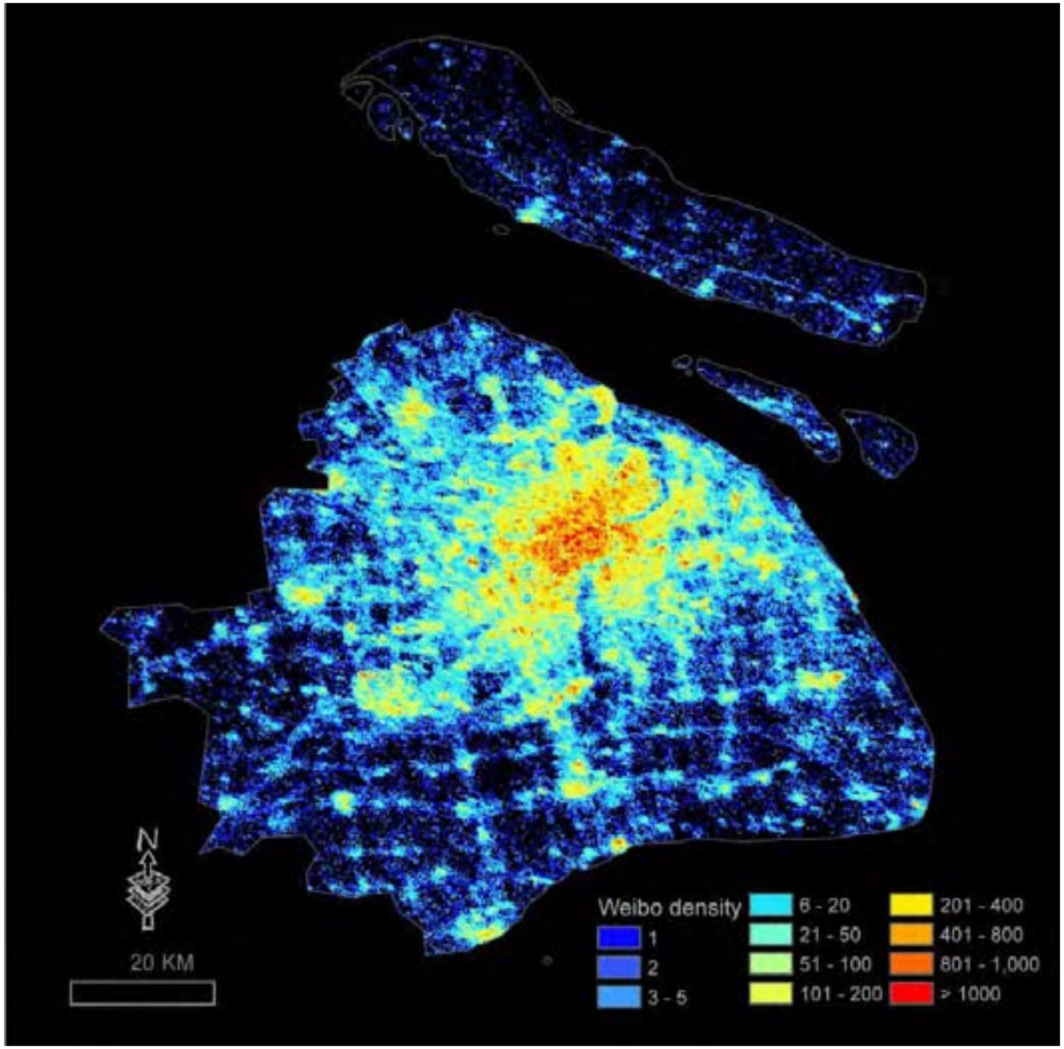


- Event detection
  - Disaster management
  - Disease/Health
  - Traffic
- Location Inference
- Social Network





社交足迹地理学研究



上海微博空间分布

1. 城市活力研究
2. 突发事件监测与预警
3. **微观城市人口社会经济模拟**
4. 人群时空间行为研究
5. **城市空间流动与相互作用机制**
6. 城市规划方案设计与优化
7. ……

技术方法：分析流程与工具

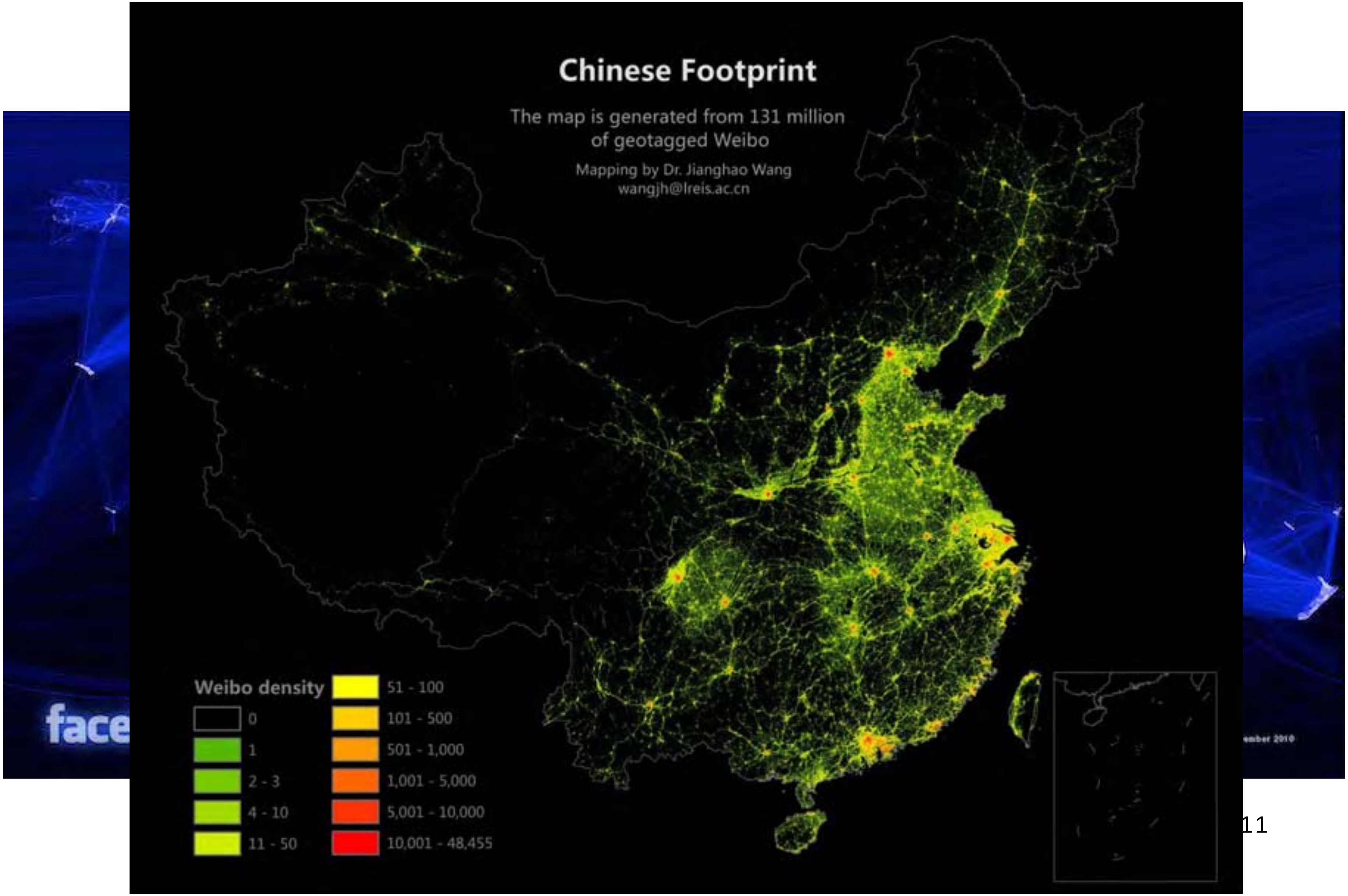
- □ **数据获取**
  - ■ 爬虫系统

- □ **空间数据库**

- □ **时空数据分析与挖掘**

- □ **数据可视化**

社交媒体数据代表性问题



应用1 人口流动、文化与城际联系

Jianghao Wang
IGSNRR, CAS, CN

Wenjie Wu
Heriot-Watt University, UK

Weiyang Zhang
GaWC, Belgium

- Periphery to Core: Mining China's Urban Social Interaction Footprint Patterns Using Big Data
- The Geography of Cultural Ties and Human Mobility: New Evidence based on Social Media from China
- Assessing spatial patterns using Chinese location-based social media: the case of Weibo-users' intercity connections in the Yangtze River Delta

研究背景

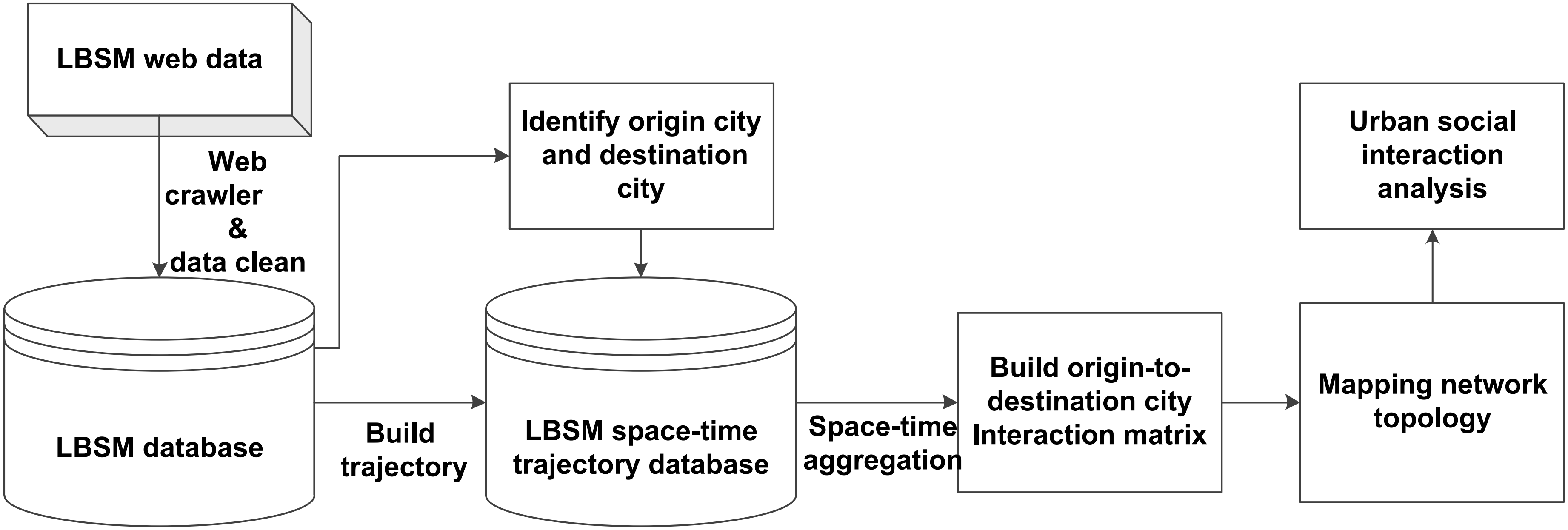
采用社交网络大数据构建人口流动与城际联系网络，分析和评价城市联系强度与影响要素。

- 传统的统计数据并没有提供城际联系、人口流动的普查数据
- Weibo 等社交网络记录人的流动轨迹，为研究城际联系提供了一个新的视角
- 通过与方言文化的比较，分析人口流动的驱动力

研究回顾：

- 传统城际联系一般是基于交通流数据、人流调查，也有基于LBS的方式（如百度迁徙）、但尚无基于社交媒体的中国区域内城际联系研究
- Twitter等社交媒体数据已广泛应用与社交网络数据挖掘、公共健康预警、自然灾害过程识别、城市活力研究、以及人口流动研究；
- 中国范围内由于数据局限，研究受限

数据获取与处理流程


LBSM web data
Web crawler & data clean
LBSM database
Identify origin city and destination city
Build trajectory
LBSM space-time trajectory database
Space-time aggregation
Build origin-to-destination city Interaction matrix
Mapping network topology
Urban social interaction analysis

Weibo重要属性：longitude、latitude、sendTime、Content、registration

构建**trajectory**：$WT_i=\{(s_i^j, t_i^j, c_i^j),(s_i^{j+1}, t_i^{j+1}, c_i^{j+1}),\ldots(s_i^{j+k}, t_i^{j+k}, c_i^{j+k})\ldots\}$

统计人口流动：$OutC(p_m) = \sum_{i=1}^{k} OutC(p_{m,i}) - \sum_{i=1}^{k} F(p_{m,i}) \quad InC(p_m) = \sum_{i=1}^{k} InC(p_{m,i}) - \sum_{i=1}^{k} F(p_{m,i})$

$$F(p_1, p_2) = \sum_{i=1}^{k} \sum_{j=1}^{k} F(p_{1,i}, p_{2,j})$$

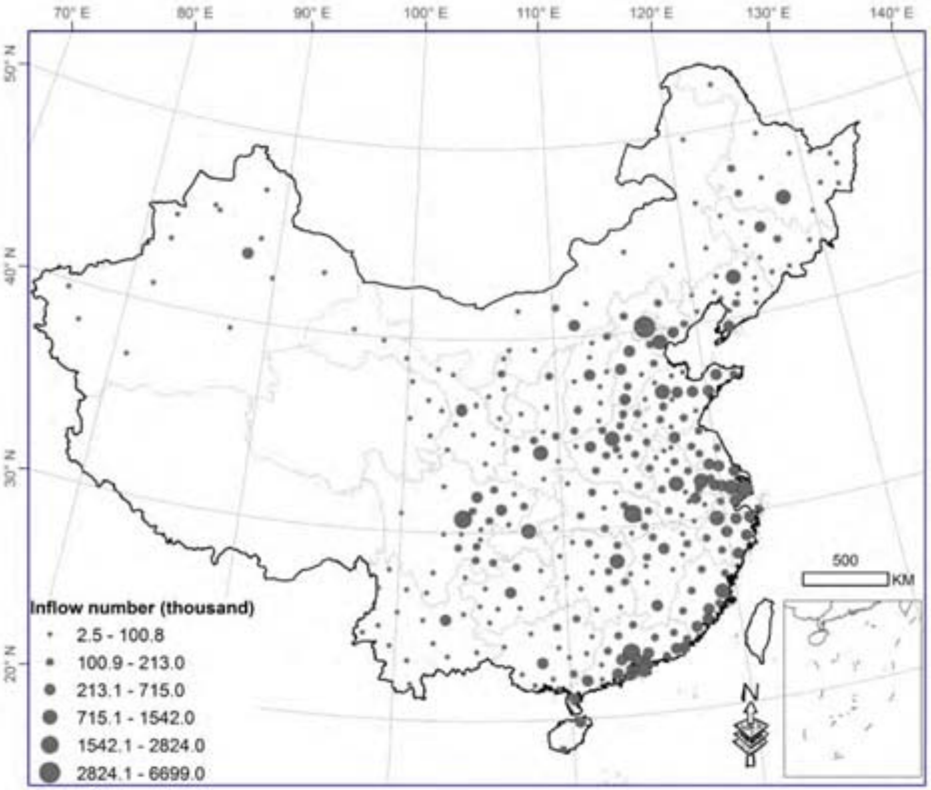
构建**origin-to-destination city matrix：328*328**

构建人口流动与城际联系指标

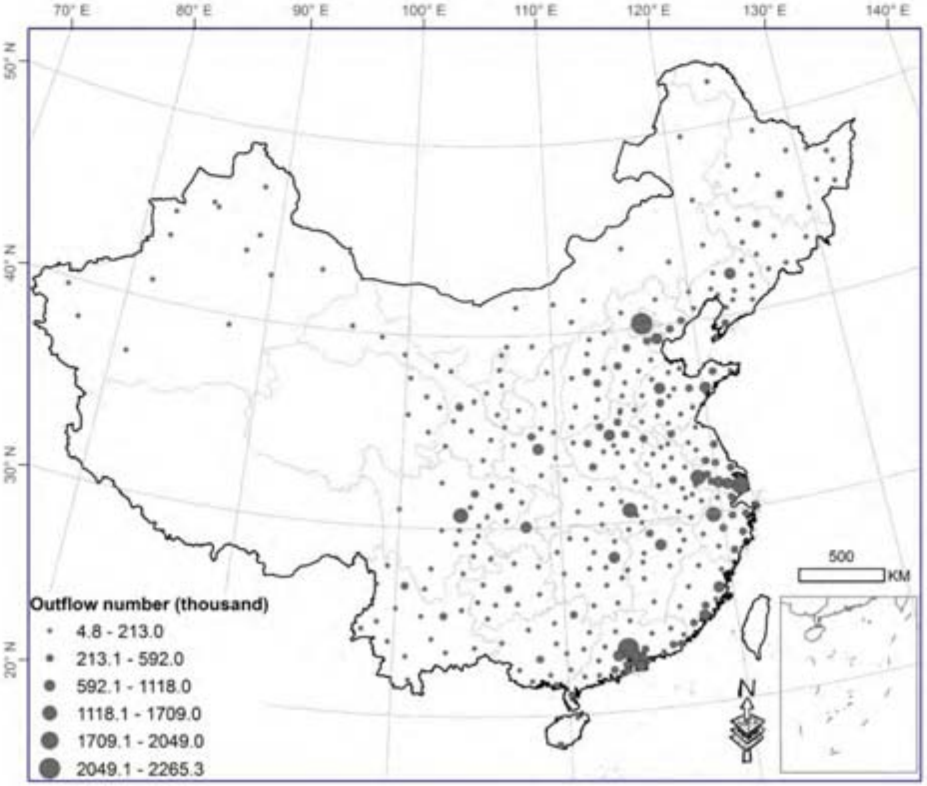
□城市人口流动指标

- 流入量：外地人口进入本城市发送的地理微博量
- 流出量：本地人口在其他城市发送的地理微博量
- 总流动量： sum = 流入量 + 流出量
- 本地人本地量：本地人口在本地发送的地理微博量
- 流入流出比：ratio = 流入量 / 流出量

Urban interaction index

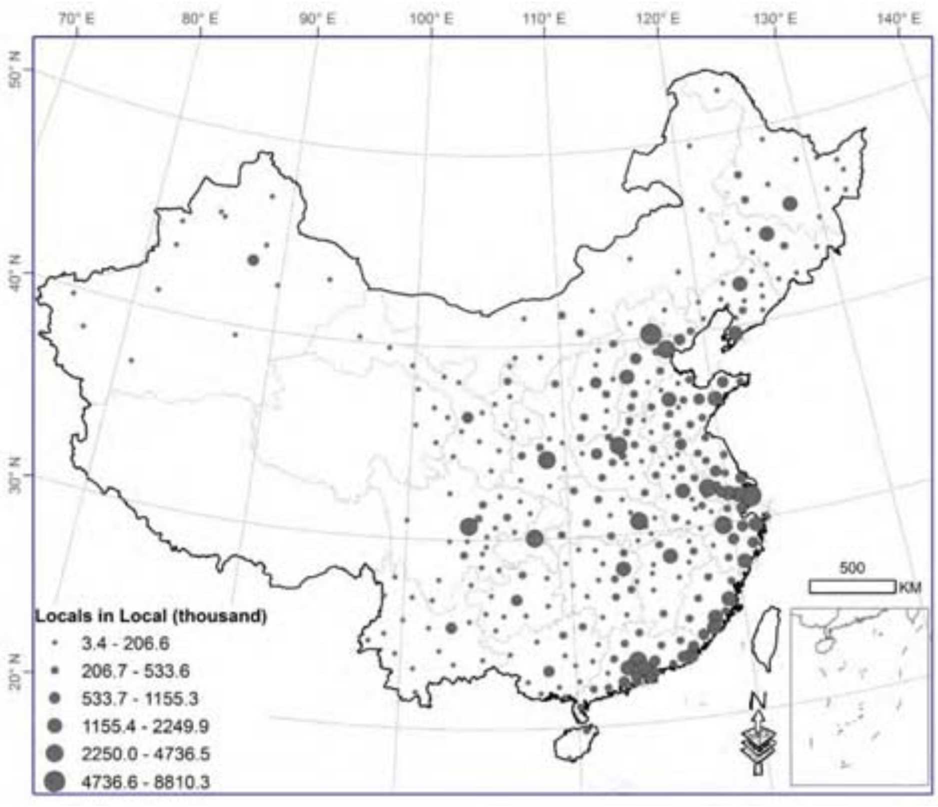


Inflow

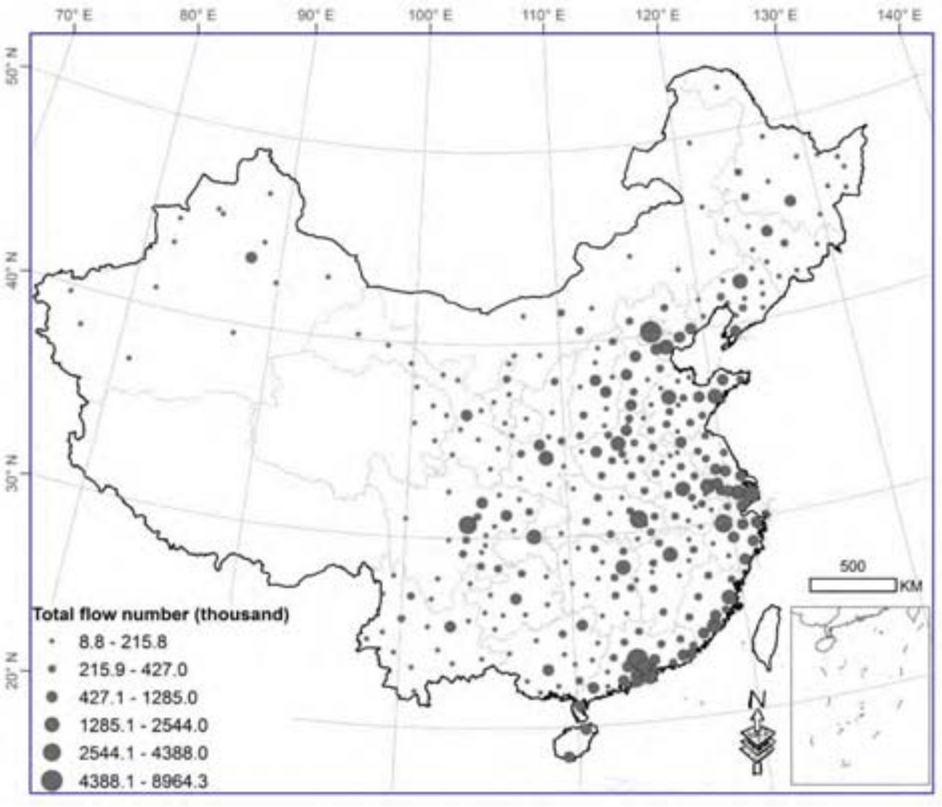


Outflow

Urban interaction index



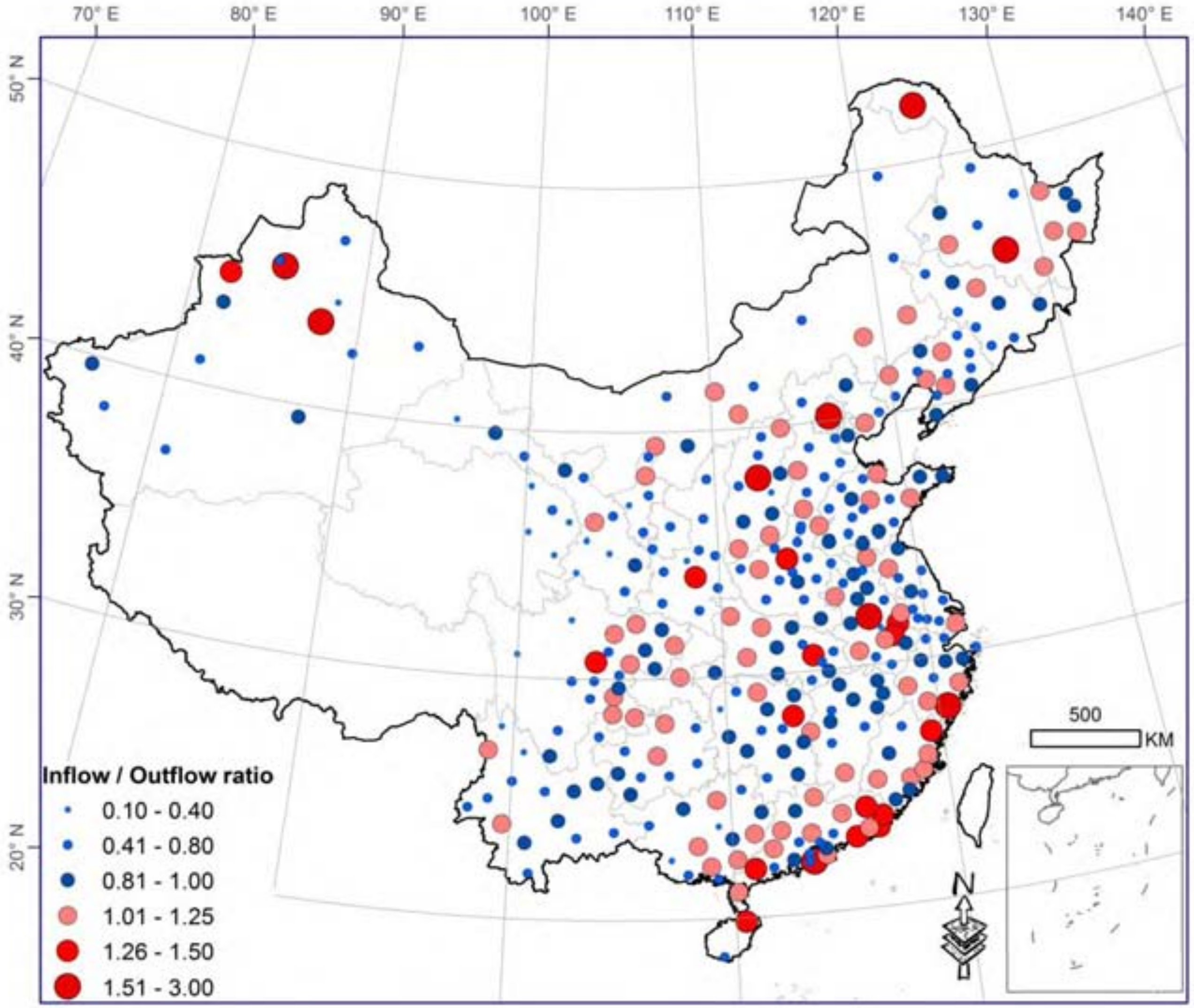
Local in Local



Total flow

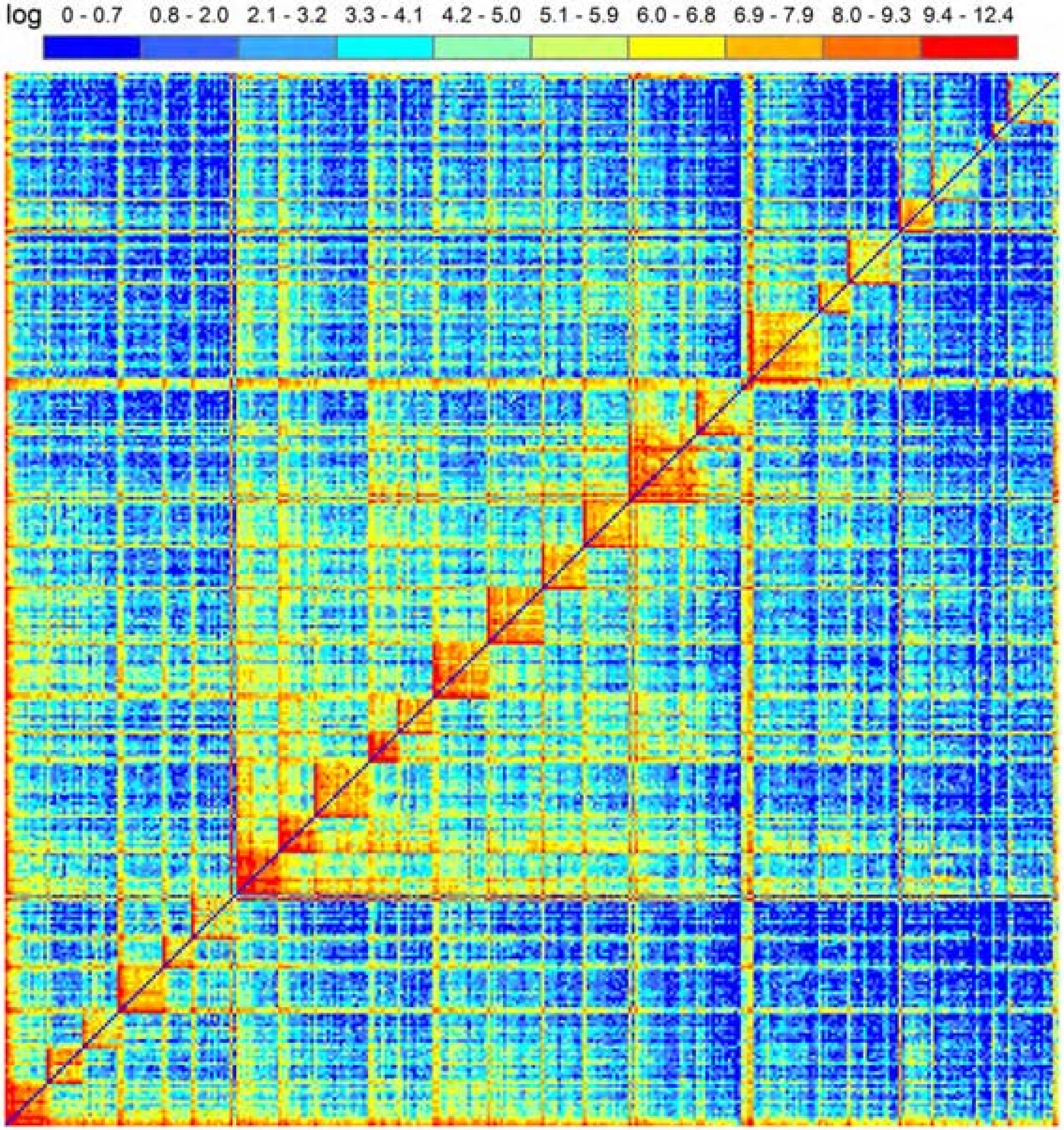
Urban interaction index

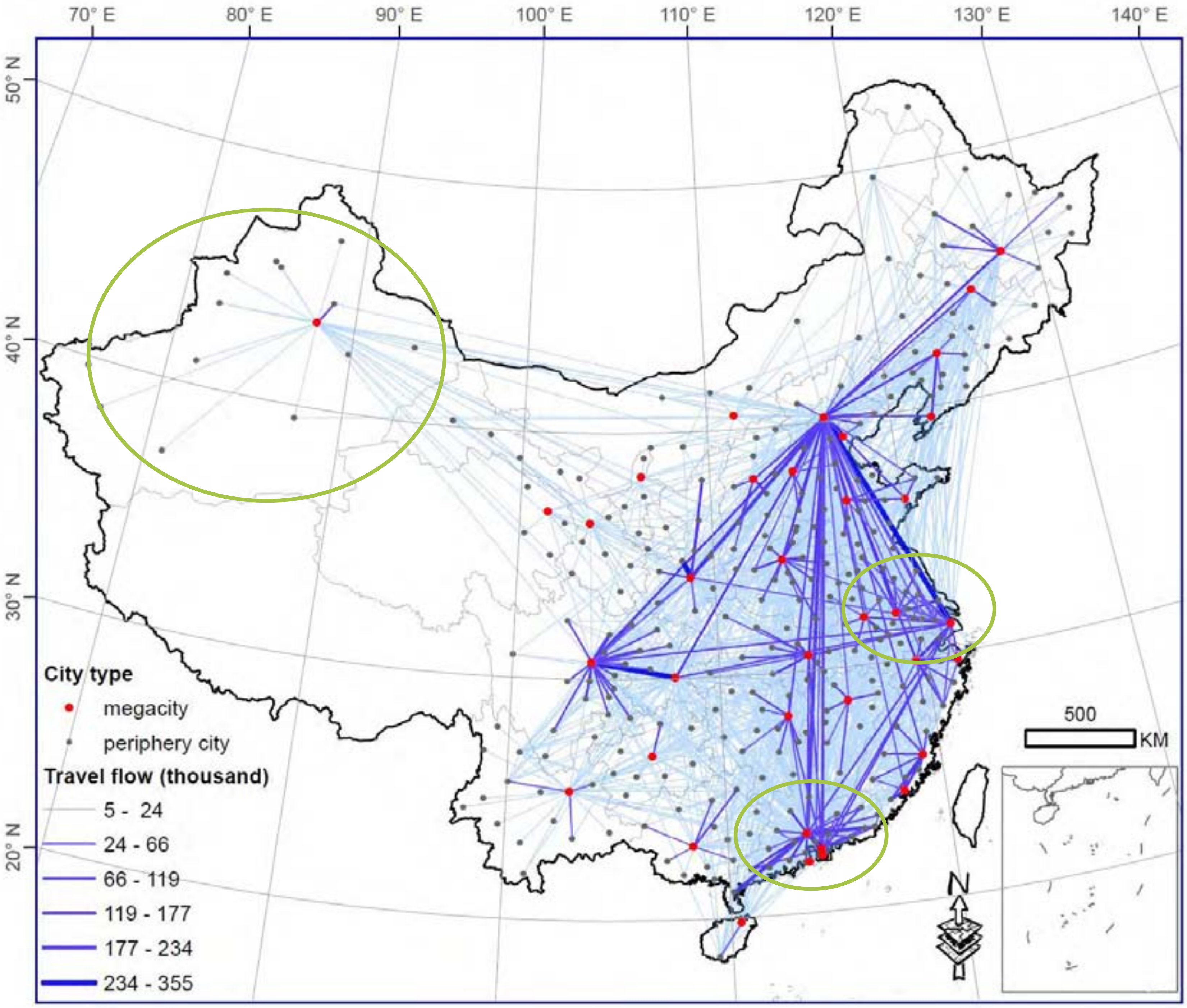
Ratio = Inflow / Outflow



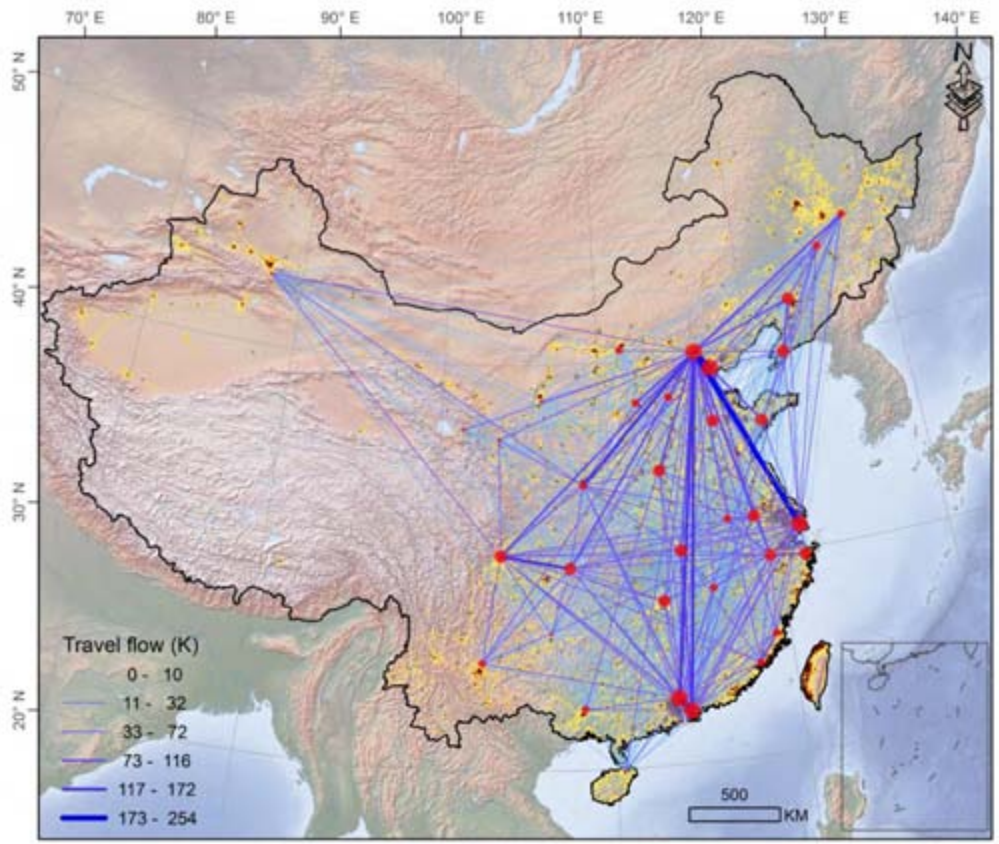
流入量、流出量、总的流动量与人口普查中省级流动人口指标相关系数 **84%**

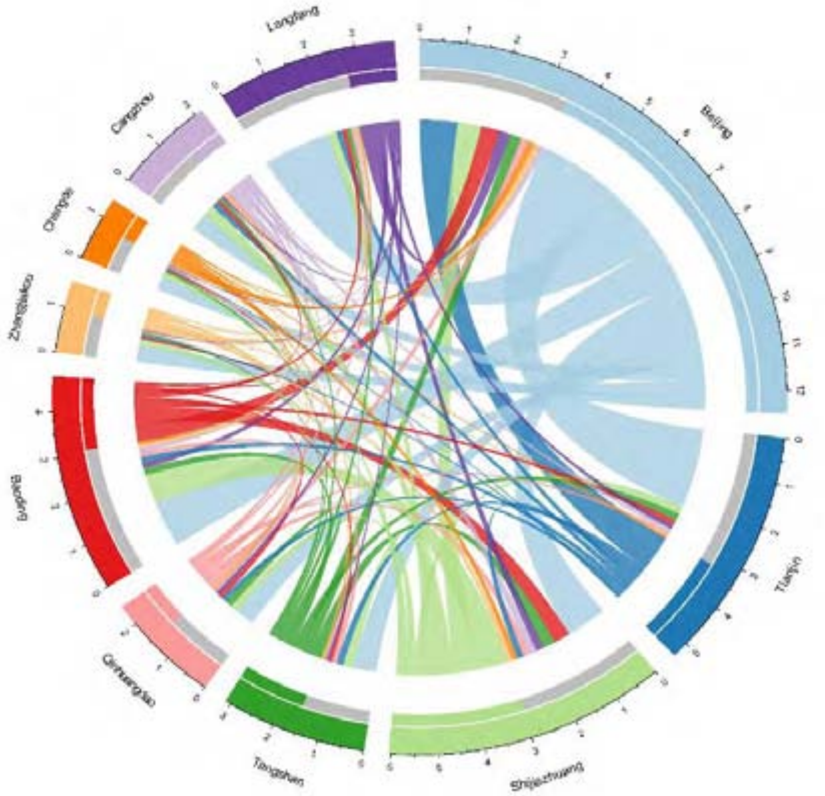
Urban interaction matrix



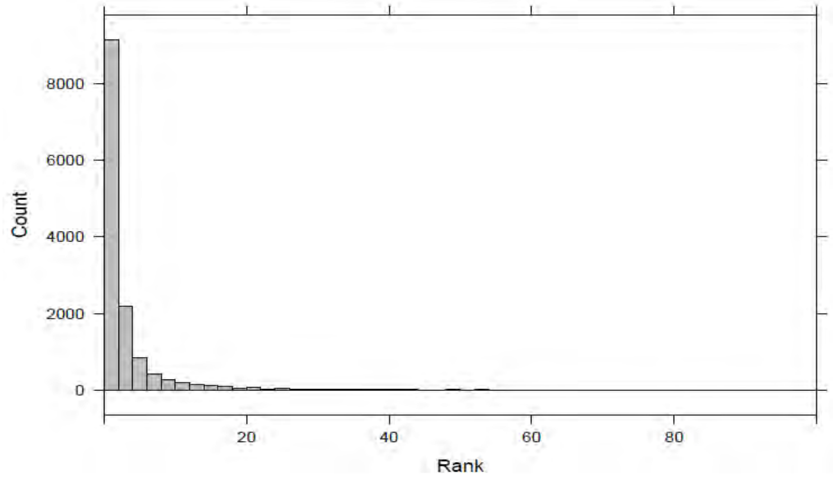

70° E
80° E
90° E
100° E
110° E
120° E
130° E
140° E
50° N
40° N
30° N
20° N
City type
● megacity
● periphery city
Travel flow (thousand)
5 - 24
24 - 66
66 - 119
119 - 177
177 - 234
234 - 355
500
KM
N

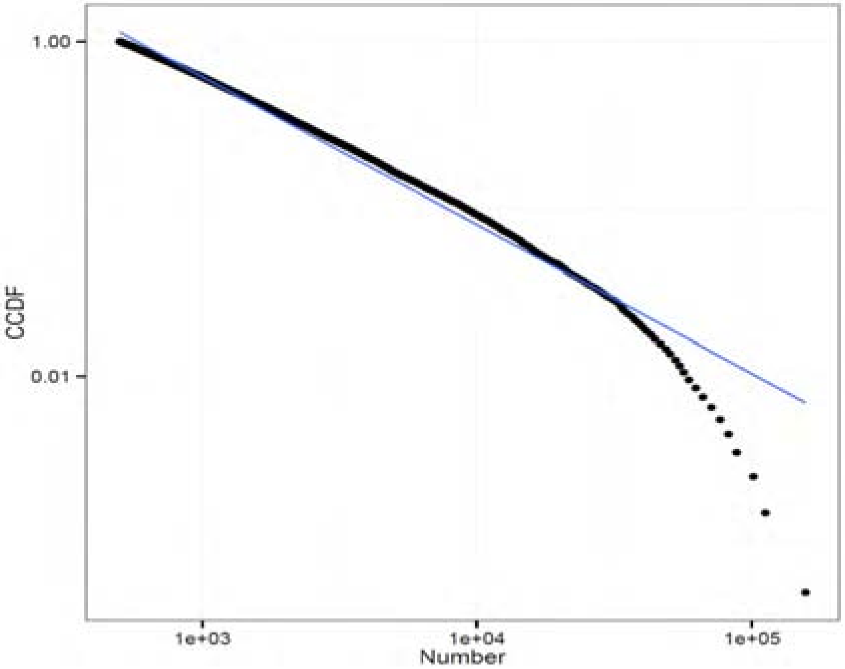
Urban interaction





Power law distribution





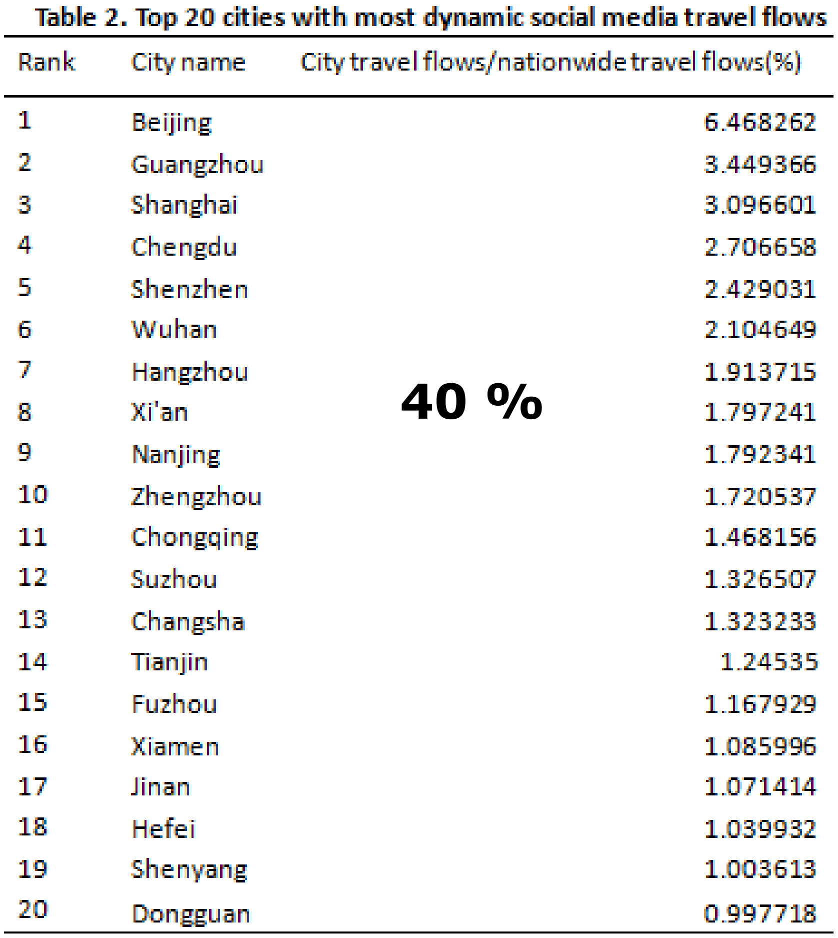


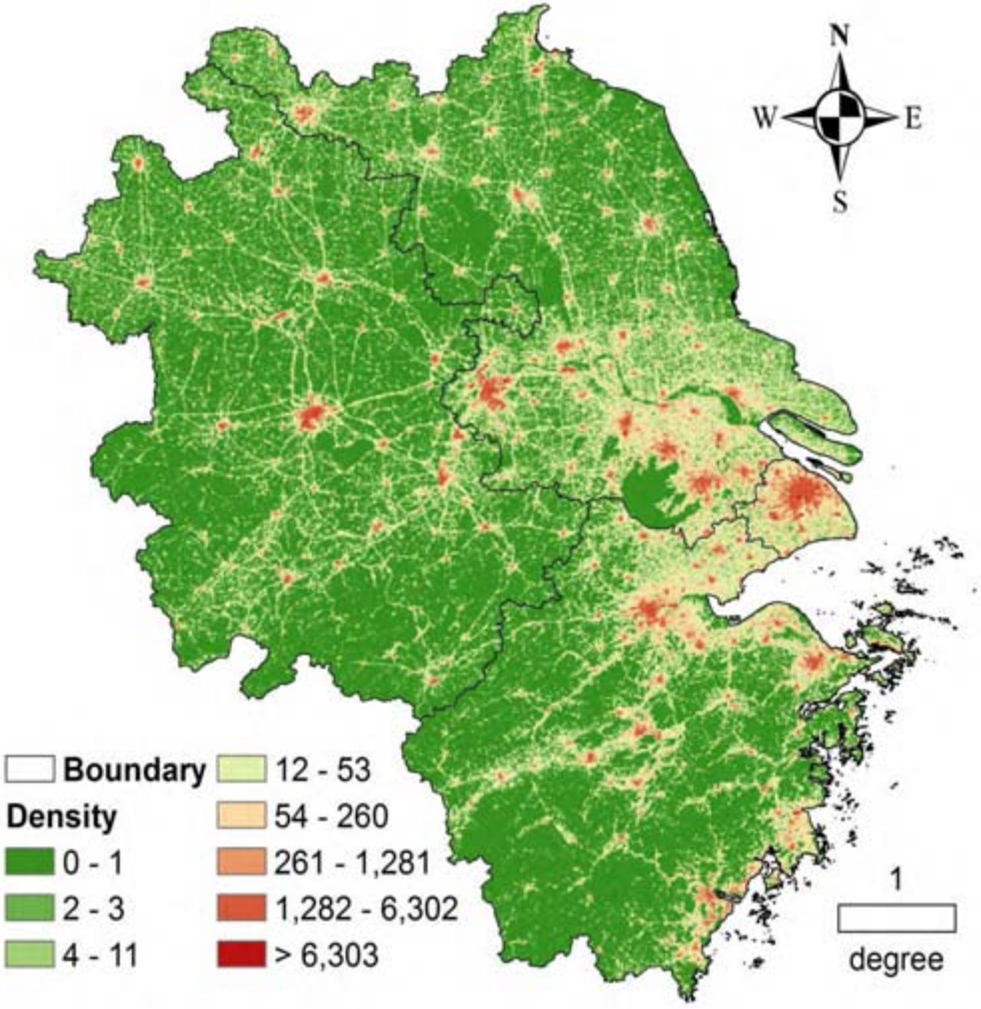
Table 2. Top 20 cities with most dynamic social media travel flows

Rank	City name	City travel flows/nationwide travel flows(%)
1	Beijing	6.468262
2	Guangzhou	3.449366
3	Shanghai	3.096601
4	Chengdu	2.706658
5	Shenzhen	2.429031
6	Wuhan	2.104649
7	Hangzhou	1.913715
8	Xi'an	1.797241
9	Nanjing	1.792341
10	Zhengzhou	1.720537
11	Chongqing	1.468156
12	Suzhou	1.326507
13	Changsha	1.323233
14	Tianjin	1.24535
15	Fuzhou	1.167929
16	Xiamen	1.085996
17	Jinan	1.071414
18	Hefei	1.039932
19	Shenyang	1.003613
20	Dongguan	0.997718

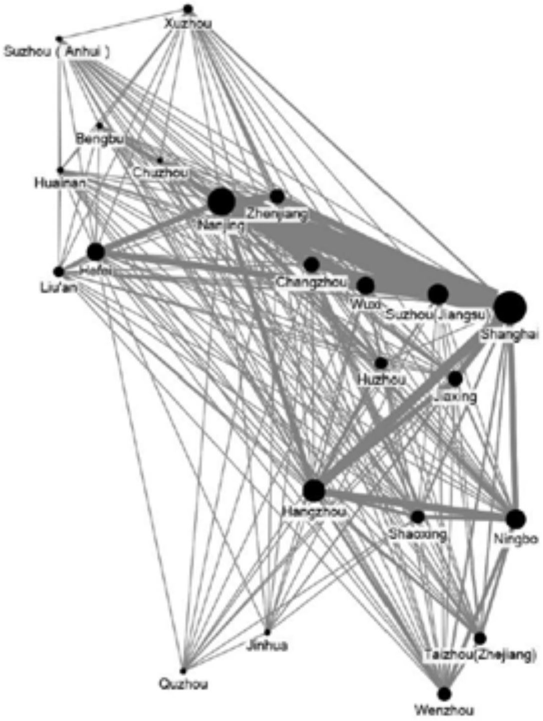
40 %

right-skewed/heavy-tailed

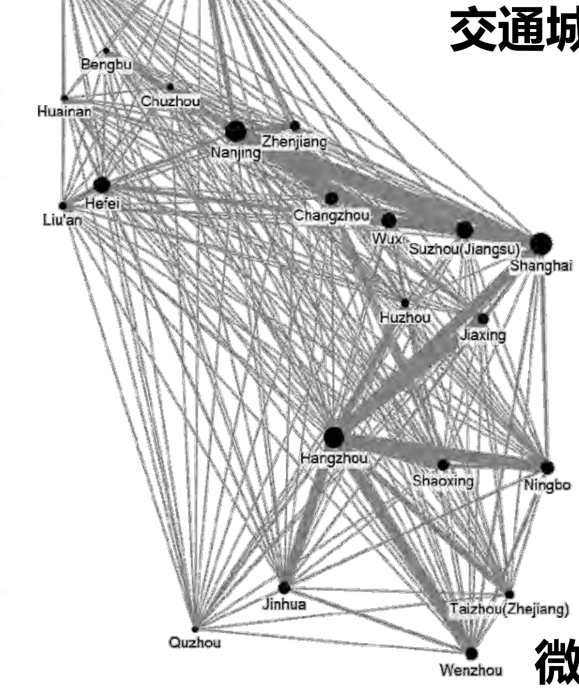
长三角城际联系度量



长三角地理微博密度图



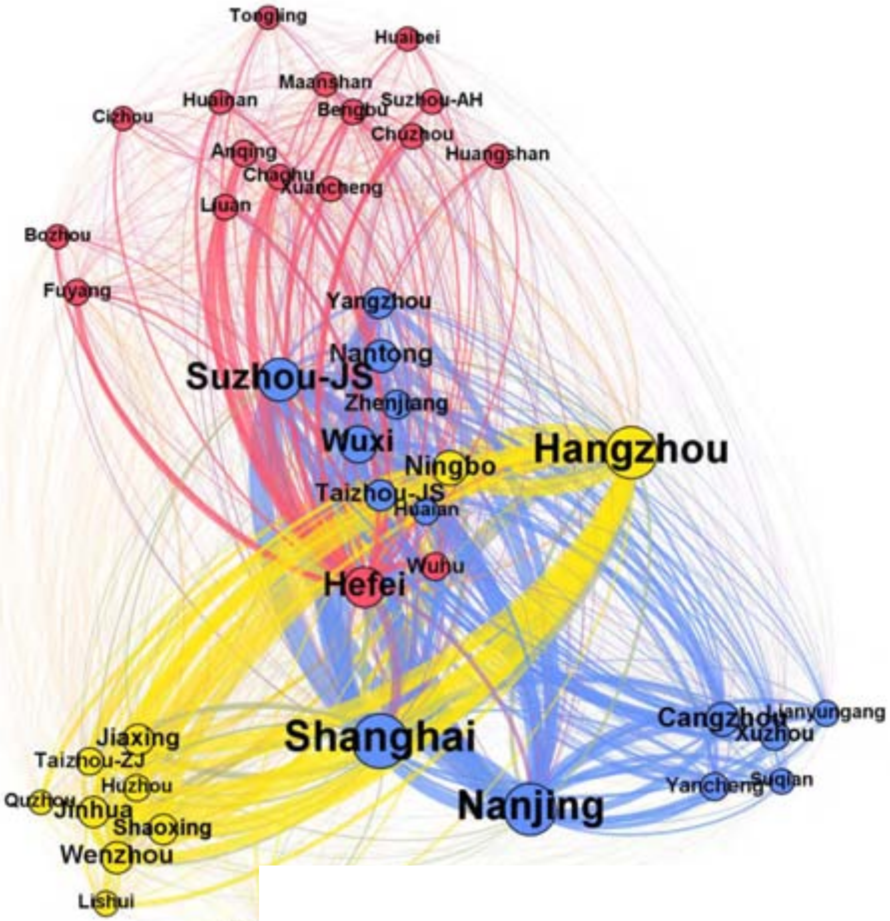
交通城际联系

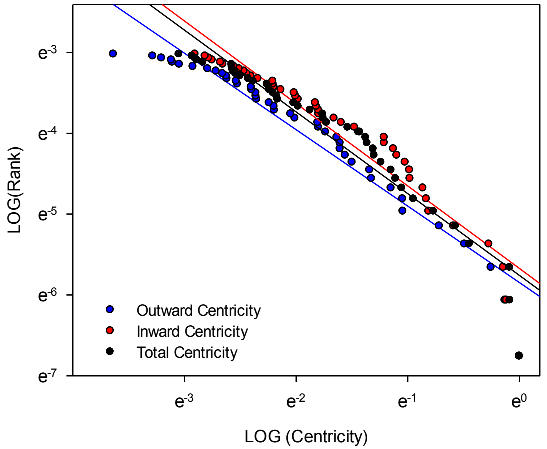


微博城际联系

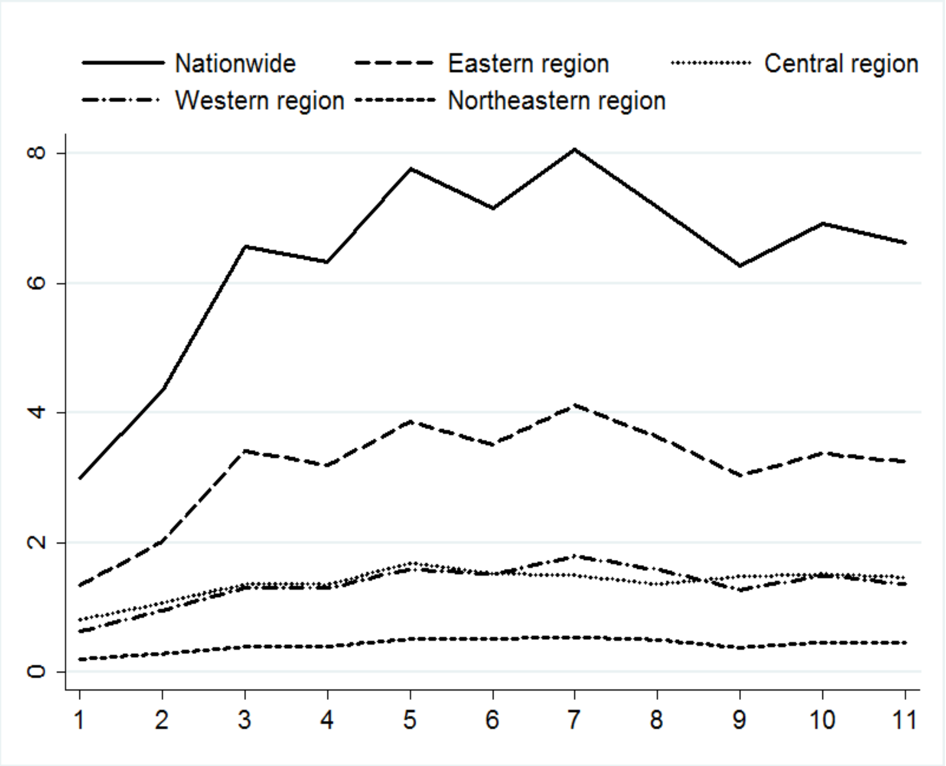
Table 1: The top 20 city-dyads for intercity gross links, monodirectional links and net links

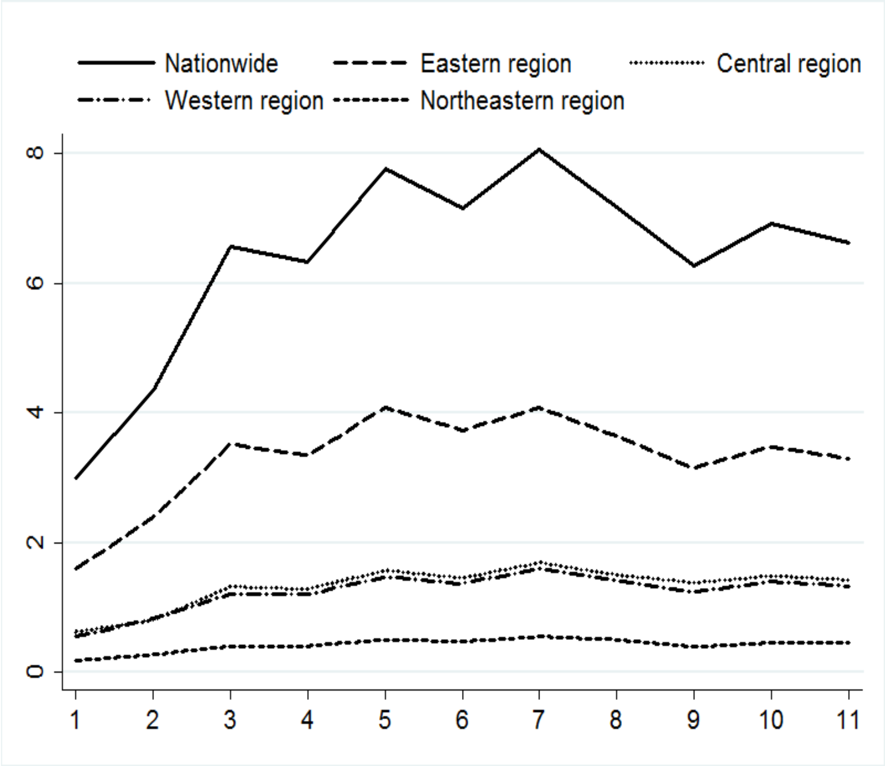
Rank	Gross link	Monodirectional link	Net link
1	Shanghai—Suzhou (J)	Shanghai→Suzhou (J)	Shanghai→Wuxi
2	Ningbo—Hangzhou	Shanghai→Wuxi	Shanghai→Taizhou (J)
3	Nanjing—Suzhou (J)	Wenzhou→Hangzhou	Shanghai→Suzhou (J)
4	Shanghai—Hangzhou	Hangzhou→Jinhua	Hefei→Chaohu
5	Hangzhou—Shaoxing	Shanghai→Hangzhou	Wenzhou→Hangzhou
6	Wenzhou—Hangzhou	Ningbo→Hangzhou	Hangzhou→Jinhua
7	Hangzhou—Jiaxing	Hangzhou→Shaoxing	Nanjing→Zhenjiang
8	Hangzhou—Jinhua	Hangzhou→Jiaxing	Shanghai→Nantong
9	Shanghai—Wuxi	Hangzhou→Ningbo	Wenzhou→Jinhua
10	Nanjing—Nantong	Nanjing→Suzhou (J)	Shanghai→Zhenjiang
11	Nanjing—Wuxi	Suzhou (J)→Nanjing	Shanghai→Cangzhou
12	Nanjing—Shanghai	Nanjing→Nantong	Nanjing→Taizhou (J)
13	Shanghai—Nantong	Nanjing→Wuxi	Hefei→Liu'an
14	Nanjing—Cangzhou	Shanghai→Nantong	Shanghai→Hangzhou
15	Wuxi—Suzhou (J)	Shanghai→Taizhou (J)	Hefei→Shanghai
16	Nanjing—Yangzhou	Hangzhou→Shanghai	Hangzhou→Shaoxing
17	Hangzhou—Huzhou	Shaoxing→Hangzhou	Taizhou (J)→Zhenjiang
18	Taizhou—Hangzhou	Jiaxing→Hangzhou	Wuhu→Chaohu
19	Nanjing—Zhenjiang	Nanjing→Shanghai	Nanjing→Huaian
20	Nanjing—Xuzhou	Wuxi→Nanjing	Shanghai→Yangzhou





时间序列分析





Holiday、weather

方言地理分布

□ 度量城际之间方言的相似度

□ Micro linguistic data from Atlas of Chinese Dialects (ACD)

- Documented all direct and indirect linguistic characteristics
- Collected from Institute of Linguistic, Chinese Academy of Science
- Contemporary and historical dialect data for 2010s, 1980s and 1960s

□ Defining dialect distance (non-similarity) between city pairs:

$$LD_{AB} = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(s_{Ai} \times s_{Bj} \times \delta_{ij} \right)$$

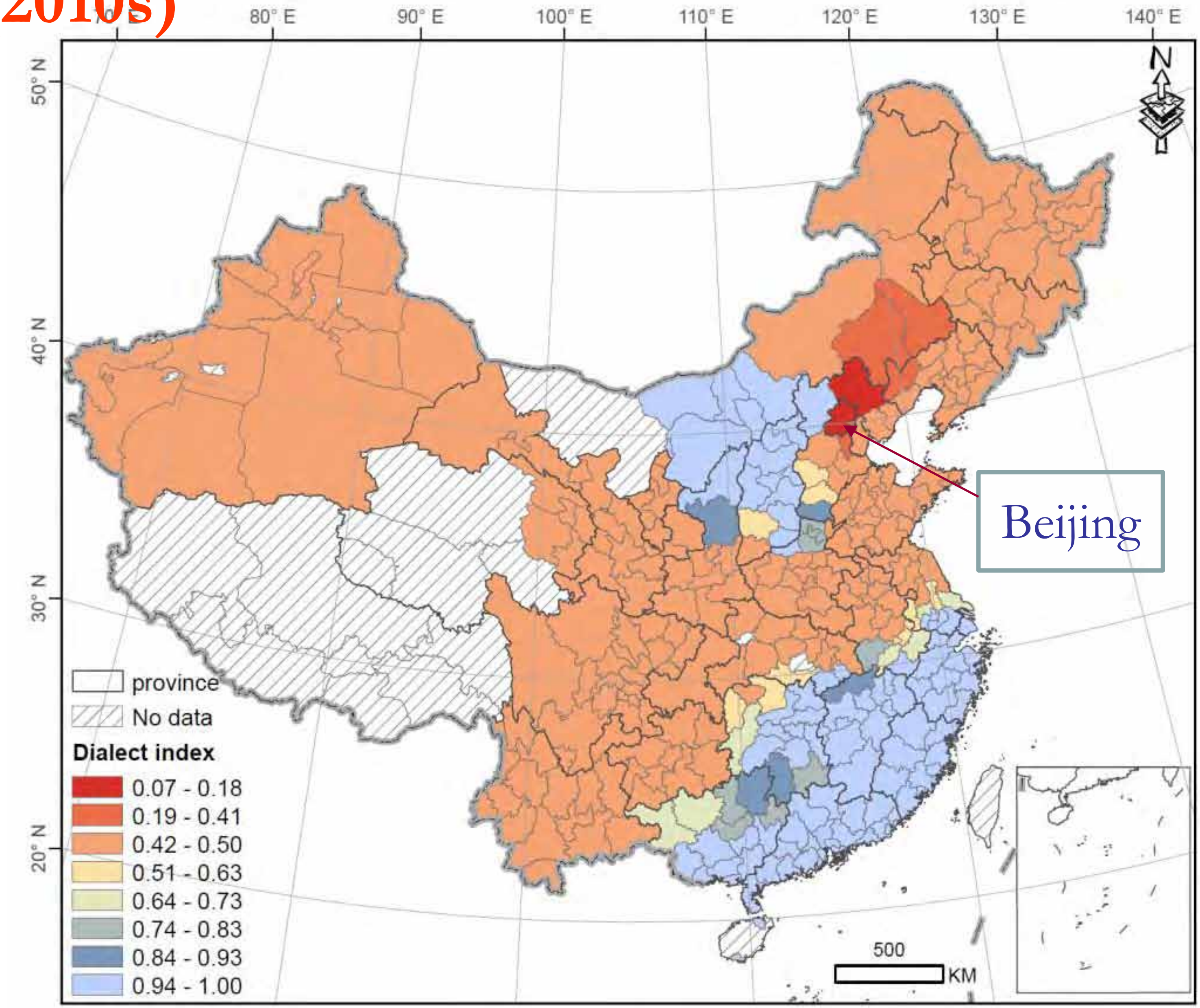
I indicates the linguistic of city A; J indicates the linguistic of city B;

s_{Ai} is the proportion of population in city A who speak the linguistic I;

s_{Bj} is the proportion of population in city B who speak the linguistic J;

δ_{ij} is the linguistic non-similarity between linguistic I and linguistic J.

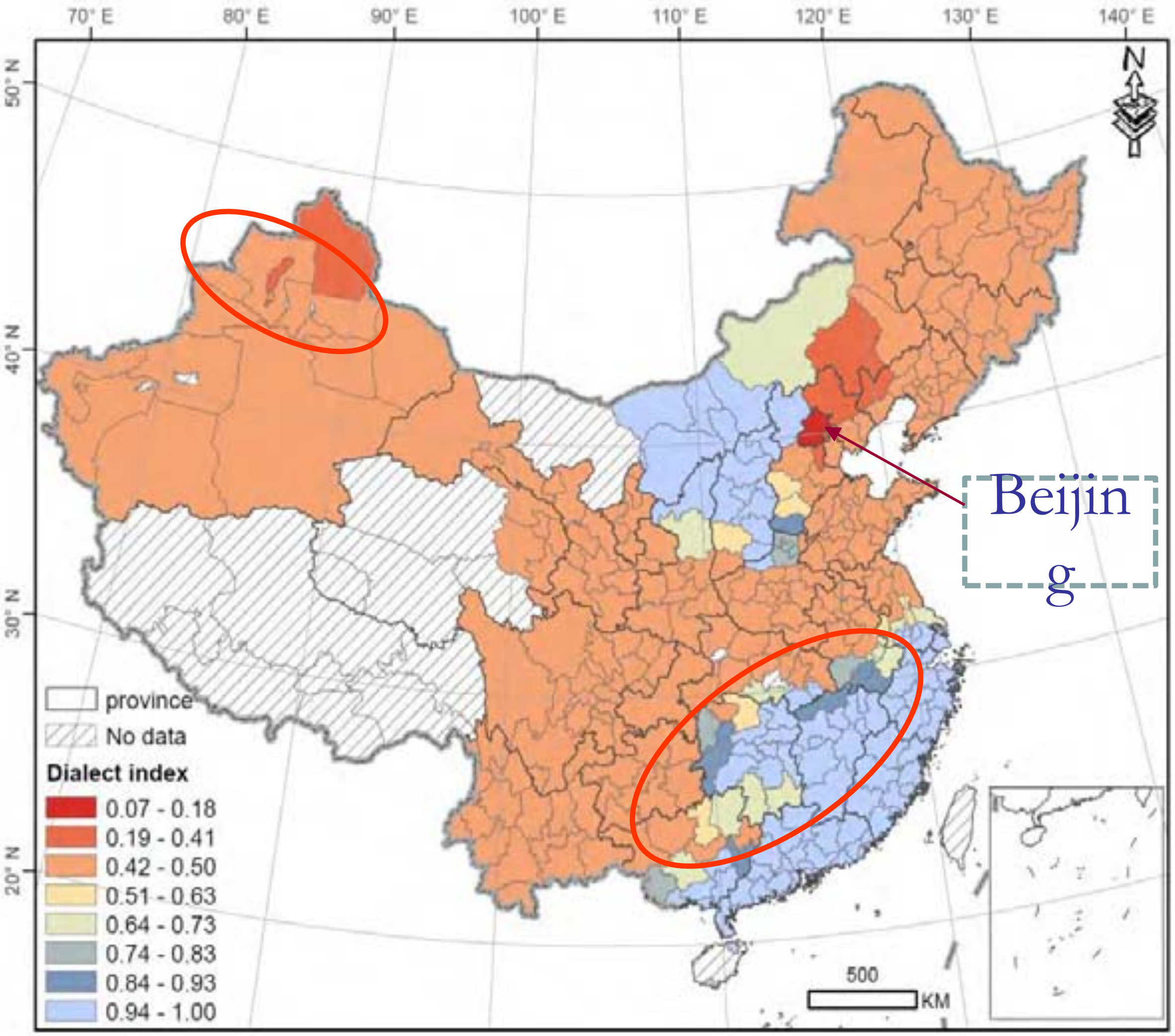
Contemporary dialect distance from other cities to Beijing (2010s)



Beijing

Historical dialect distance from other cities to Beijing (1960-1980s)

Correlation coefficient: 0.7



Beijin g

Estimation strategy

- Cross-sectional OLS regression (in logs), city pairs

$$\log(T_{od} / L_o) = \beta_1 \cdot \log[Commuting_{od}] + \beta_2 \cdot \log[Dialect_{od}] + F_o + F_d + controls + \varepsilon_{od}$$

T_{od} / L_o: mobility flows between city pairs/total Weibo user flows of the origin city

- [Commuting]: commuting distance&time-(pecuniary mobility costs)
- [Dialect]: dialect distance- (non-pecuniary mobility costs)
- Fr: origin city fixed effect
- Fs: destination city fixed effect
- IV: historical dialect distance

Additional Robustness checks:

Heterogeneity effects by imputed travel motivations

- **Family reunion:** Spring Festival Season (Chun Jie) sample
- **Tourism:** National public holidays sample (Qingming, Duanwu, Labor Day, Zhongqiu, National Day)
- **Business v.s. Leisure:**
- *Weekdays trip* sample v.s. *Weekends trip* sample
- high-frequency visited cities per months (visited more than once per month)

Conclusions

- **Key findings:**
  - The rise of 1 percent in a city-pair's contemporary dialect distance would increase human mobility flows by 4.8 percent
  - Effects are not distributed evenly over time, and between metropolitan regions and periphery regions
- **Big data:** we economists could ride the wave of social media data availability and develop people/place-based policy analysis
- **Future works:** Impacts from changes in market potentials (induced by High-Speed Rails) on changes in mobility flows

应用2 华人足迹与人口估算研究

Jianghao Wang

IGSNRR, CAS, CN

Xingjian Liu

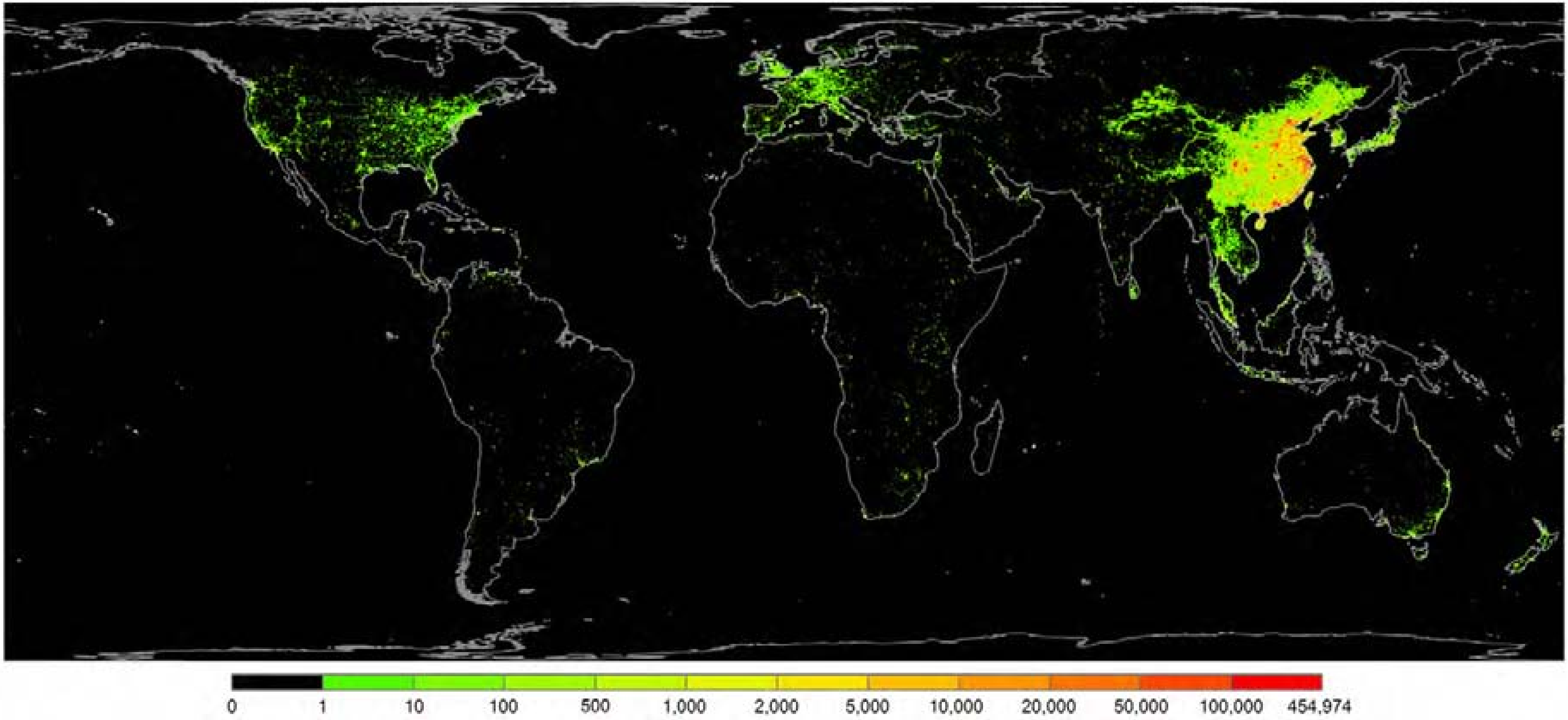
Hong Kong University

The geographic of Weibo: Where are the Chinese? EPA.

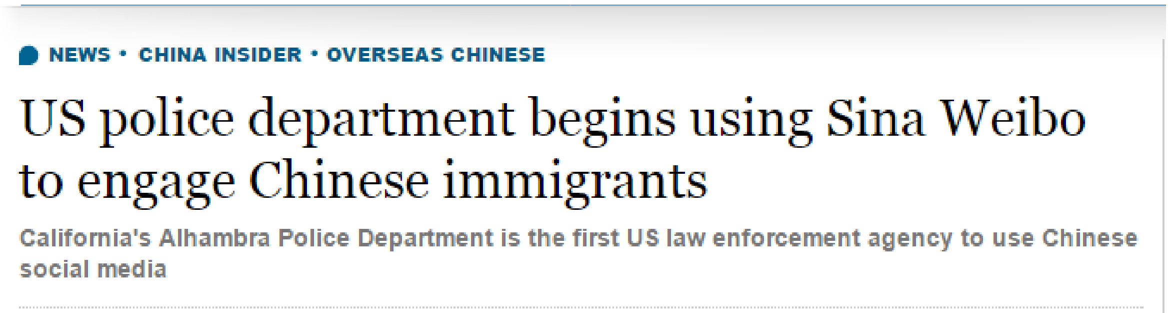
Where are the Chinese?

Bias is endogenous, but sometimes bias can be useful

2014年上半年世界范围内所有地理微博热点分布， 共计**1.49**亿



A new way of small-area population estimation



NEWS • CHINA INSIDER • OVERSEAS CHINESE

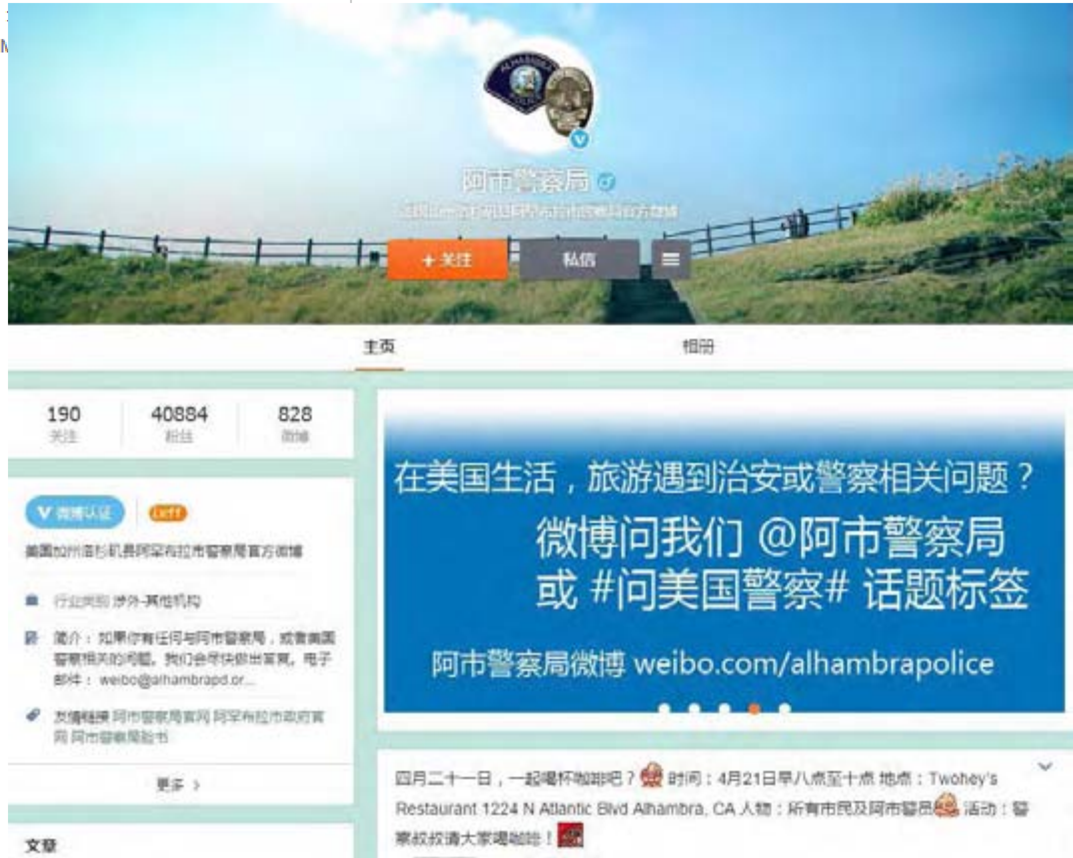
US police department begins using Sina Weibo to engage Chinese immigrants

California's Alhambra Police Department is the first US law enforcement agency to use Chinese social media

Jeremy Blum
jeremy.blum@scmp.com

PUBLISHED :
UPDATED : M

As the basis for public service provision?



美国加州洛杉矶县阿罕布拉市警察局官方微博

行业类别 涉外-其他机构

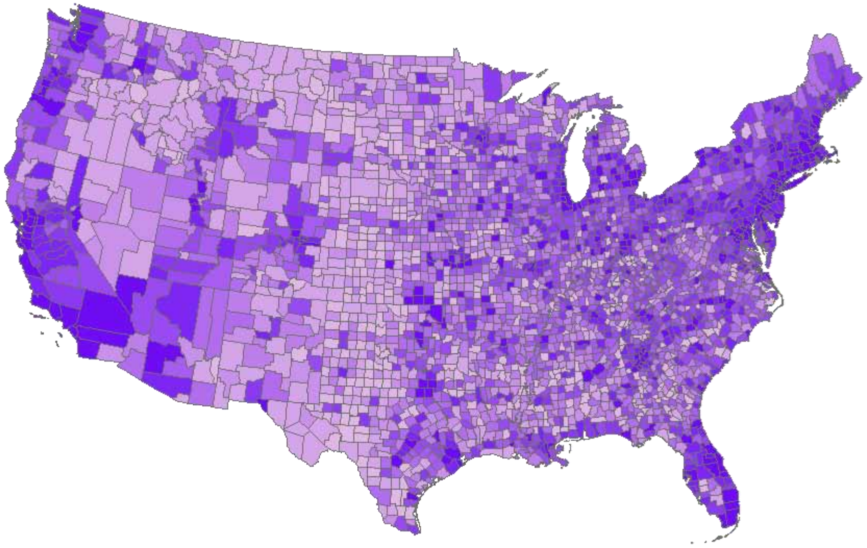
简介：如果你有任何与阿市警察局，或者美国警察相关的问题，我们会尽快做出答复。电子邮件：weibo@alhambrapd.or...

友情链接 阿市警察局官网 阿罕布拉市政府官网 阿市警察局脸书

在美国生活，旅游遇到治安或警察相关问题？
微博问我们 @阿市警察局
或 #问美国警察# 话题标签
阿市警察局微博 weibo.com/alhambrapolice

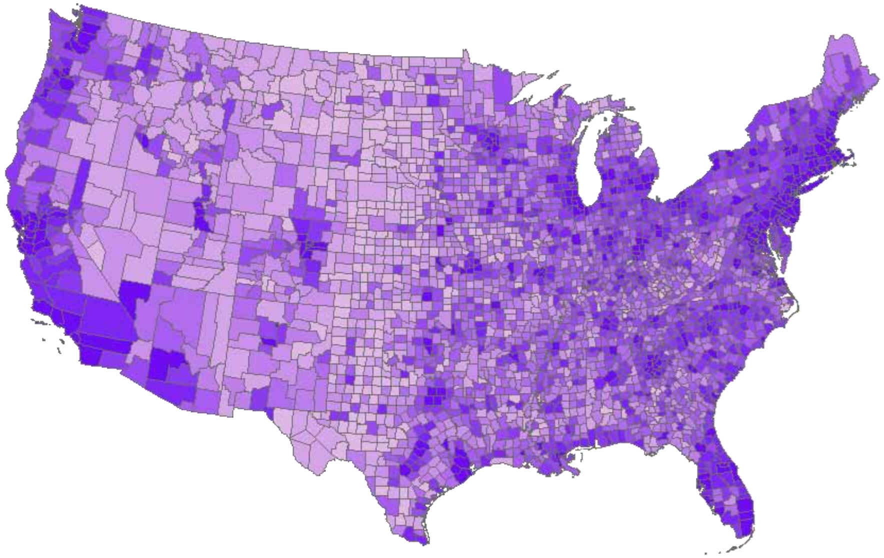
四月二十一日，一起喝杯咖啡吧？ 时间：4月21日早八点至十点 地点：Twohey's Restaurant 1224 N Atlantic Blvd Alhambra, CA 人物：所有市民及阿市警员 活动：警察叔叔请大家喝咖啡！

Census vs. geotagged Weibo estimation



Chinese American

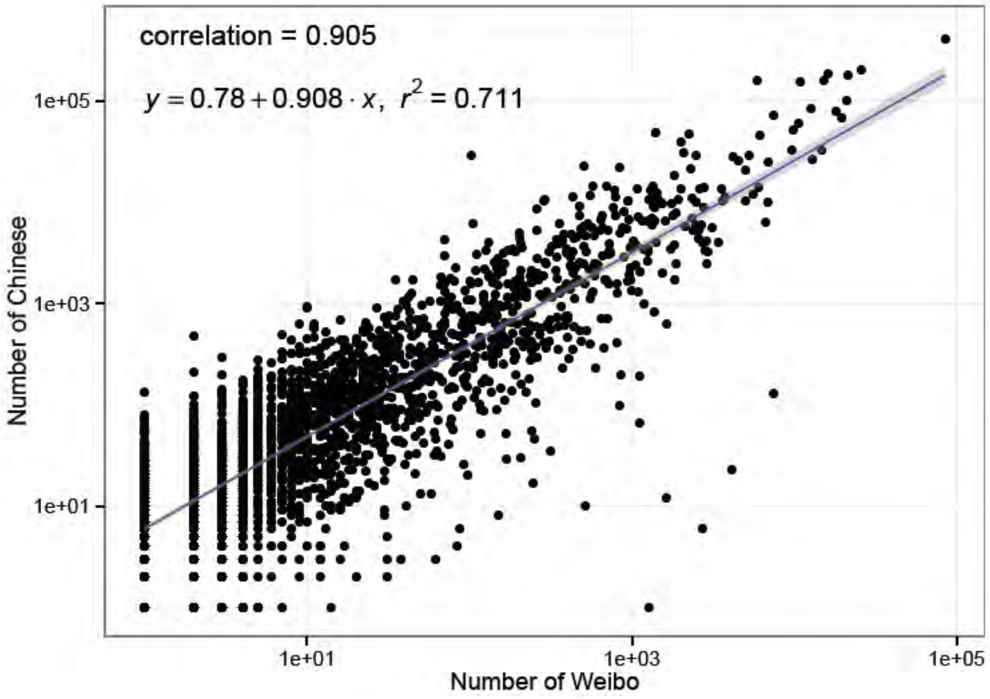
(Pew Research, based on 2010 Census)

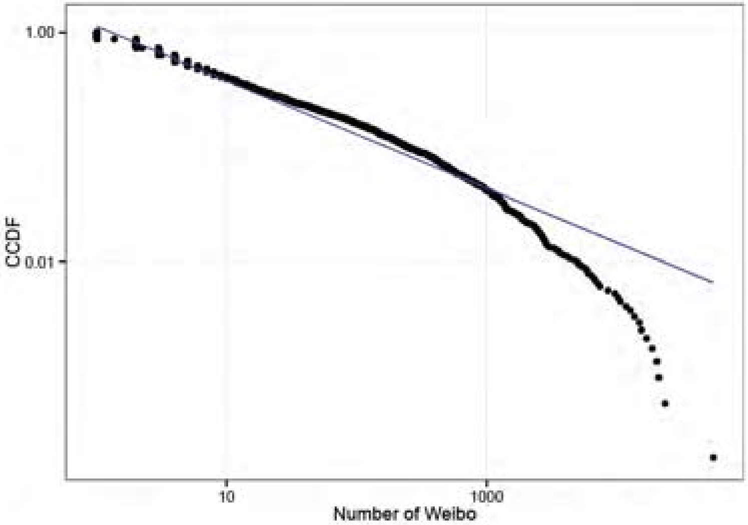


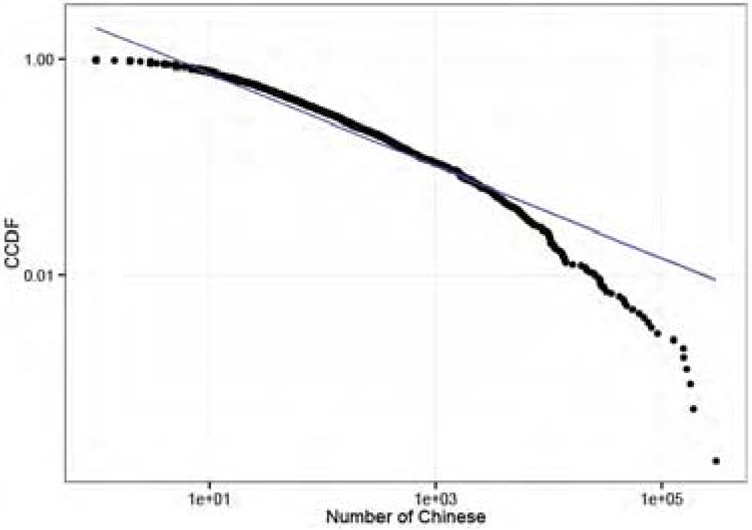
Chinese American

(Estimated based on geotagged Weibo)

Census vs. Geotagged Weibo estimation







心得和体会

□ 社交媒体空间大数据研究的缺陷与优势

- 大样本代表性问题、去伪存真、纠偏
- 相比于轨迹数据、粒度高、信息量大
- 地理学、城市规划、经济学研究提供新的视角

□ 大数据挖掘的三要素

- 数据数据数据 ＋ 技术技术 ＋ 学科背景
- 鼓励学科交叉和合作

Thanks！

Q & A

王江浩CAS

http://jianghao.wang