北京城市实验室
Beijing City Lab

ID of the slides

25

Slides of BCL

www.beijingcitylab.com

How to cite

Author(s), Year, Title, Slides at Beijing City Lab, http://www.beijingcitylab.com

E.g. Long Y, 2014, Automated identification and characterization of parcels (AICP) with OpenStreetMap and Points of Interest, Slides at Beijing City Lab, http://www.beijingcitylab.com

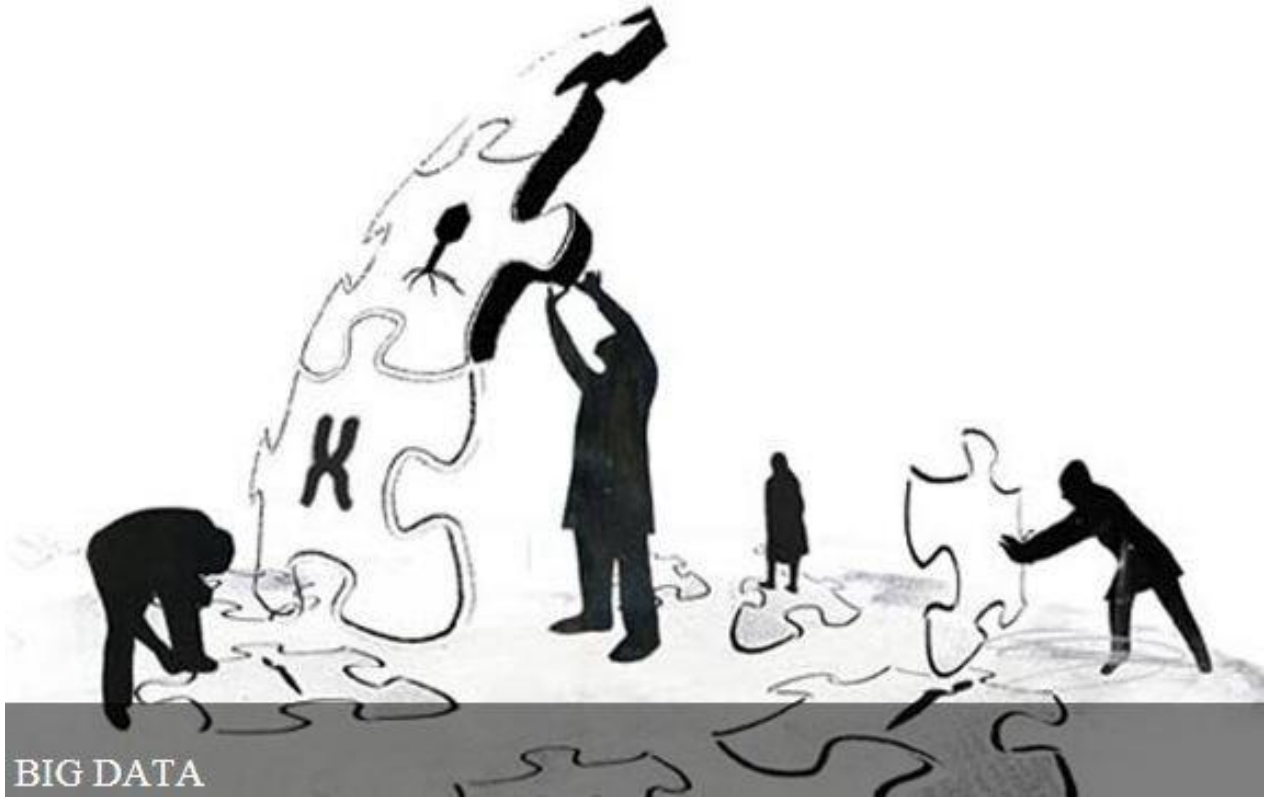
大数据给城市研究带来的机遇和挑战

姚晓白， 佐治亚大学地理系

清华大学建筑学院城市规划系学术交流报告
2014年6月18日

内容提纲

- 大数据的特点
- 大数据给城市研究带来了什么样的研究机会？
- 研究实例
- 大数据的挑战



Science
AAAS

The World's Technological Capacity to Store, Communicate, and Compute Information

Martin Hilbert and Priscila López

Science **332**, 60 (2011);
DOI: 10.1126/science.1200970

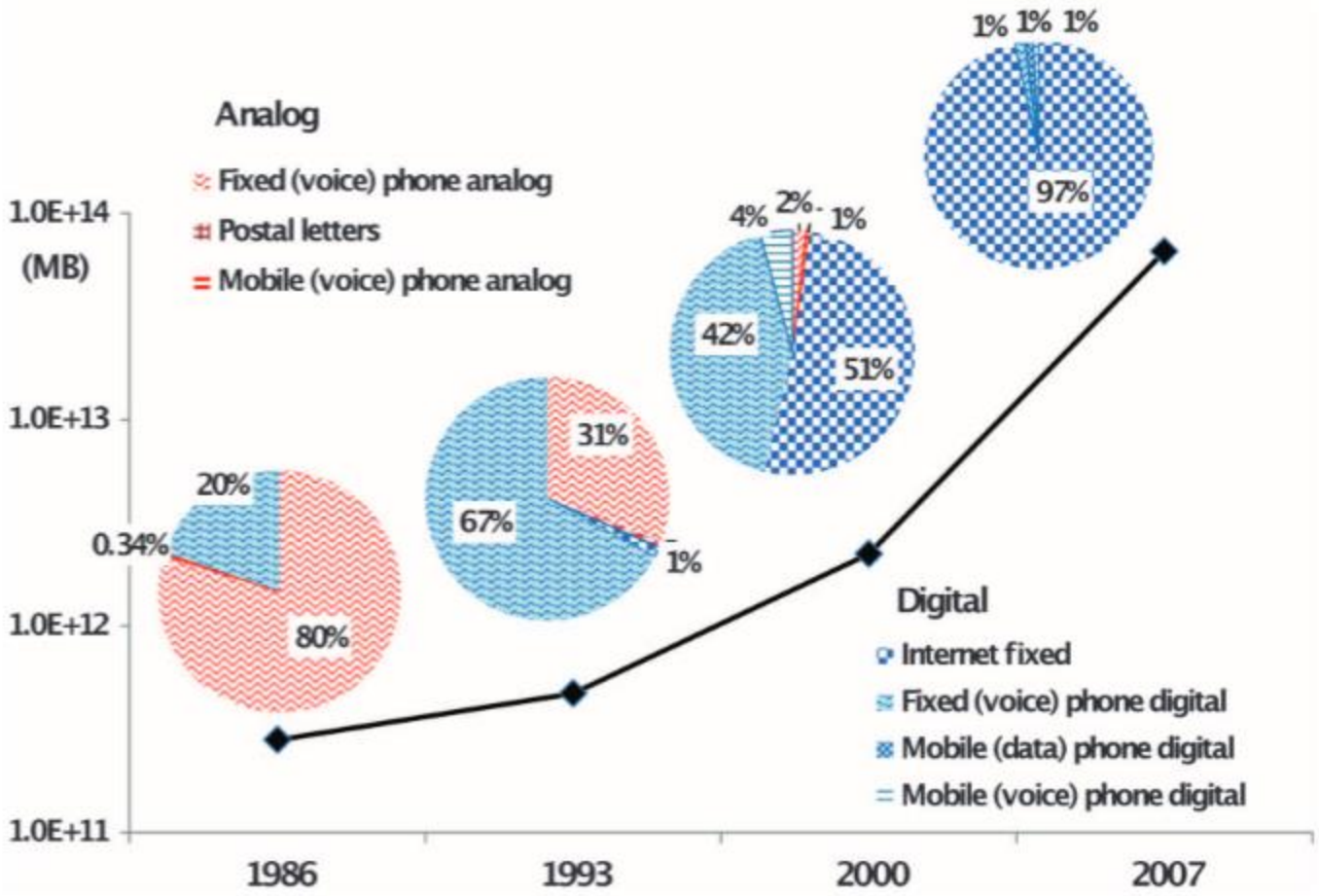
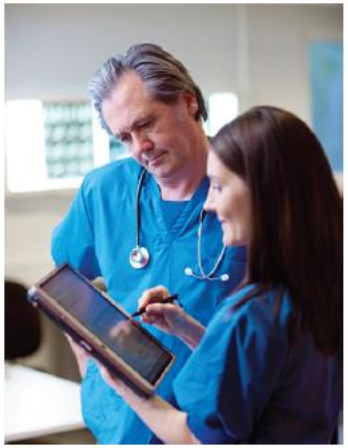


Fig. 4. World's technological effective capacity to telecommunicate information (table SA2) (*16*)

McKinsey Global Institute

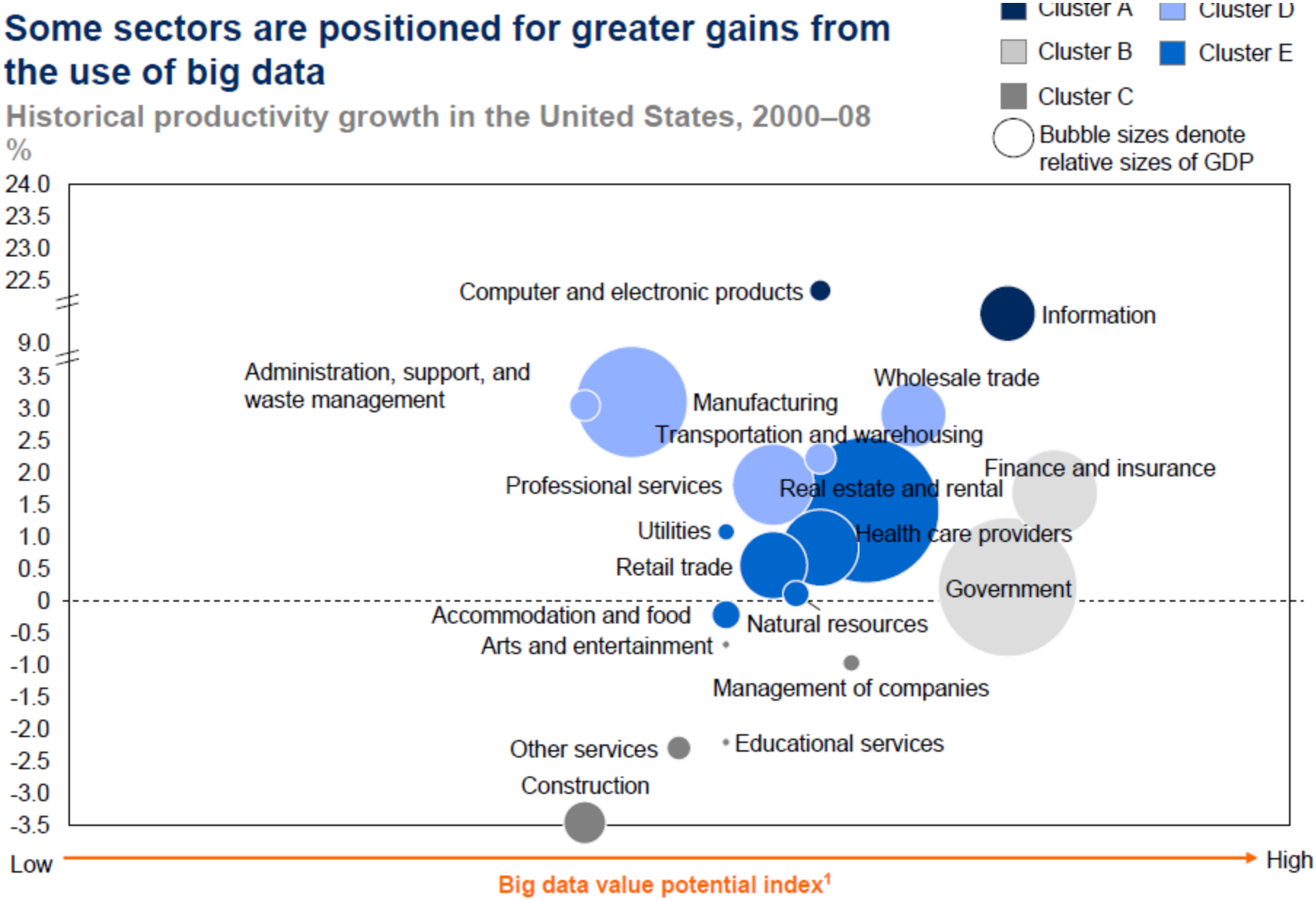



June 2011

Big data: The next frontier for innovation, competition, and productivity

Some sectors are positioned for greater gains from the use of big data

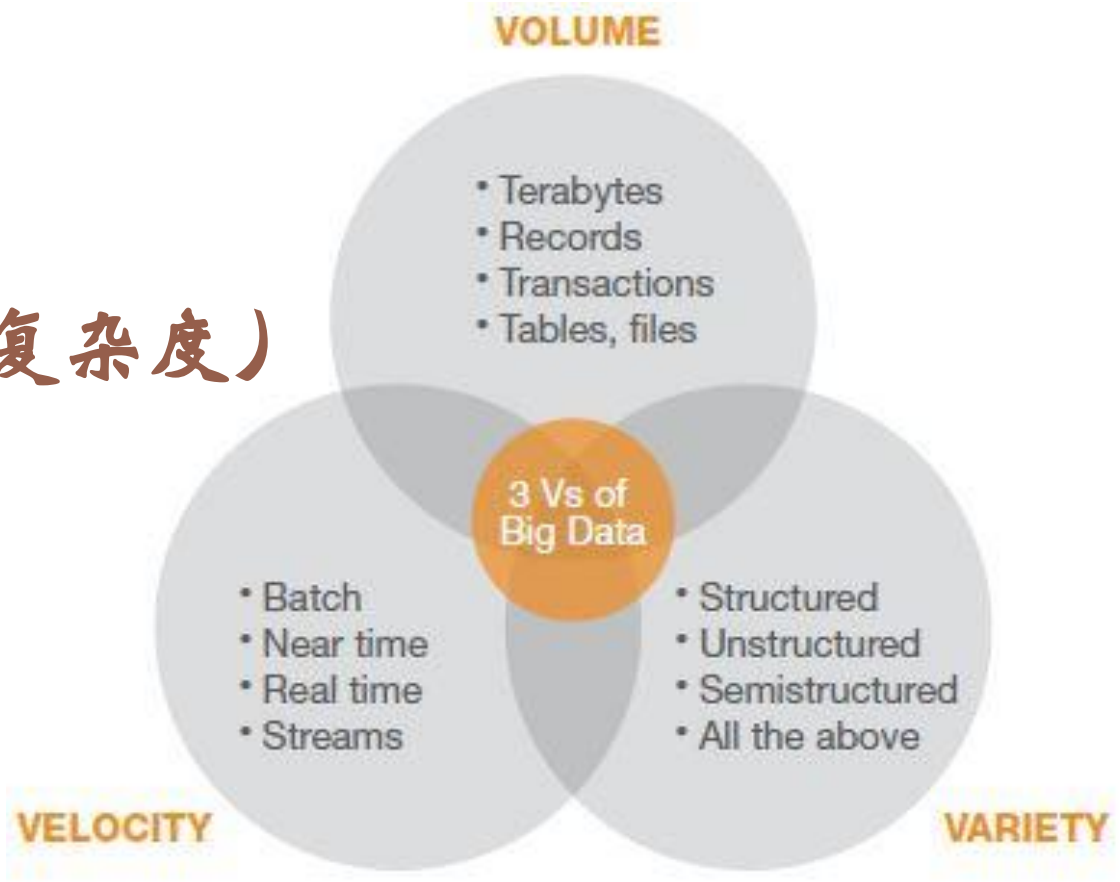
Historical productivity growth in the United States, 2000–08



1 See appendix for detailed definitions and metrics used for value potential index.
SOURCE: US Bureau of Labor Statistics; McKinsey Global Institute analysis

大数据的特点

1. 容量
2. 速度
3. 多样性（复杂度）



来源：Meta Group 2001

城市中常见的大数据

社交网络或通信数据

- Facebook, Twitter
- 微博，微信，QQ
- 淘宝, Amazon, …

物联网数据

- 空气质量检测数据
- 公交卡数据
- …

大数据的应用潜力举例

城市服务

- 智慧城市

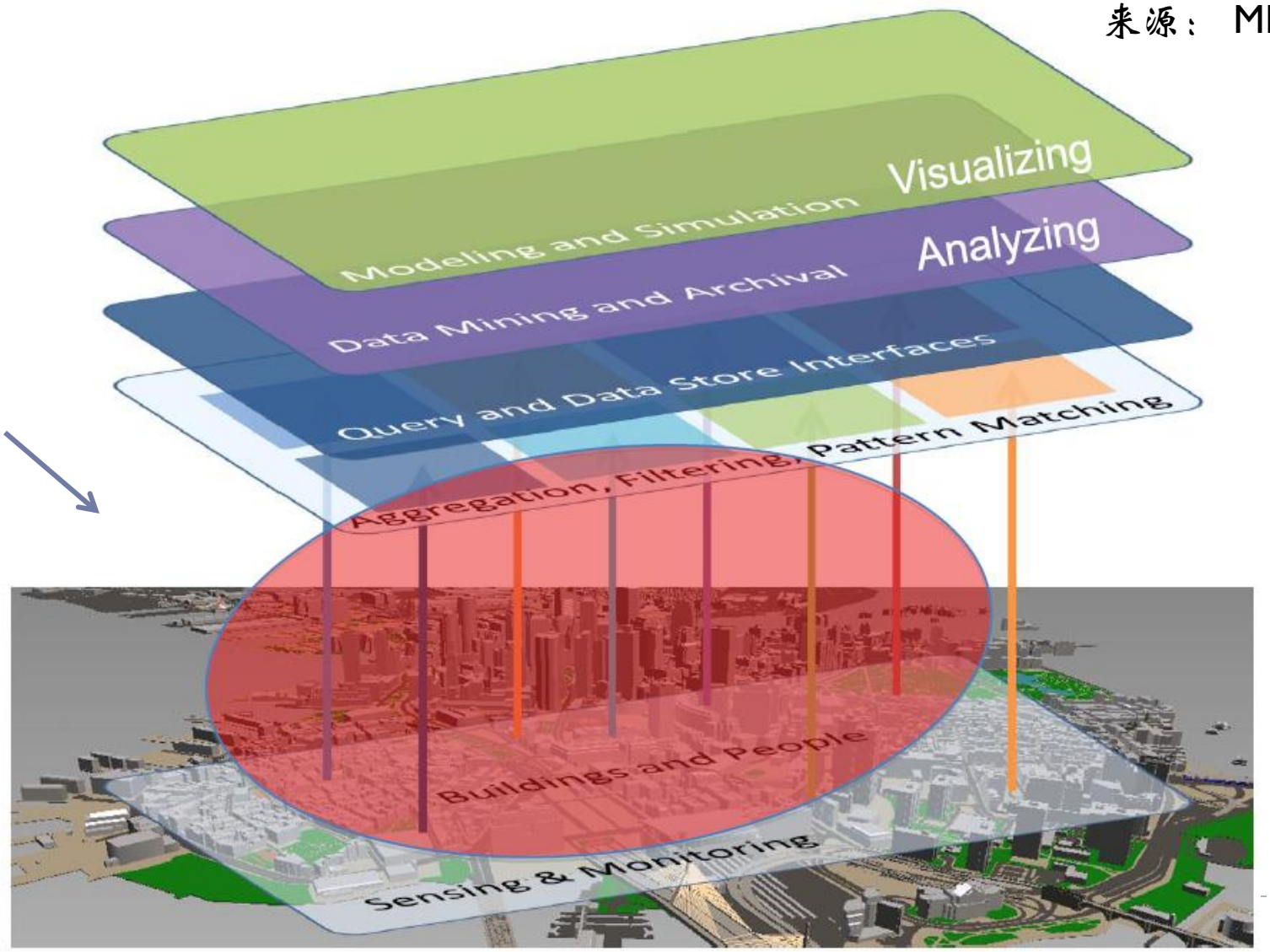
城市研究

- 人的行为
- 城市的行为

例： MIT提出的GIS为核心的大数据平台

来源： MIT

辅以
物联网



London Tweets by Language

大数据给我们带来的机遇

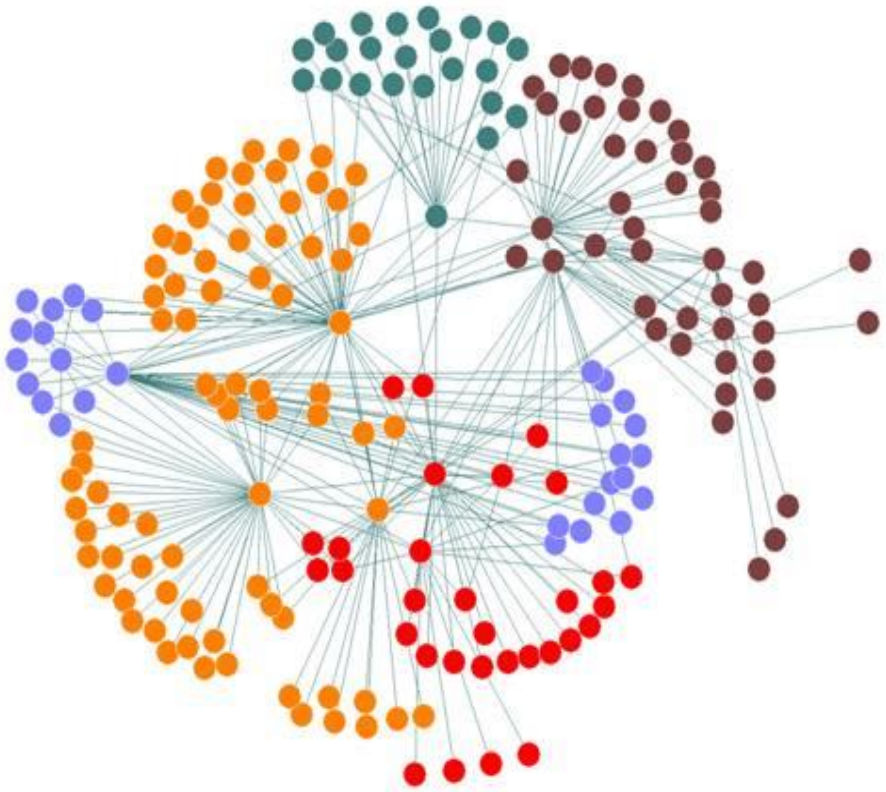
- 动态看世界的机会
  - 这个世界是什么样的 (how it looks)
  - 这个世界如何运行 （how it works)

- 微观看世界的机会
- 客观看世界 (objective, inductive)
- 及时反馈的机会
- 基于地理场所的研究 （place-based research）
- …

大数据在城市研究中应用的实例1

传统的研究方法还有用吗？

社交网络拓扑结构分析- 研究实例1

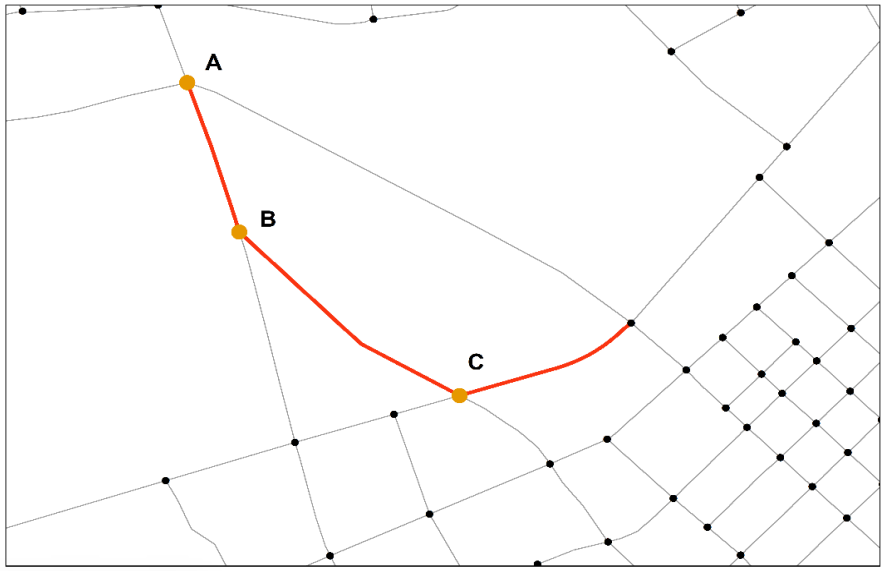


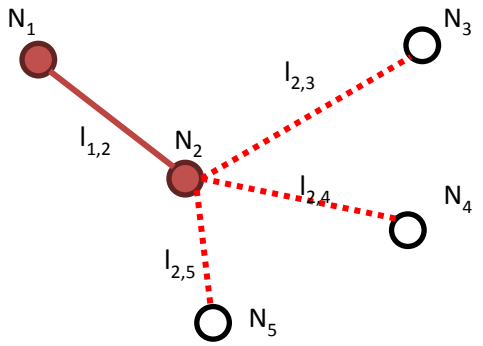
各种Centrality 指标（中心性）

- Degree
- Closeness
- Betweenness
- …
- PageRank
- 加权PageRank

一个新的方法- 随机行走指数

模拟若干随机行走路径后，计算每个结点（或链接）的被访问数



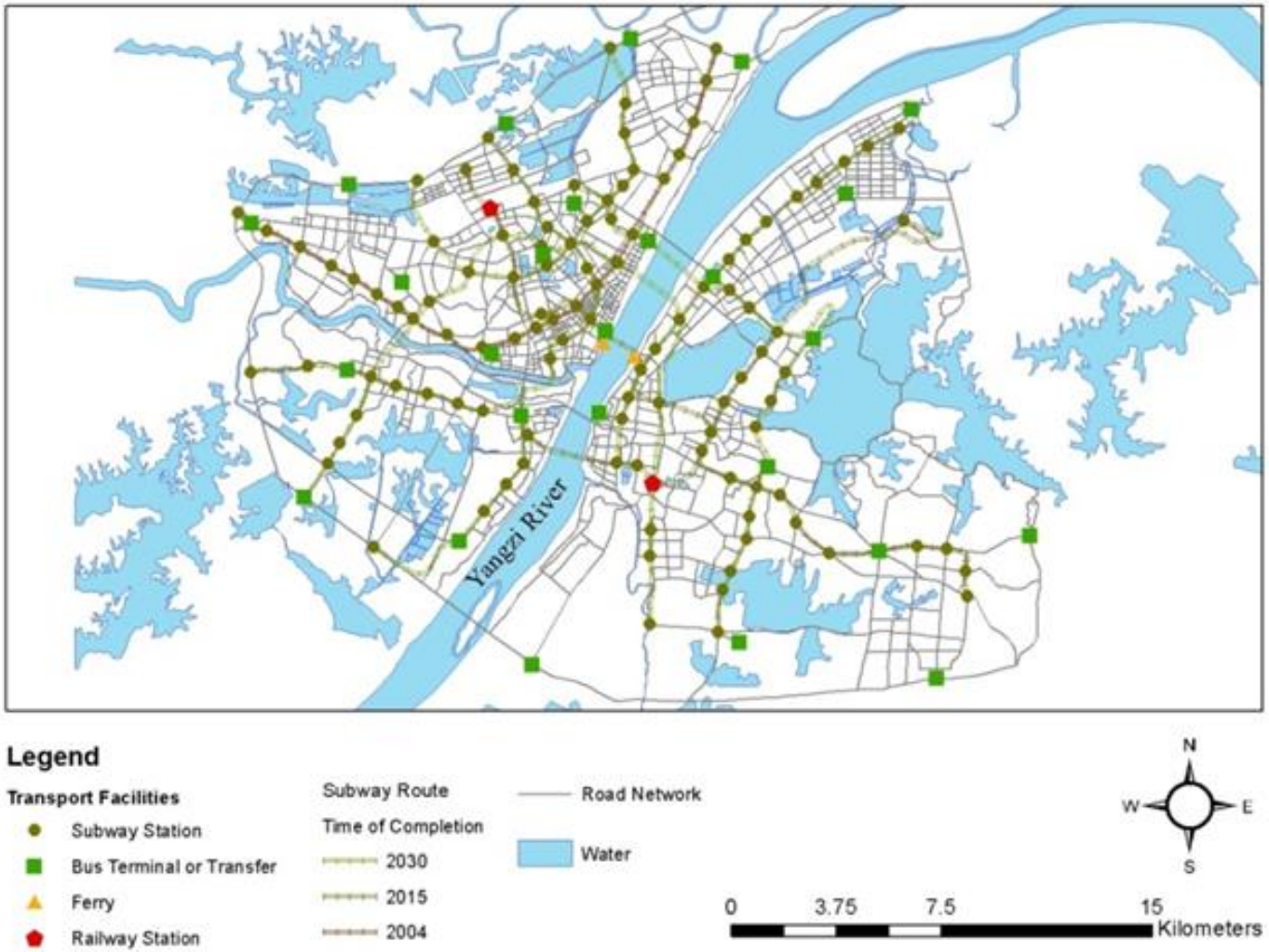


Length threshold

$$= \text{Expectation} + \text{Standard deviation} \times \sqrt{2 \times \log\left(\frac{1}{1-R_1}\right)} \times \cos(2 \times \text{Pi} \times R_2)$$

$$p_{l_{2,3}} = \frac{w_{l_{2,3}}}{w_{l_{2,3}} + w_{l_{2,4}} + w_{l_{2,5}}}$$

The Case Study City and Data



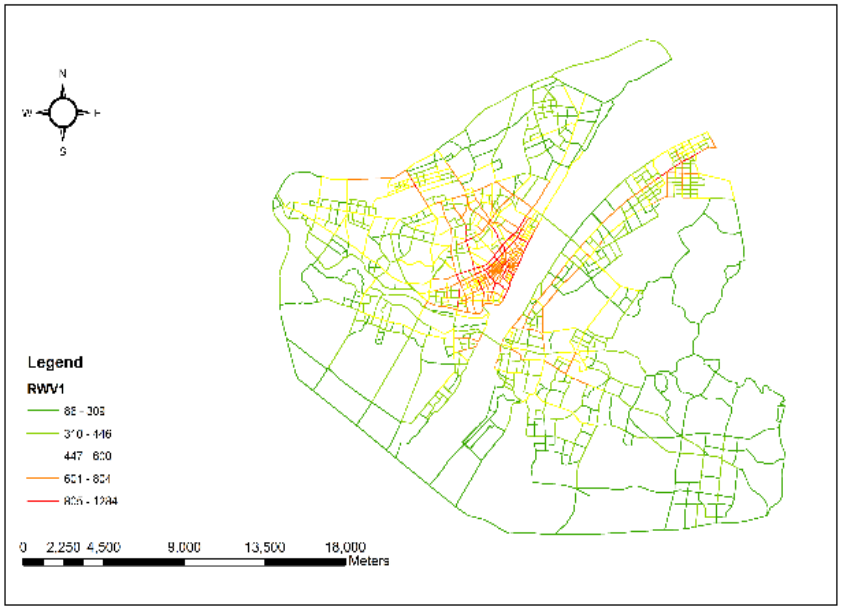
Wuhan, China

Data:
road network

随机行走指数

- 用于预测个节点的中心性（或其它空间特征）
- 在武汉和亚特兰大的实例研究

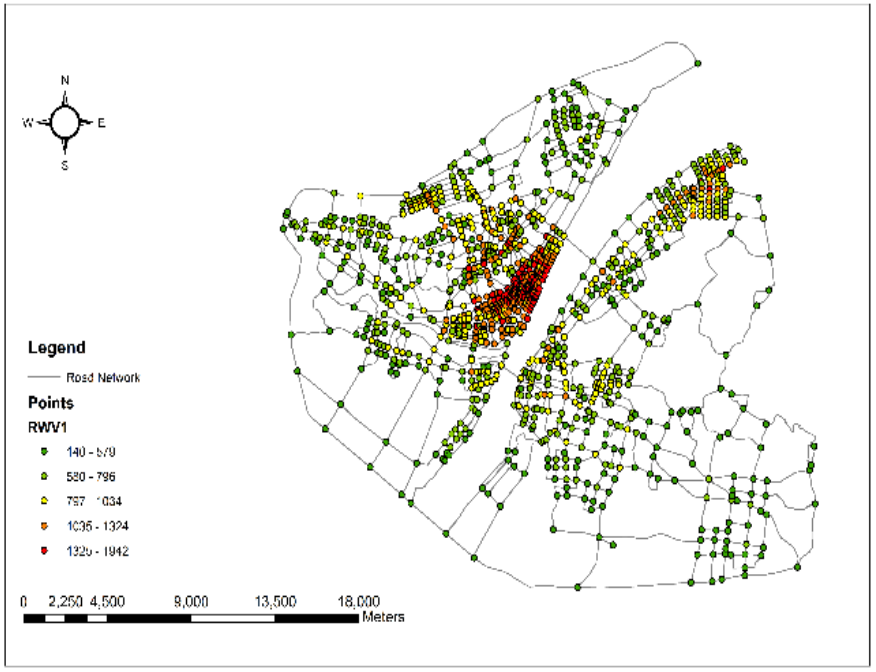
RWVs of nodes



RWVs of links

随机行走指数

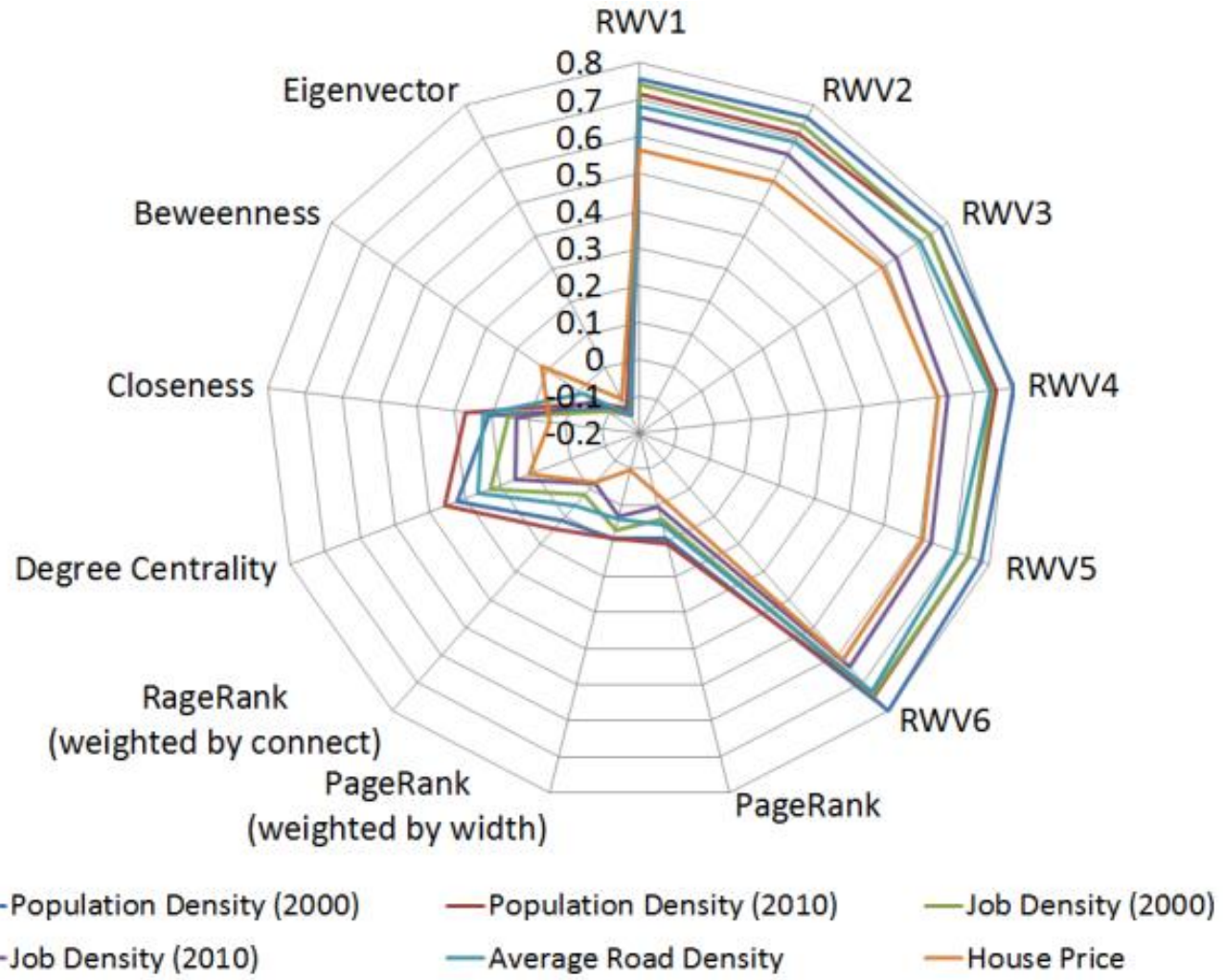
- 在武汉和亚特兰大的实例研究



RWVs of nodes

人口分布

RWV与其它常用特征指数的比较分析



大数据实例研究2

大数据不神秘

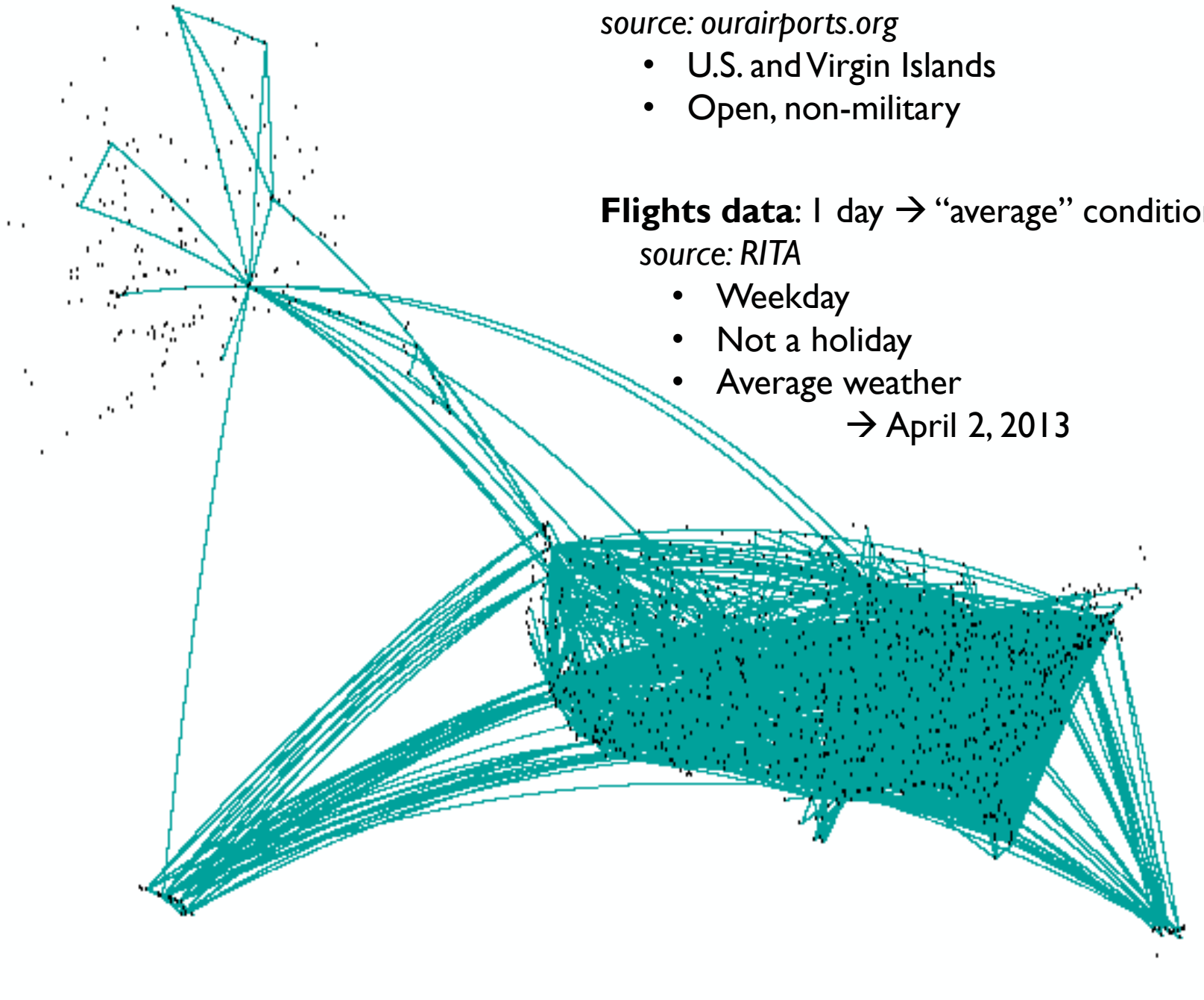
美国六大航空公司运行网络与航班准点运行状况分析

数据

- 网上下载的实时航班运行状况数据（大数据）

研究方法

- 建立各自的运营网络
- 提取航班运行的时空数据
- 分析晚点航班的空间分布
- 分析网络特征于航班晚点状况之间的关系



Airports data:

source: ourairports.org

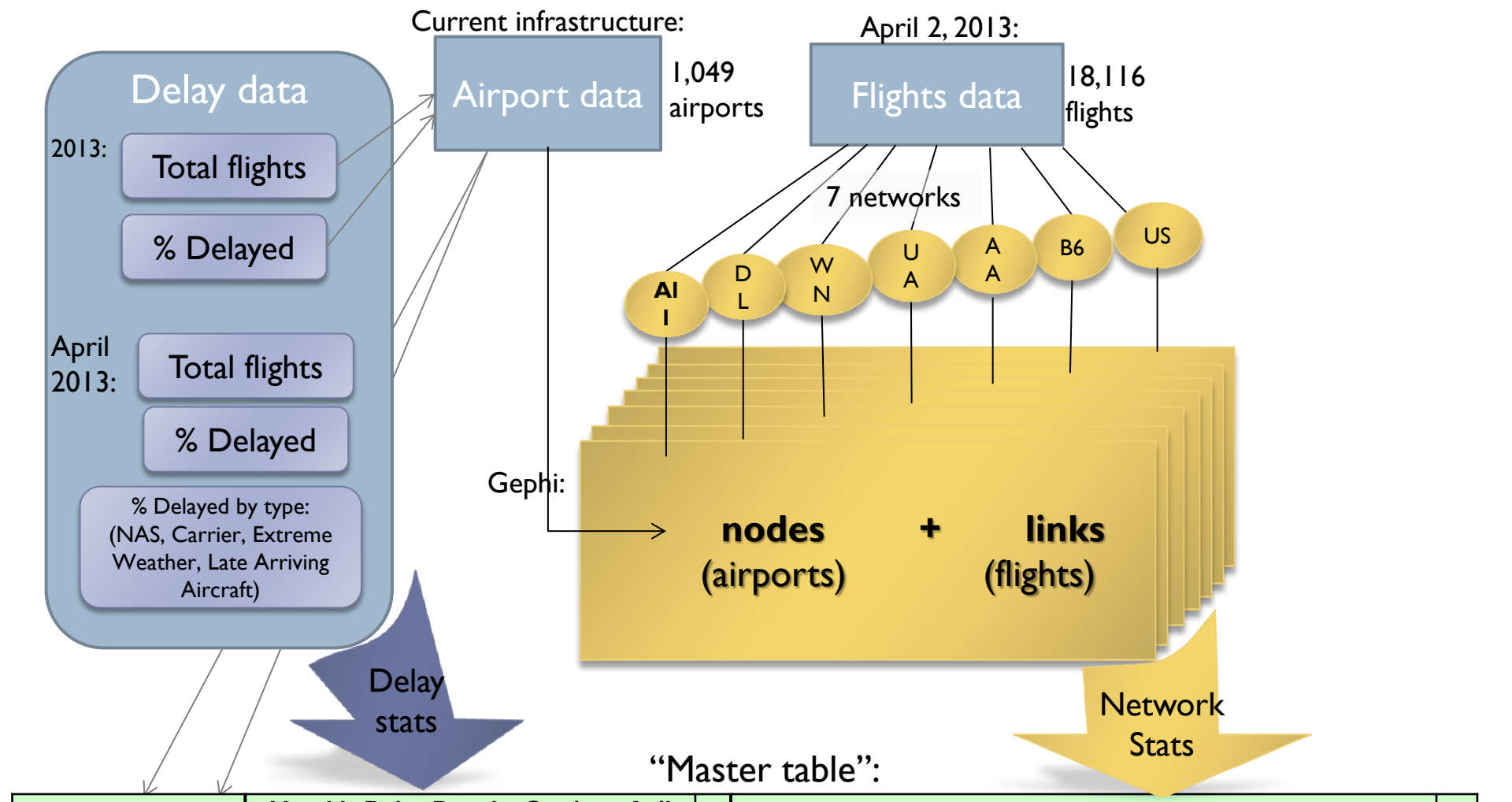
- U.S. and Virgin Islands
- Open, non-military

Flights data: 1 day → "average" conditions

source: RITA

- Weekday
- Not a holiday
- Average weather

→ April 2, 2013

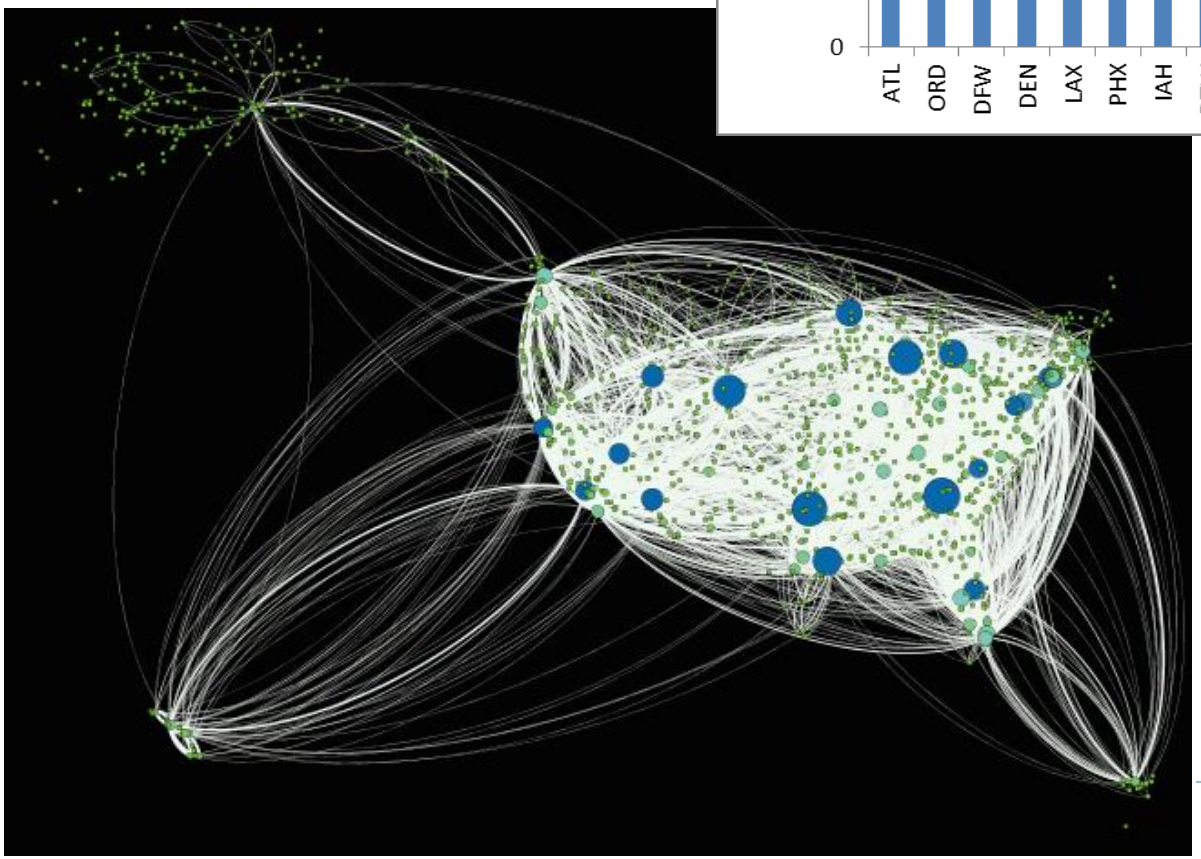


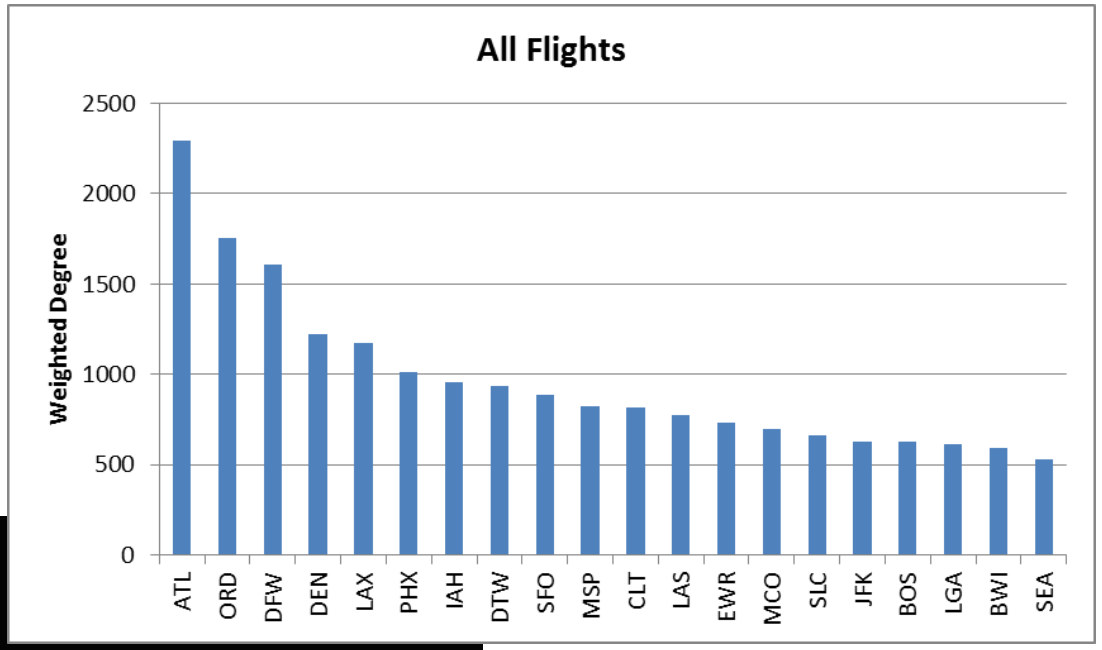
"Master table":

Airports info				Monthly Delay Data by Carrier – Apil 2013							April 2, 2013 Networks							
Code	Coordinates	2013 Total flights	2013 % Delayed	Total Flights	Total % Delayed	% Carrier Delay	% NAS Delay	% Extreme weather Delay	% Late arriving aircraft Delay	→	Degree	Weighted Degree	Betweenness Centrality	Page Rank	Eigenvector Centrality	Closeness Centrality	Clustering Coefficient	Hub (Edge Importance)	→
ABQ										→									→
ATL										→									→
BOS										→									→
↓										→									→

Columns x 6 for each air carrier

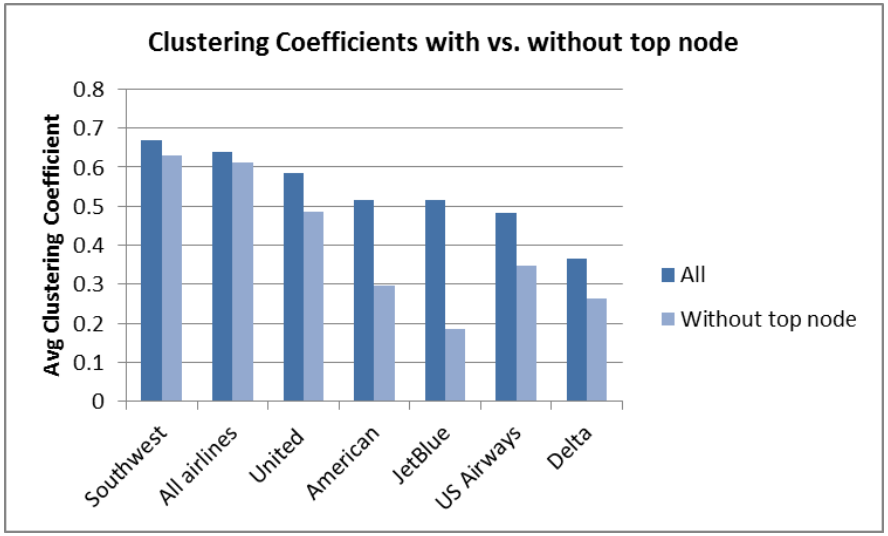
Columns x 7 for total network, plus each air carrier



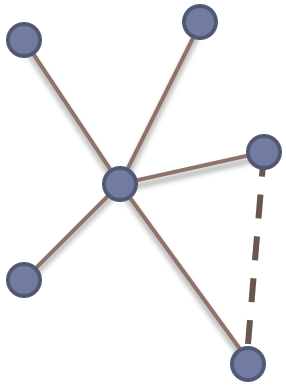

All Flights
Weighted Degree
2500
2000
1500
1000
500
0
ATL
ORD
DFW
DEN
LAX
PHX
IAH
DTW
SFO
MSP
CLT
LAS
EWR
MCO
SLC
JFK
BOS
LGA
BWI
SEA

topological network measures

	April 2, 2013											April, 2013
	# Nodes	# Edges	Total Flights	Highest Degree	Diameter	Avg Degree	Avg Weighted Degree	Graph Density	Modularity	Avg Clust coeff	Avg Path Length	% On-Time
All	294	3932	18116	2294 (ATL)	4	13.4	61.6	0.046	0.3	0.639	2.407	77%
Delta	127	562	2095	1160 (ATL)	3	4.4	16.5	0.035	0.166	0.366	2.143	86%
Southwest	79	1012	3396	466 (MDW)	3	12.8	43.0	0.164	0.352	0.669	1.938	78%
United	74	413	1333	332 (IAH)	4	5.6	18.0	0.076	0.171	0.586	2.187	76%
American	77	317	1482	856 (DFW)	3	4.1	19.2	0.054	0.082	0.517	2.014	72%
US Airways	73	279	1180	509 (CLT)	3	3.8	16.2	0.053	0.25	0.484	2.139	81%
JetBlue	52	253	674	237 (JFK)	3	4.9	13.0	0.095	0.192	0.515	2.114	72%



Clustering coefficient:



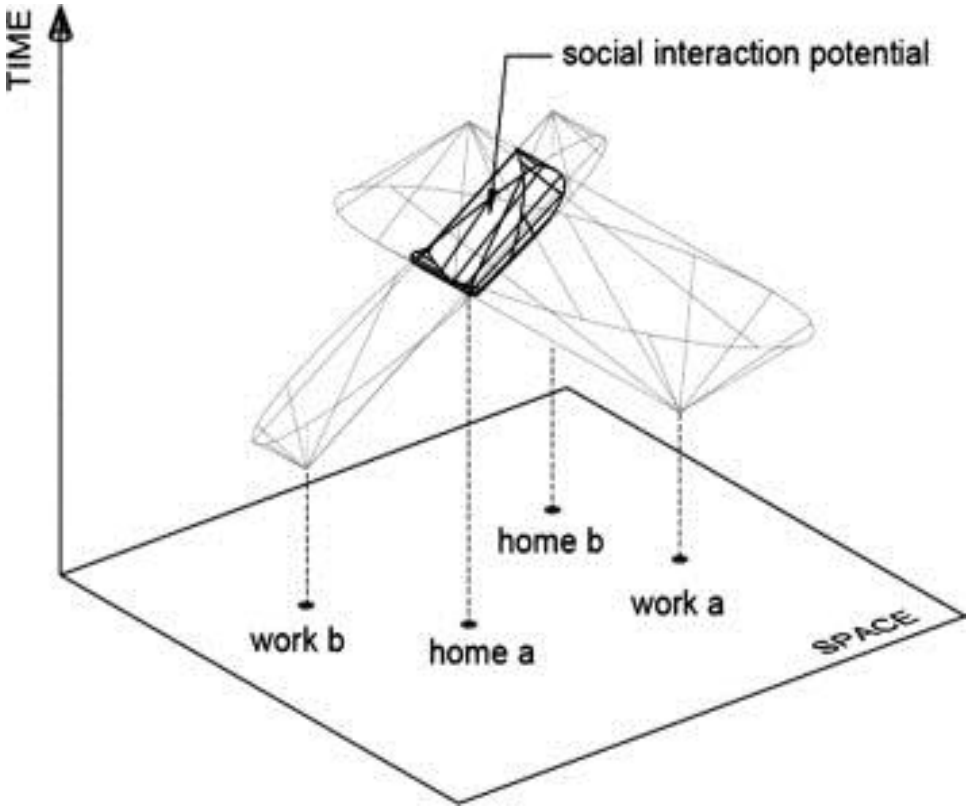
Probability that two nodes that are directly connected to a third node are also directly connected to each other

实例研究3 – 社交空间和物理空间活动之间的关联性研究

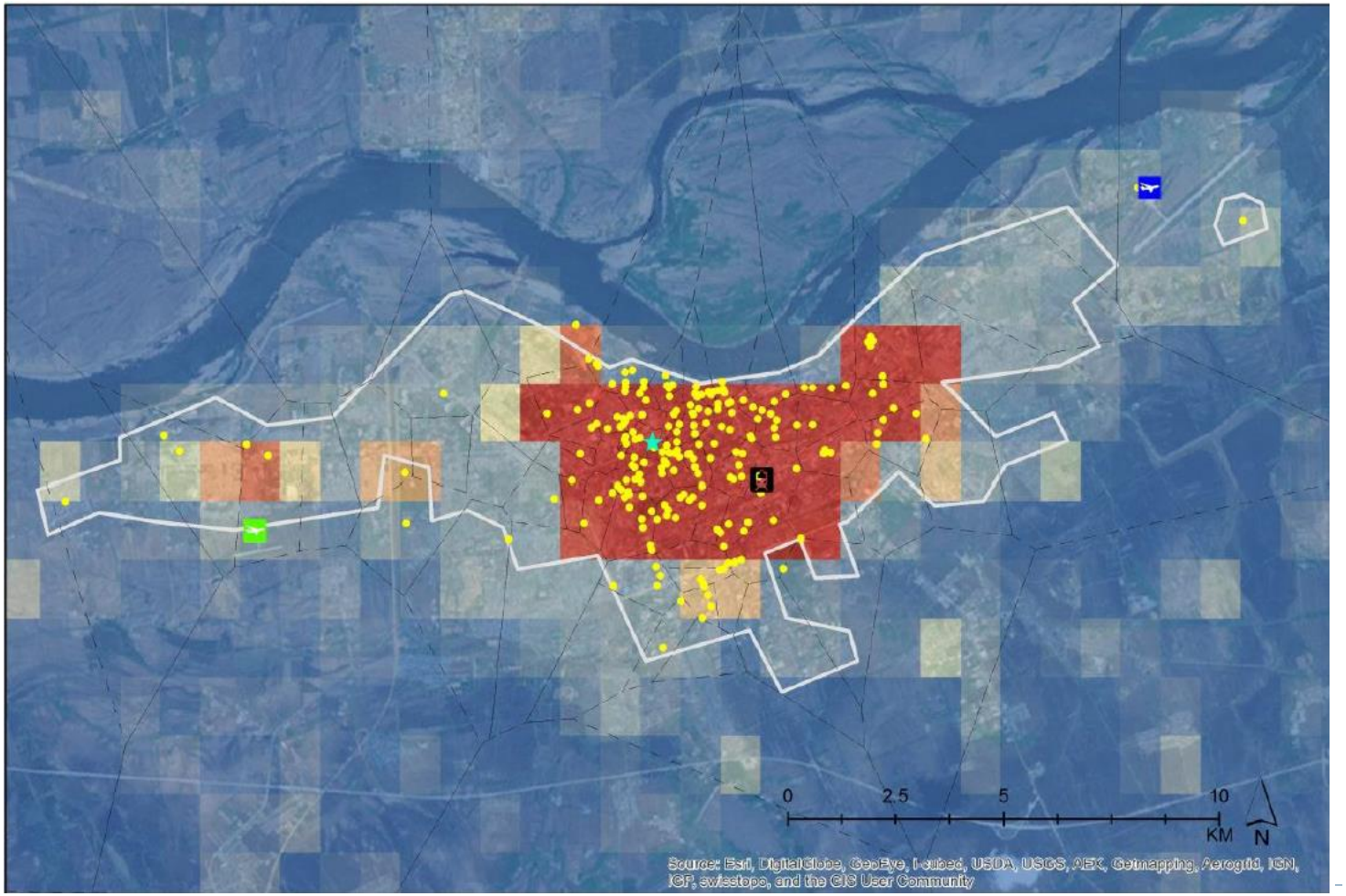
大数据让我们可以提出并回答新的问题

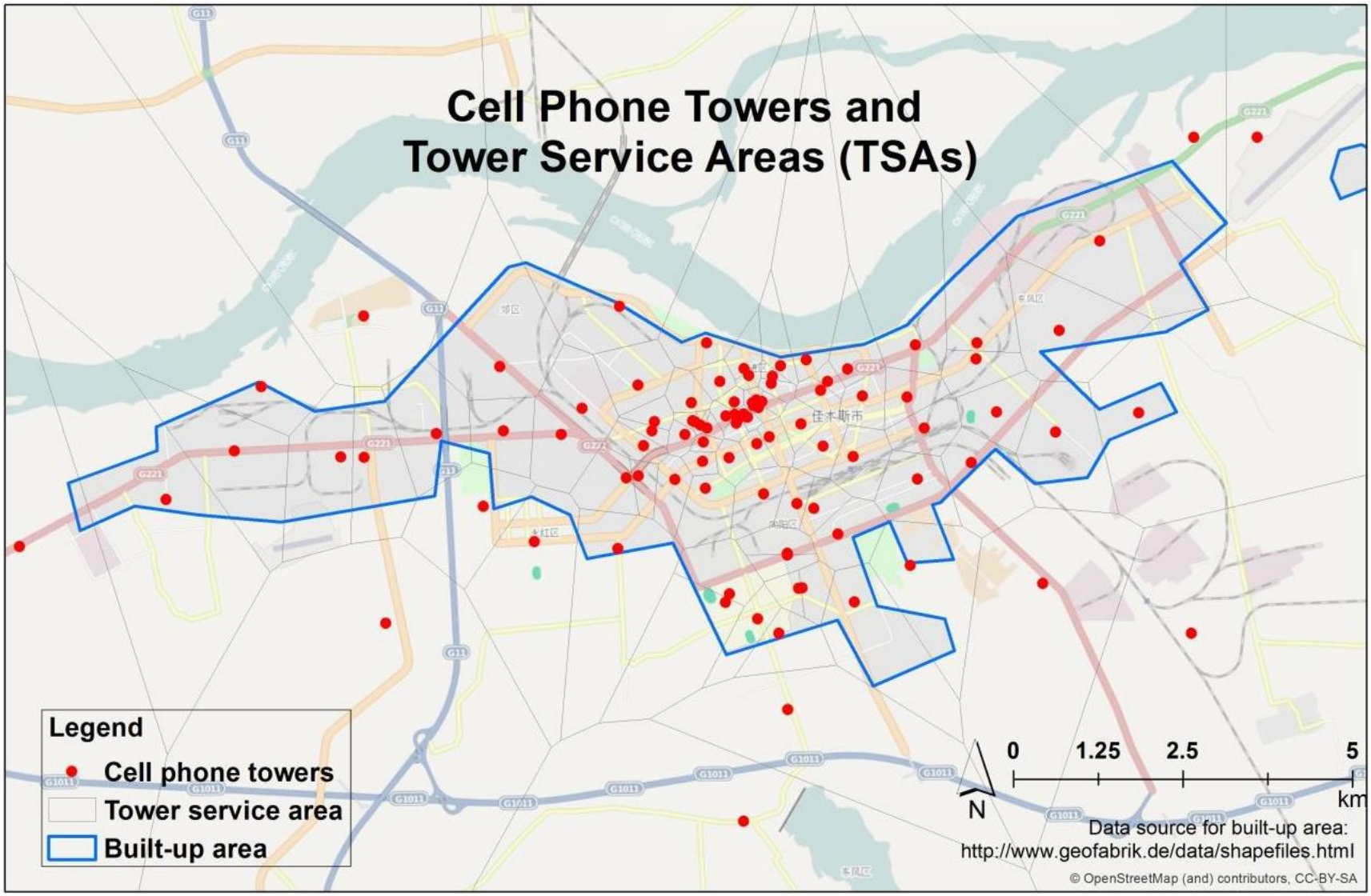
社交空间中关系紧密的朋友是否在物理空间中的活动规律也类似？

- 使用modularity-based 聚类分析
- 数据： 手机数据

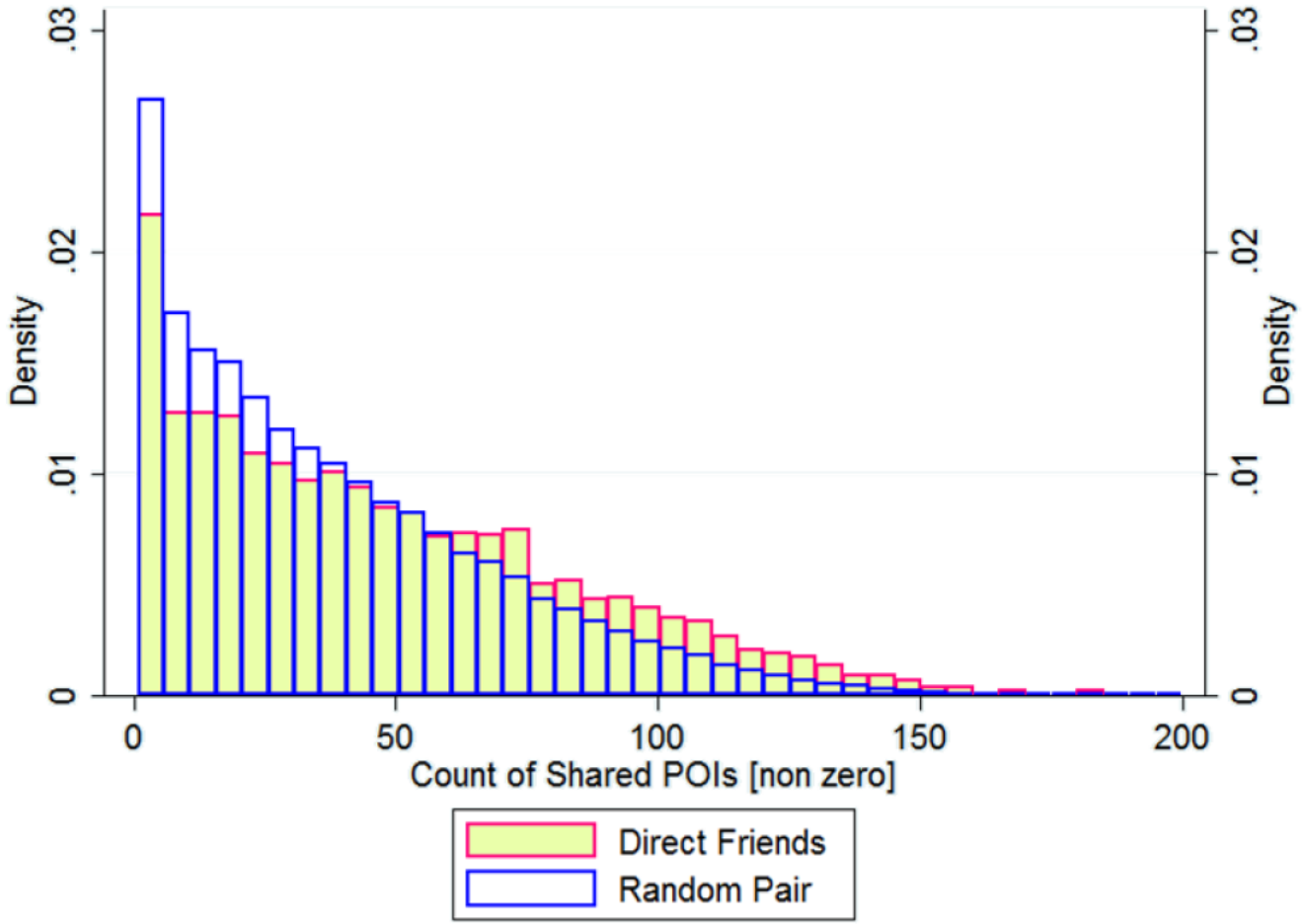


实例研究城市 – 佳木斯

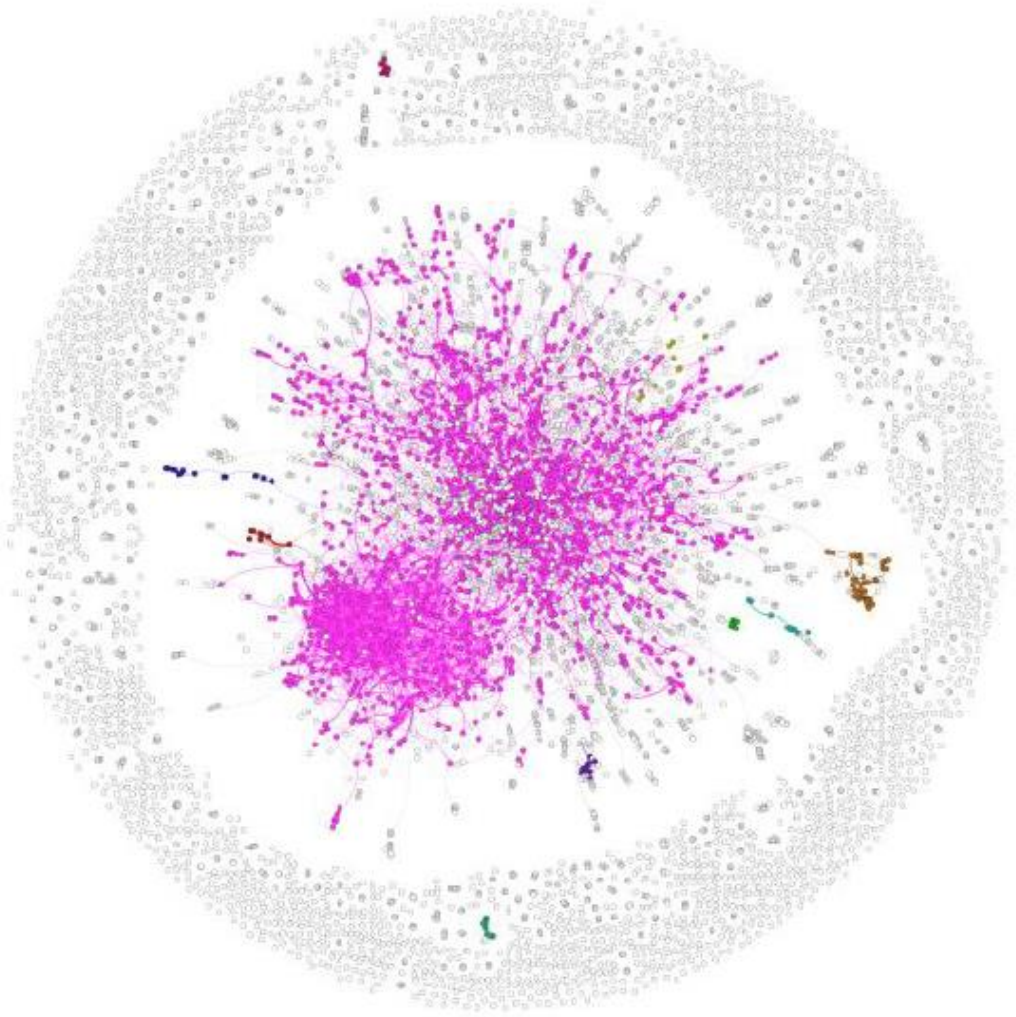

0
2.5
5
10
KM
N
Source: Esri, DigitalGlobe, GeoEye, i-cubed, USDA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community


Cell Phone Towers and
Tower Service Areas (TSAs)
Legend
Cell phone towers
Tower service area
Built-up area
0
1.25
2.5
5
km
N
Data source for built-up area:
http://www.geofabrik.de/data/shapefiles.html
© OpenStreetMap (and) contributors, CC-BY-SA

共享同一公共空间的可能性对比（朋友vs任意两个人）



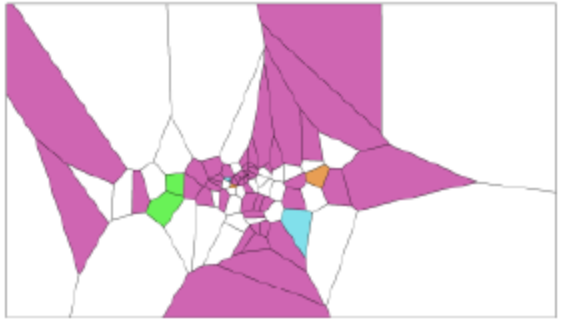
社交空间聚类分析结果 （Day2）



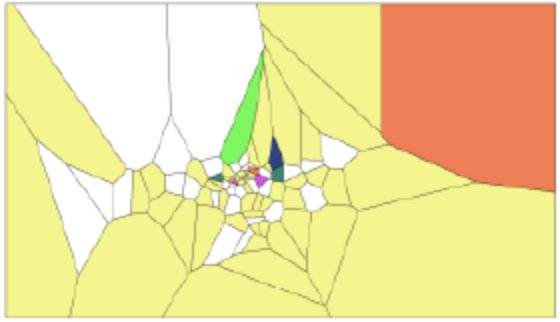
反映到城市空间

Day 1 Pattern

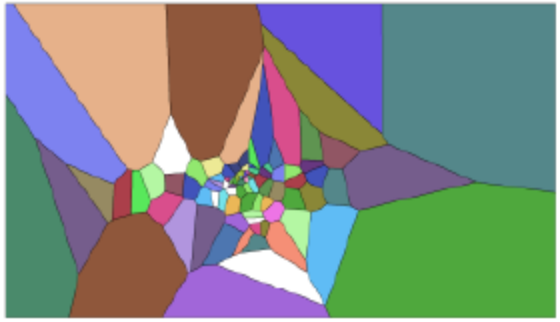
[A]



[B]



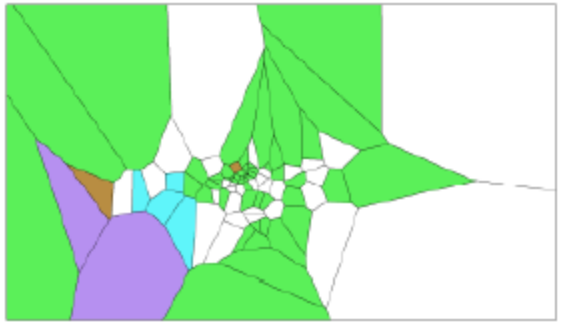
[C]



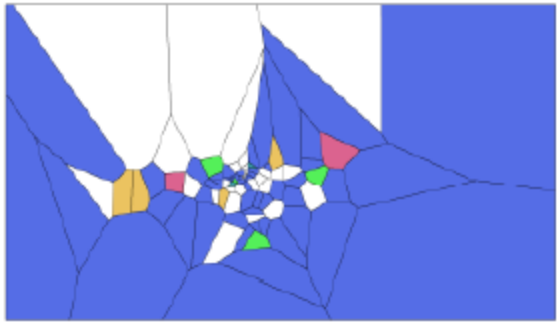
Colors are randomly assigned;
the same color stands for the same community.

Day 2 Pattern

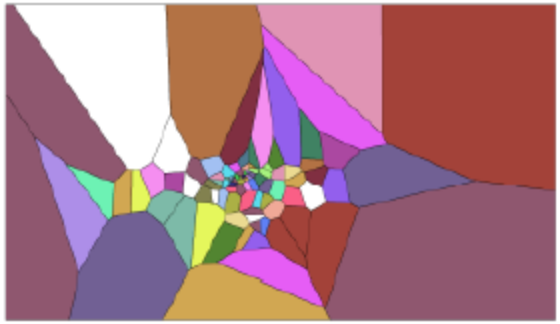
[A]



[B]



[C]



Colors are randomly assigned;
the same color stands for the same community.

分析结果比较

Table 2.5 t-test (two-tail) result for the *similarity coefficient* (*SC*)

(Sample size = 100, degree of freedom = 99)

Day	**Mean of the *rSC*s**	**[95% Conf. Interval of *rSC*]**		***oSC***
D1	0.0154	0.0152	0.0156	0.0335*
D2	0.0168	0.0166	0.0170	0.0356*

大数据带来的挑战 （Big Is Bad ?)

- 数据质量问题
- 数据代表性的问题
- 数据分析方法
- 数据结构
- 隐私保护

总结

- 大数据时代已经到来，我们无法选择是否面对它，只能选择如何面对－通过迎接挑战，抓住机遇。
- 大数据给城市研究带来了从研究内容到研究方法等多方面的研究机遇。势必给城市研究开启一个新时代。城市中的个体行为模型，互动模型，空间集合模型，时空分析模型等等将会有史无前例的全新数据支持。
- 传统的分析方法和研究方法在大数据时代大有用武之地。必要时可以结合大数据的特点做出相应调整。新方法也需要应运而生。
- 大数据的多样性呈现在质量，分析方法等各个方面。不能笼统批评或神话大数据及其数据产品。需要针对审慎对待每一个项目。
- 会出现更多多学科性质的合作研究，需要更多复合型人才
- 。。。

Acknowledgement

报告中提及的三个案例分别基于下面三个研究生与我合作或独立研究的工作：

Xuebin Wei

Lauren Anderson

Yaoli Wang

谢谢，欢迎讨论！