Towards Big Graph Search:

Challenges & Techniques

马 帅

北京大数据与脑机智能高精尖中心　　软件开发环境国家重点实验室

北京航空航天大學
BEIHANG UNIVERSITY

国家重点基础研究发展计划

- 网络信息空间大数据计算的基础研究(2014-2018)
  - Chief Scientist: Prof. Jinpeng Huai.
  - 8 institutes involved
  - Focus on “computing theory and practice on Big Data”
  - http://cnbigdata.org/

北京航空航天大學 BEIHANG UNIVERSITY

國防科学技術大学 National University of Defense Technology

ISCAS

HUAWEI

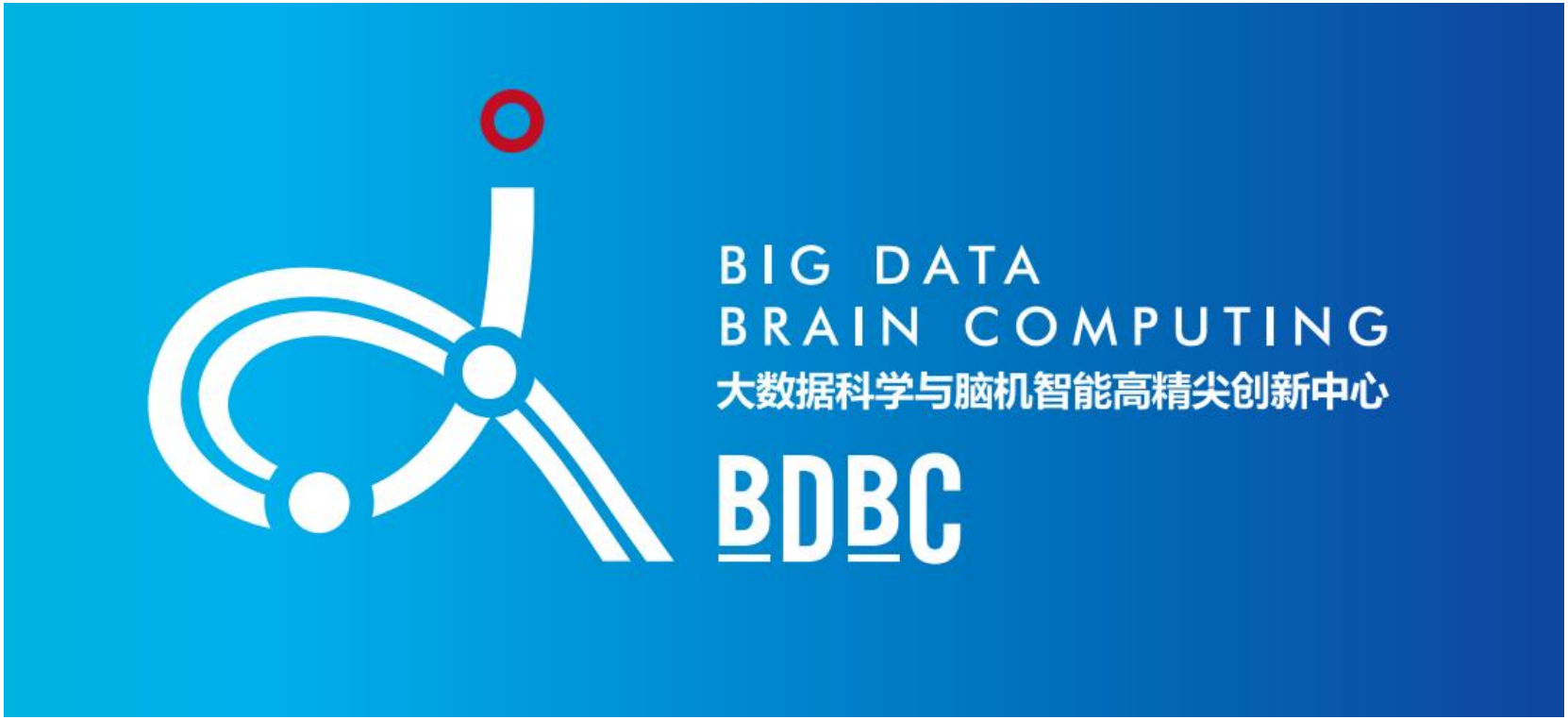
香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

CNCERT/CC 国家计算机网络应急技术处理协调中心

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Baidu 百度

北京市大数据科学与脑机智能创新中心

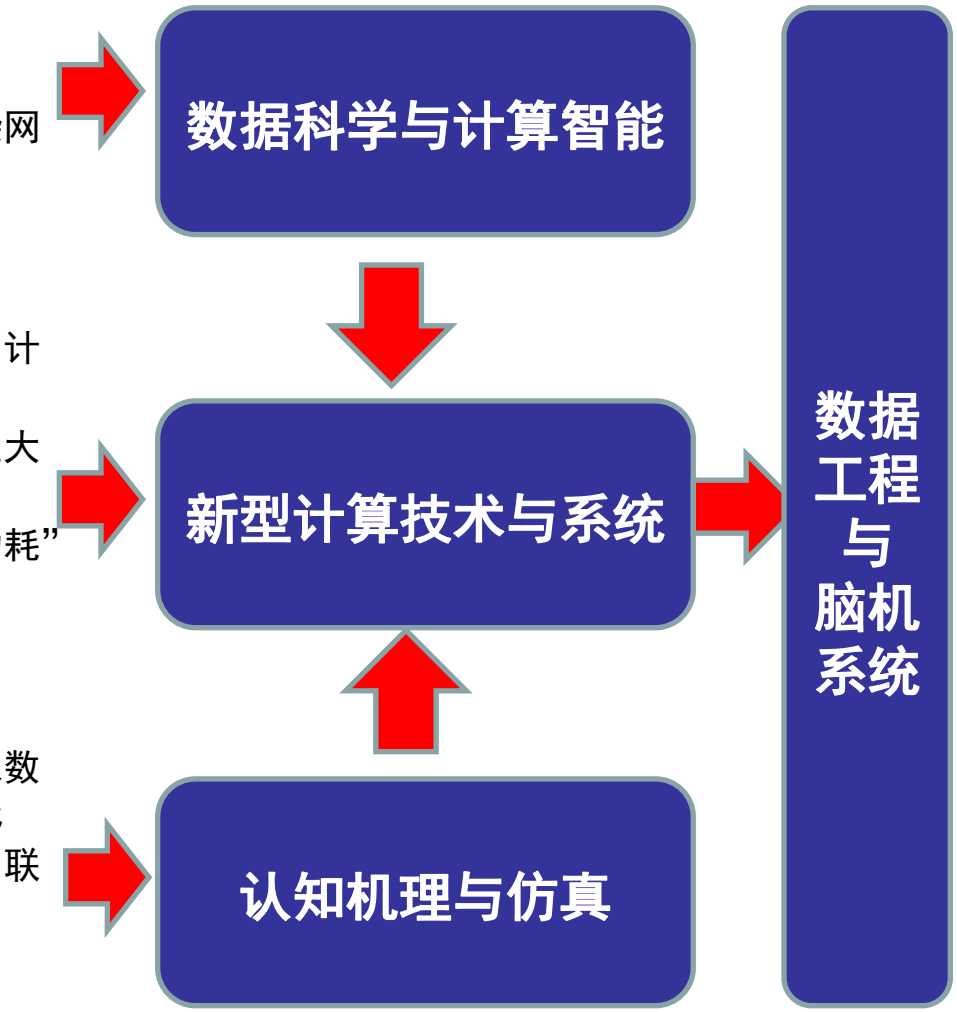


- 2015年，北京市首批北京高校高精尖创新中心
- 引领未来数据科学与计算智能的研究与应用方向
- 加速计算科学、数据科学与脑科学的交叉研究
- 促进高效智能的下一代计算与数据分析技术创新
- 通过以数据为中心的智能机器、系统及应用改变未来

研究方向与机构设置

- 瓶颈1：计算的有效性遇到障碍
  - 计算的有效性：
  - 认识数据的内在特征，复杂网络、数学（统计）方法

- 瓶颈2：能耗成为突出问题
  - 随着规模增大，调度复杂，计算系统功耗问题日益突出
  - 传统存算分离的结构，产生大量的数据搬移开销
  - 传统的计算和存储器件“功耗”不友好
- 瓶颈3：学习效率和灵活性
  - 学习效率：需要大量的输入数据及标定数据，学习效率低
  - 灵活性：普遍缺乏“类比、联想”等学习功能

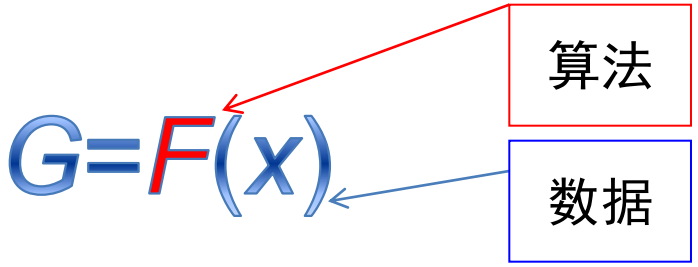


http://www.bdbc.org.cn/

BIG DATA
BRAIN COMPUTING
大数据科学与脑机智能高精尖创新中心

- 问题：是否有坚实的理论基础
- （大）数据科学是否能真的成为一种“科学”？

- 其中一个可能性：计算问题、复杂性与算法
  - 计算问题是计算机科学的本质问题，而算法是一切计算问题的核心



70年代前	•算法研究
70年代	•确定性多项式时间算法 •发现NP困难性
80年代	•随机化算法 •随机性能加速算法
90年代	•近似算法 •后期发现近似困难性

John E Hopcroft Robert Tarjan (1986)　Stephen Cook (1982)　Donald Knuth (1974)

Leslie Valiant (2010)　Manuel Blum (1995)　Juris Hartmanis , Richard Edwin Stearns (1993)

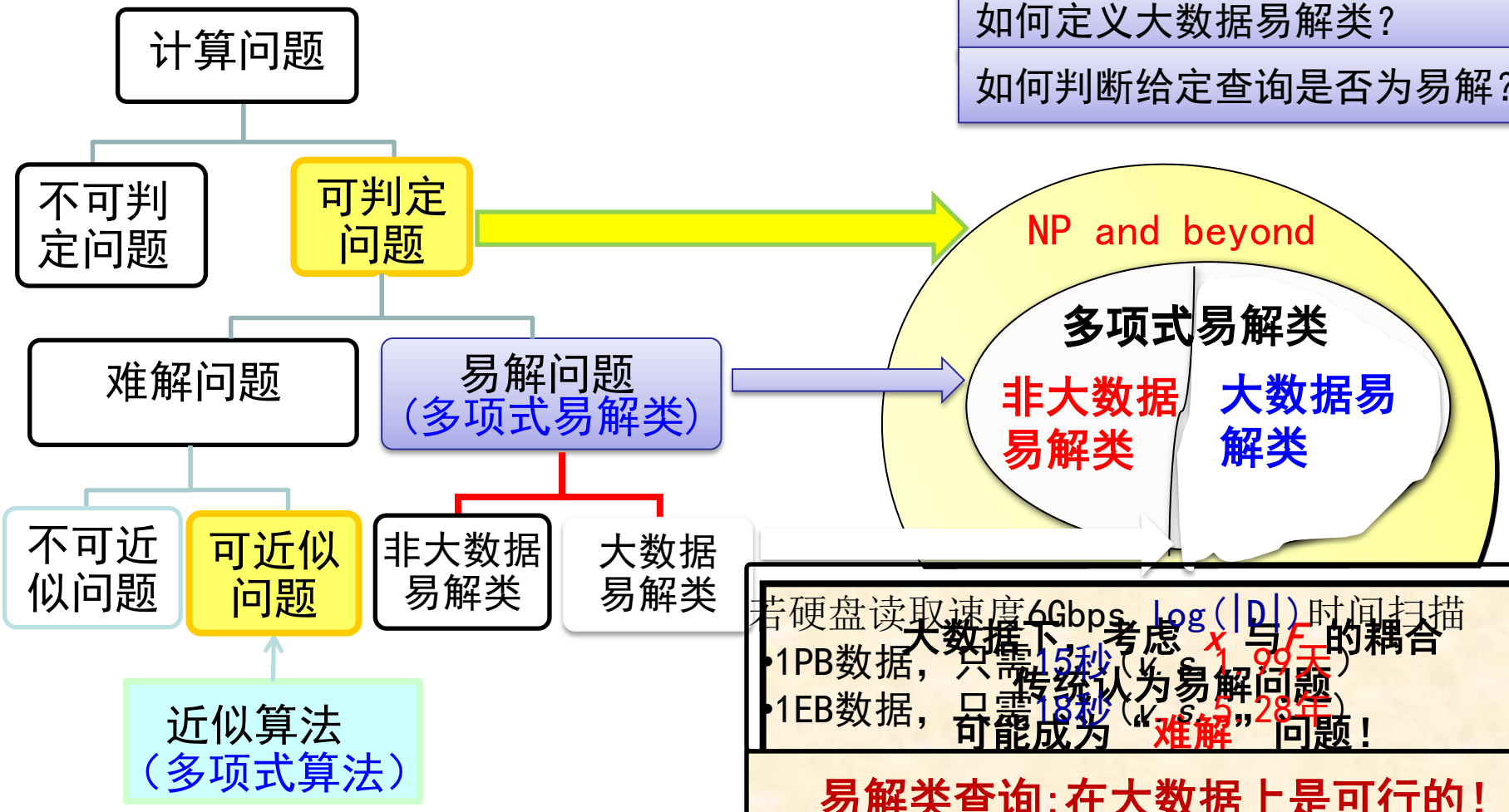
21世纪一大数据时代：计算复杂度与算法理论是否有新的理论问题和新方法？

回答“可计算”问题(1)

$G=F(x)$

任务1：易解类复杂性理论

- 如何定义大数据易解类？
- 如何判断给定查询是否为易解？



若硬盘读取速度6Gbps，$\log(|D|)$时间扫描

- 1PB数据，只需15秒（$x=1.99$天）
- 1EB数据，只需18秒（$x=5.28$年）

大数据下，考虑x与F的耦合

传统认为易解问题

可能成为“难解”问题！

易解类查询：在大数据上是可行的！

针对传统易解成为实际难解问题，**提出大数据易解类复杂性理论；**发表在数据库领域顶级会议VLDB，审稿专家认为：“*The paper is going to start a new line of research and products*”

回答“可计算”问题(2)

$G=F(x)$

计算问题

不可判定问题

可判定问题

难解问题

易解问题（多项式易解类）

不可近似问题

可近似问题

非大数据易解类

大数据易解类

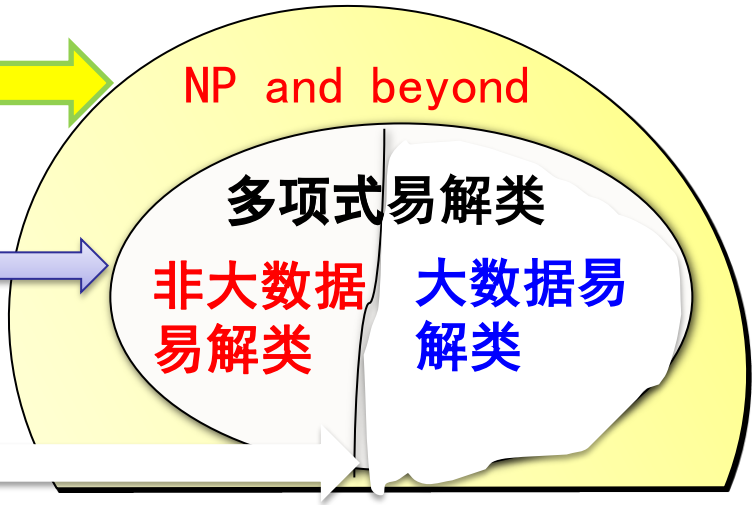
大数据不可近似问题

大数据可近似问题

任务2：数据驱动的近似算法理论

什么是大数据计算中可近似问题？

如何衡量数据量与近似效果的关系？



大数据下，传统近似方法局限性

近似的新挑战： $F \rightarrow F'$ ⊕ $X \rightarrow X'$

回答“可计算”问题(3)

$G=F(x)$

计算问题

不可判定问题

可判定问题

难解问题

易解问题（多项式易解类）

不可近似问题

可近似问题

非大数据易解类

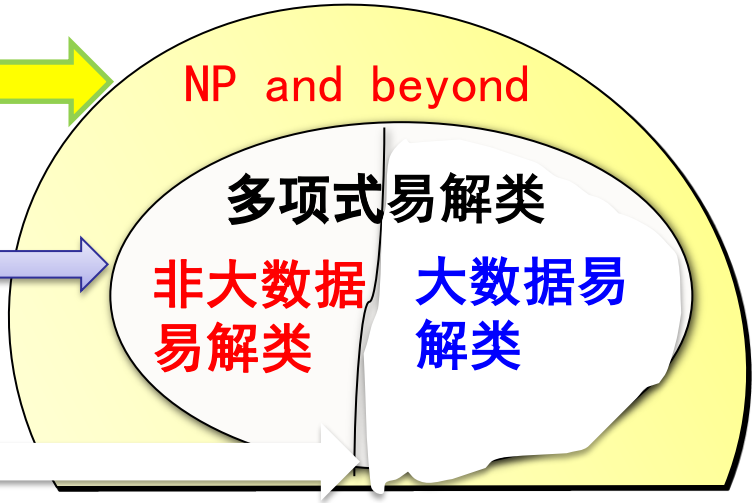
大数据易解类

近似算法（多项式算法）

任务3：大数据高效算法理论

如何设计更有效的算法？

如何以“以局部观全局”？



针对大数据非易解类问题，提出高效算法理论与算法！

大图的查询近似技术

$$R = Q(D)$$

查询近似技术

主要思想：对一类查询复杂性高的查询语言Q，变换为一类查询复杂性低的查询语言Q’，并且尽量不影响查询结果的准确性。

挑战：平衡查询的复杂性和查询的准确性!

如一，强模拟图查询

子图同构[11]:

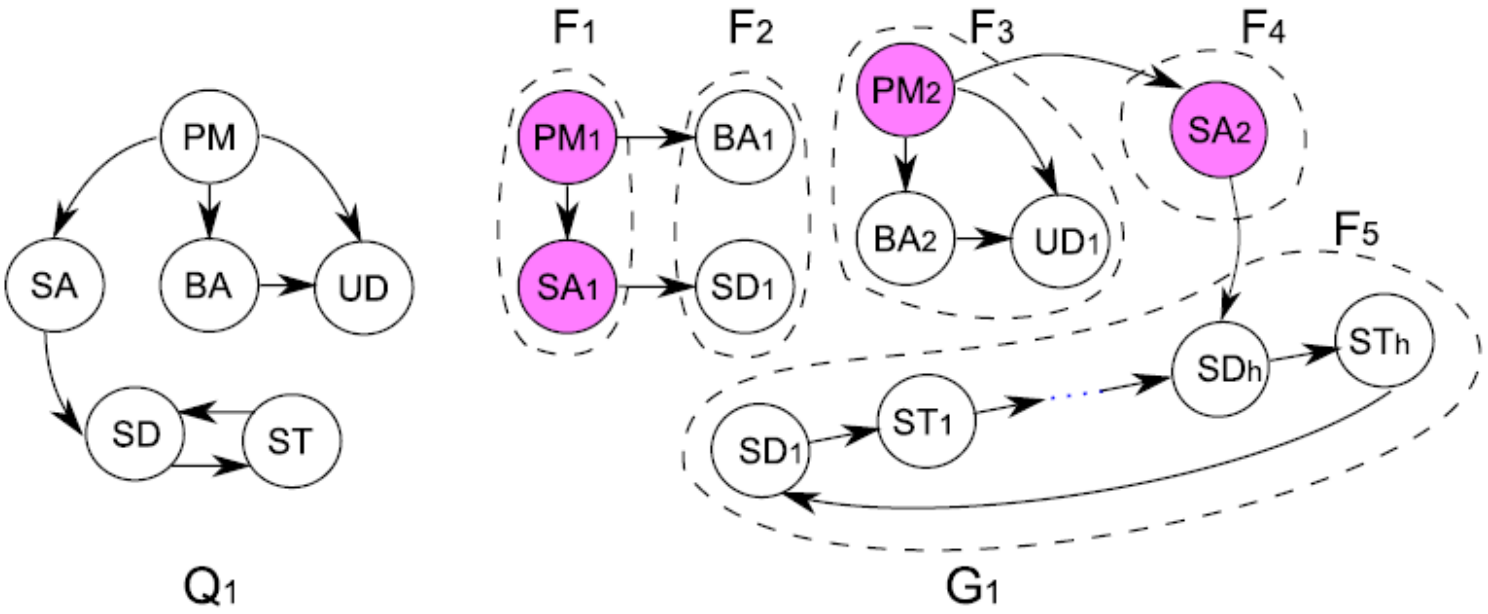
- 给定模式图Q, 数据图G的子图G_s:
  - Q 图同构 G_s 如果存在一一映射函数f: $V_Q \rightarrow V_{Gs}$ 满足：
    - ✓ 对Q中的任何顶点u，u 和f(u) 有相同的标签
    - ✓ 边(u, u')在Q当且仅当 (f(u), f(u')) 在G_s
  - Q 子图同构G，如果G中存在如上子图Gs

优点：Q 和 G_s一模一样

缺点：NP完全问题；最坏情况下指数个匹配子图；约束过于严格

Shuai Ma, Yang Cao, Wenfei Fan, Jinpeng Huai, and Tianyu Wo. Strong Simulation: Capturing Topology in Graph Pattern Matching. **TODS 2014**.
Shuai Ma, Yang Cao, Wenfei Fan, Jinpeng Huai, and Tianyu Wo, Capturing Topology in Graph Pattern Matching. **VLDB 2012.**

子图同构图查询

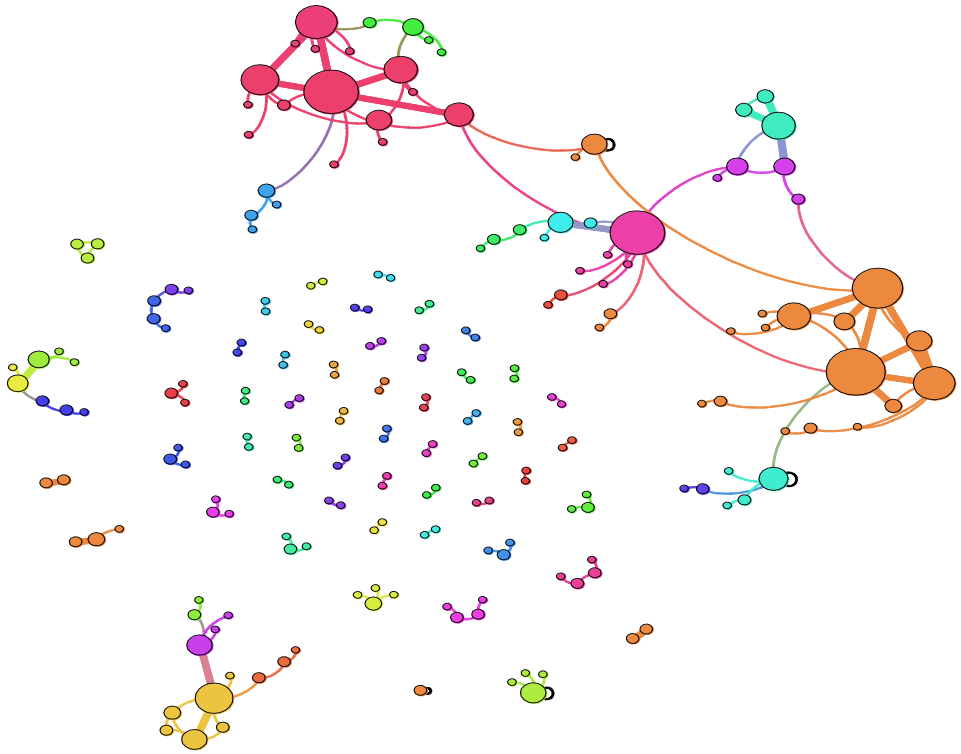


组成一个软件开发团队

强模拟：返回$F_3 + F_4 + F_5$;
子图同构：返回空集！

子图同构约束过于严格，并不适合一些新型应用！

Terrorist Collaboration Network



“Those who were trained to fly didn’t know the others. One group of people did not know the other group.” (Osama Bin Laden, 2001)

Build upon (revised) strong simulation to aid the detection of homegrown violent extremists (HVEs) who seek to commit acts of terrorism in the United States and abroad, **Colorado State University, Benjamin W. K. Hung, Anura P. Jayasumana**: Investigative simulation: Towards utilizing graph pattern matching for investigative search. ASONAM 2016.

强模拟图查询

子图同构 (NP-Complete) → 强模拟 ($O(n^3)$) → 双模拟 ($O(n^2)$) → 图模拟 ($O(n^2)$)

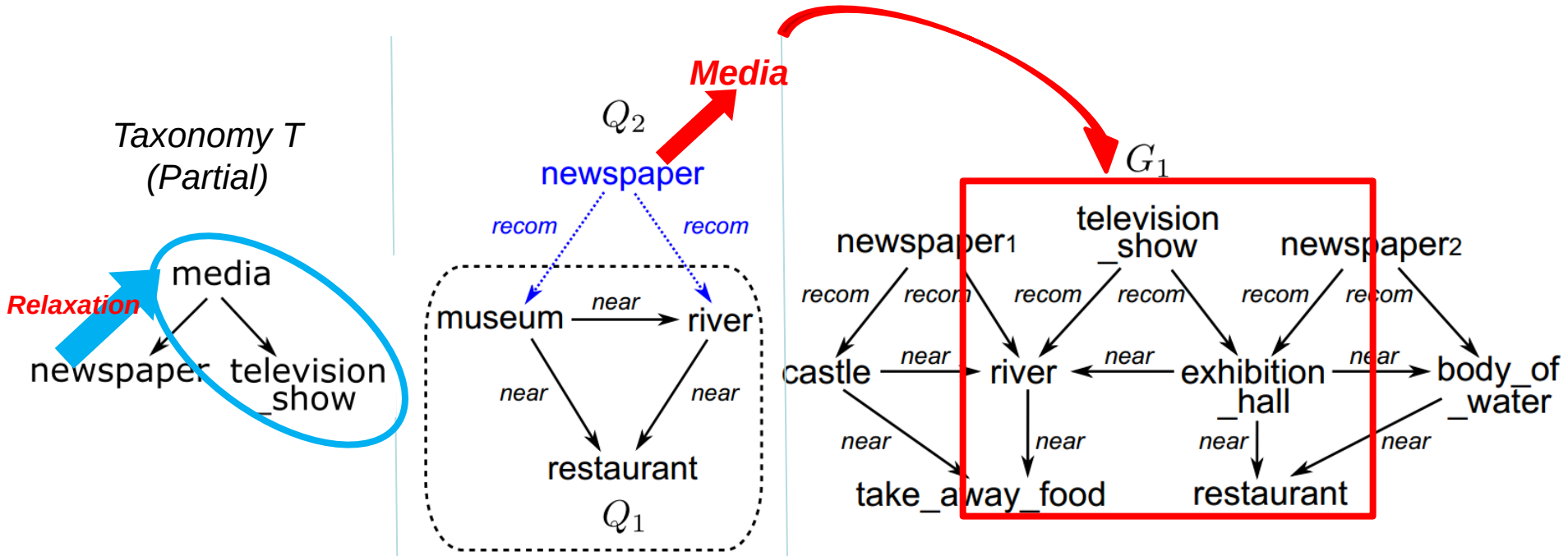
Matching	children	parents	connectivity	cycles
$\prec$	✓	×	×	✓ (directed), × (undirected)
$\prec_D$	✓	✓	✓	✓ (directed & undirected)
$\prec_D^L$	✓	✓	✓	✓ (directed & undirected)
$\lhd$	✓	✓	✓	✓ (directed & undirected)

locality	matches	Bisimilar&b'ed-cycle
×	✓	×
×	×	×
✓	✓	×
✓	×	✓

查询结果保持70-80%子图同构结构，效率提高100倍！

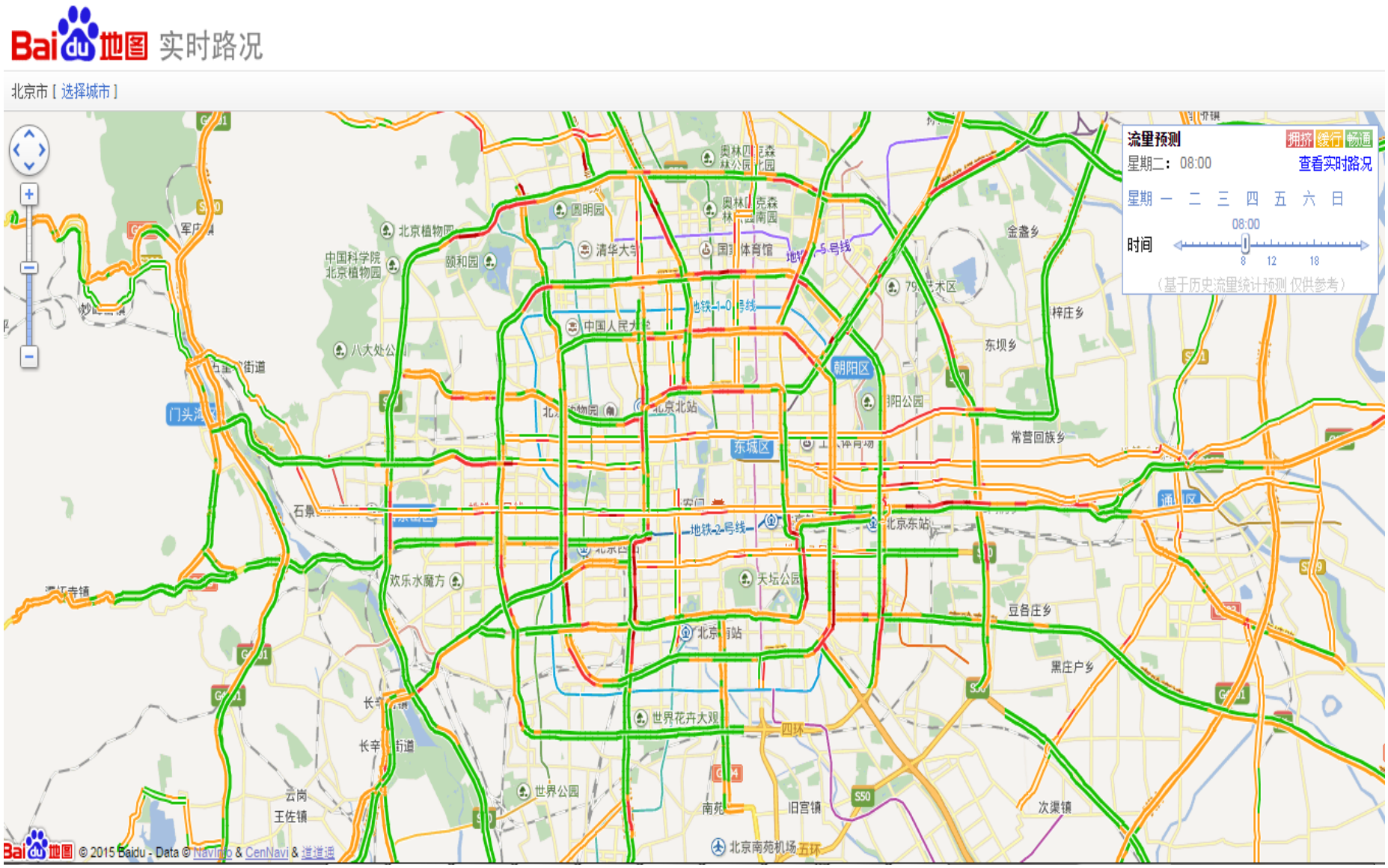
Taxonomy图模拟

$\|V_Q\|$	2	4	6	8	10
DBpedia	90%	18%	0%	0%	0%
YAGO	54%	2%	0%	0%	0%



Jia Li, Yang Cao, and Shuai Ma. Relaxing Graph Pattern Matching With Explanations. **CIKM 2017**.

如二，时态稠密图查询


Bai du 地图 实时路况
北京市 [选择城市]
流量预测
拥挤 缓行 畅通
星期二： 08:00
查看实时路况
星期 一 二 三 四 五 六 日
08:00
时间
8 12 18
（基于历史流量统计预测 仅供参考）
朝阳区
东城区
北京北站
北京东站
北京植物园
中国科学院北京植物园
颐和园
圆明园
清华大学
中国人民大学
八大处公园
欢乐水魔方
天坛公园
世界花卉大观
世界公园
北京南苑机场
南苑
旧宫镇
次渠镇
豆各庄乡
黑庄户乡
常营回族乡
东坝乡
金盏乡
梓庄乡
四环
五环
S50
© 2015 Baidu - Data © NavInfo & CenNavi & 道道通

如二，时态稠密图查询

- 筛选与验证的方法 (Filter-and-Verification)

10^4	10^5	10^6	…	10^8
5×10^2	5×10^3	5×10^4		5×10^6

过滤掉95%

- 数据驱动的查询近似方法(Data Driven Query Approximation)
  - 根据数据的特点选取k个(k是个小的常数，比如10或15)

10^4	10^5	10^6	…	10^8
k	k	k		k

- 实验结果 (With the state of the art solution[Bogdanov et al. 2011])

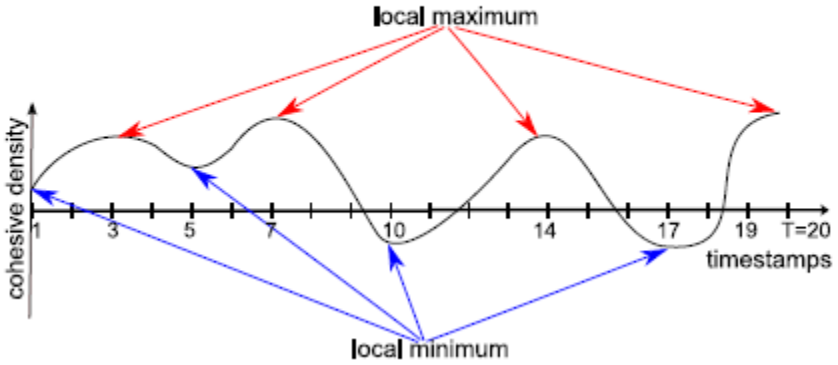
	准确性	效率
BEIJING DATA	100.25%	快4,870倍
SYNTHETIC DATA	99.69%	快1,468倍

P. Bogdanov, M. Mongiov, and A. K. Singh. Mining heavy subgraphs in time-evolving networks. In ICDM, 2011.
Haixing Huang, Jinghe Song, Xuelian Lin, Shuai Ma, Jinpeng Huai, TGraph: A Temporal Graph Data Management System (demo), **CIKM 2016.**
Shuai Ma, Renjun Hu, Luoshu Wang, Xuelian Lin, Jinpeng Huai, Fast Computation of Temporal Dense Subgraphs, **ICDE 2017**

如二，时态稠密图查询

在演化生物学中，**趋同演化**（Convergent evolution）指的是两种不具亲缘关系的动物/植物长期生活在相同或相似的环境，或曰生态系统，它们因应需要而发展出相同功能的器官的现象，即同功器官

Evolving convergence assumption



The p_{EC} are 96% on BEIJING DATA and 90% on average on all tested SYNTHETIC DATA, respectively, which justifies our observation of the evolving convergence assumption.

Proposition 2: *To find the dense subgraph, we only need to consider the time intervals $[i, j]$ such that the cohesive density curve has a local maximum at certain point between i and j under the evolving convergence assumption.* □

Fact 2: *Temporal subgraph $\mathbb{G}[i, j]$ $(i \leq j \in [1, T])$ with a higher positive cohesive density has a higher probability of containing a dense subgraph under the assumption of independent and identically distributed edge weights.* □

Shuai Ma, Renjun Hu, Luoshu Wang, Xuelian Lin, Jinpeng Huai, Finding Dense Subgraphs in Temporal Networks, **ICDE 2017**

大图的数据近似技术

$$R = Q(D)$$

数据近似技术

主要思想：对一类查询复杂性高的查询语言Q，将查询数据D变换机器能够高效处理的较小量D’，并且尽量不影响查询结果的准确性。

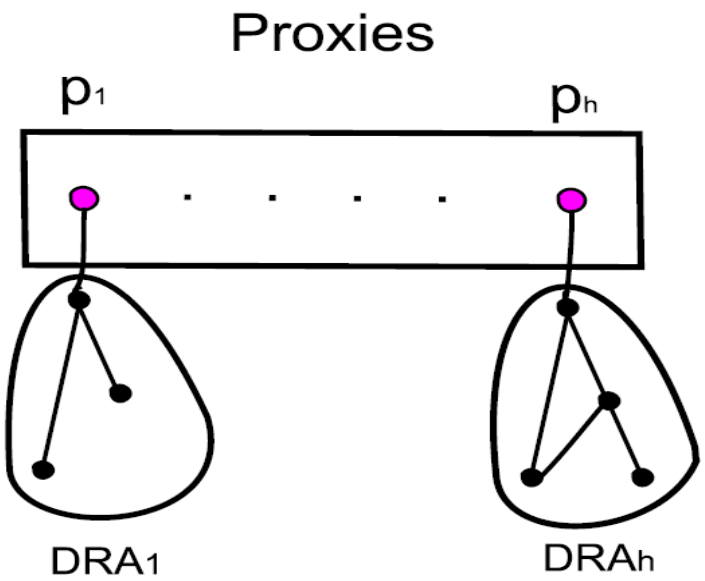
二八定律：在众多现象中，80%的结果取决于20%的原因

$$D = HARD(D) + SOFT(D)$$

挑战：平衡查询的效率和查询的准确性!

如一、最短路径/距离

- 针对有权无向图，提出 “proxies”概念
- 每个proxy代表了一个子图DRA中的顶点(互补重叠)
- Proxies 能够在O(n)时间内算出



关键性质: 给定图G、顶点u和v及其代理u_p和v_p，则：

(1) $path(u, v) = path(u, u_p) + path(u_p, v_p) + path(v_p, v)$

(2) $dist(u, v) = dist(u, u_p) + dist(u_p, v_p) + dist(v_p, v)$

在真实的公路网络和社会网络中，数据减少了1/3!
是一种针对最短路径/距离的轻量级的通用的数据缩减技术!

Shuai Ma, Kaiyu Feng, Jianxin Li, Haixun Wang, Gao Cong, and Jinpeng Huai, Proxies for Shortest Path and Distance Queries. **TKDE 2016**.

Shuai Ma, Kaiyu Feng, Jianxin Li, Haixun Wang, Gao Cong, and Jinpeng Huai, Proxies for Shortest Path and Distance Queries. **ICDE 2017 (TKDE Extended Abstract)**.

如二，网络链接预测

链接预测：

- n个顶点网络，$O(n^2)$个可能链接
- CPU速度XGHz/s，假定1个机器时钟处理1个顶点对。

Network Sizes	1 GHz	3 GHz	10 GHz
10^6 nodes	1000 sec.	333 sec.	100 sec.
10^7 nodes	27.8 hrs	9.3 hrs	2.78 hrs
10^8 nodes	$>$ 100 days	$>$ 35 days	$>$ 10 days
10^9 nodes	$>$ 10000 days	$>$ 3500 days	$>$ 1000 days

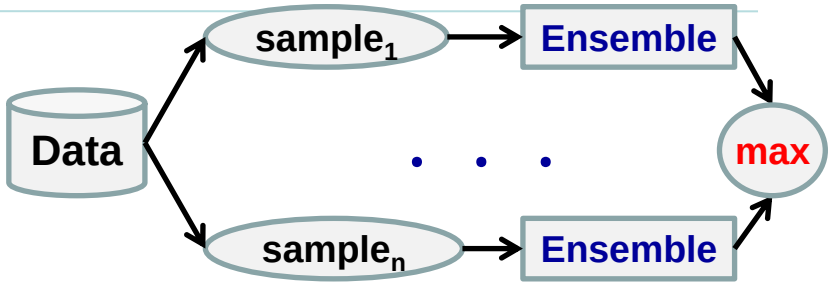
多数链接预测算法仅仅预测一个可能链接子集，而不是整个网络所有可能的链接，如[Dashun et al. 2011, Chungmok et al. 2014].

Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, Albert-László Barabási: Human mobility, social ties, and link prediction. KDD 2011.

Chungmok Lee, Minh Pham, Norman Kim, Myong K. Jeong, Dennis K. J. Lin, Wanpracha Art Chaovalitwongse. A novel link prediction approach for scale-free networks. WWW 2014.

如二，网络链接预测

- 直接采用非负矩阵分解的代价高
  - 效率低
  - 数据越稀疏，效果越差



- 数据近似技术(Ensemble Enabled Sampling)
  - 采样要保证一定的覆盖率

PROPOSITION 2. *The expected times of each node pair included in* μ/f^2 *ensemble components is at least* μ.

  - 基于链接预测特征的抽样 - Triangles
  - 结合Ensemble的思想：链接e的预测分值是所有Ensemble中的最大值

小数据	准确性	大数据	效率
YouTube	高18%	Friendster	快31倍
Wikipedia	高16%	Twitter	快21倍

同时提高了准确性和检测效率！

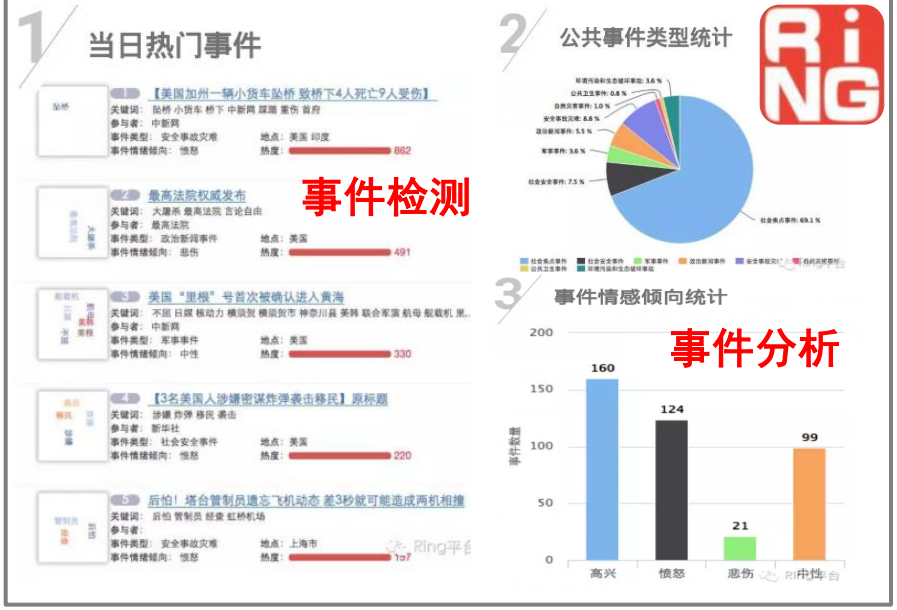
Liang Duan, Charu Aggarwal, Shuai Ma, Renjun Hu, and Jinpeng Huai, Scaling up Link Prediction with Ensembles, **WSDM 2016 - Big Data Algorithms Session**.
Liang Duan, Shuai Ma*, Charu Aggarwal, Tiejun Ma, and Jinpeng Huai, An Ensemble Approach to Link Prediction. **TKDE, 29(11): 2402-2416, 2017.**

大图搜索技术的应用

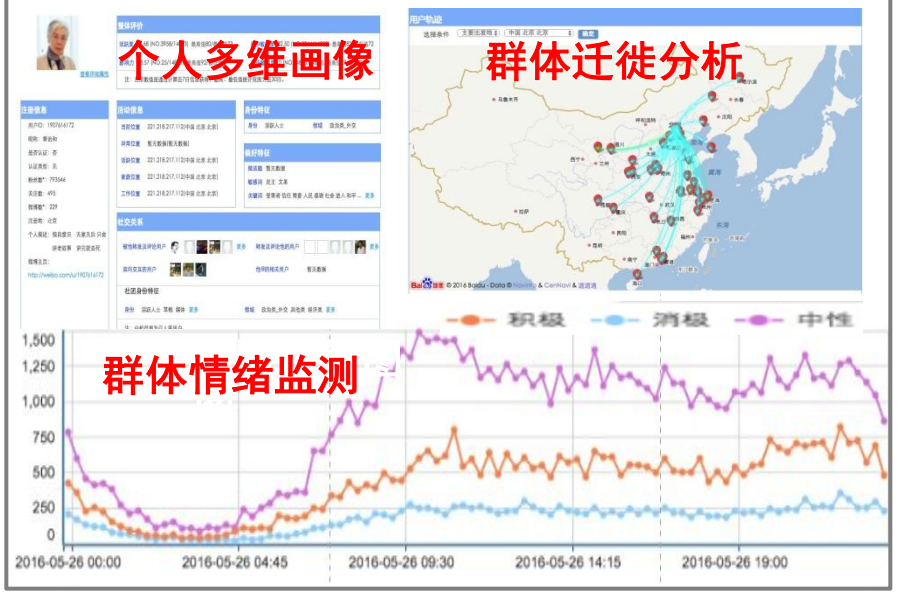
BD A

如一：面向公共事件的示范应用

安管中心：实时事件检测与分析系统



安管中心：微博人物画像系统

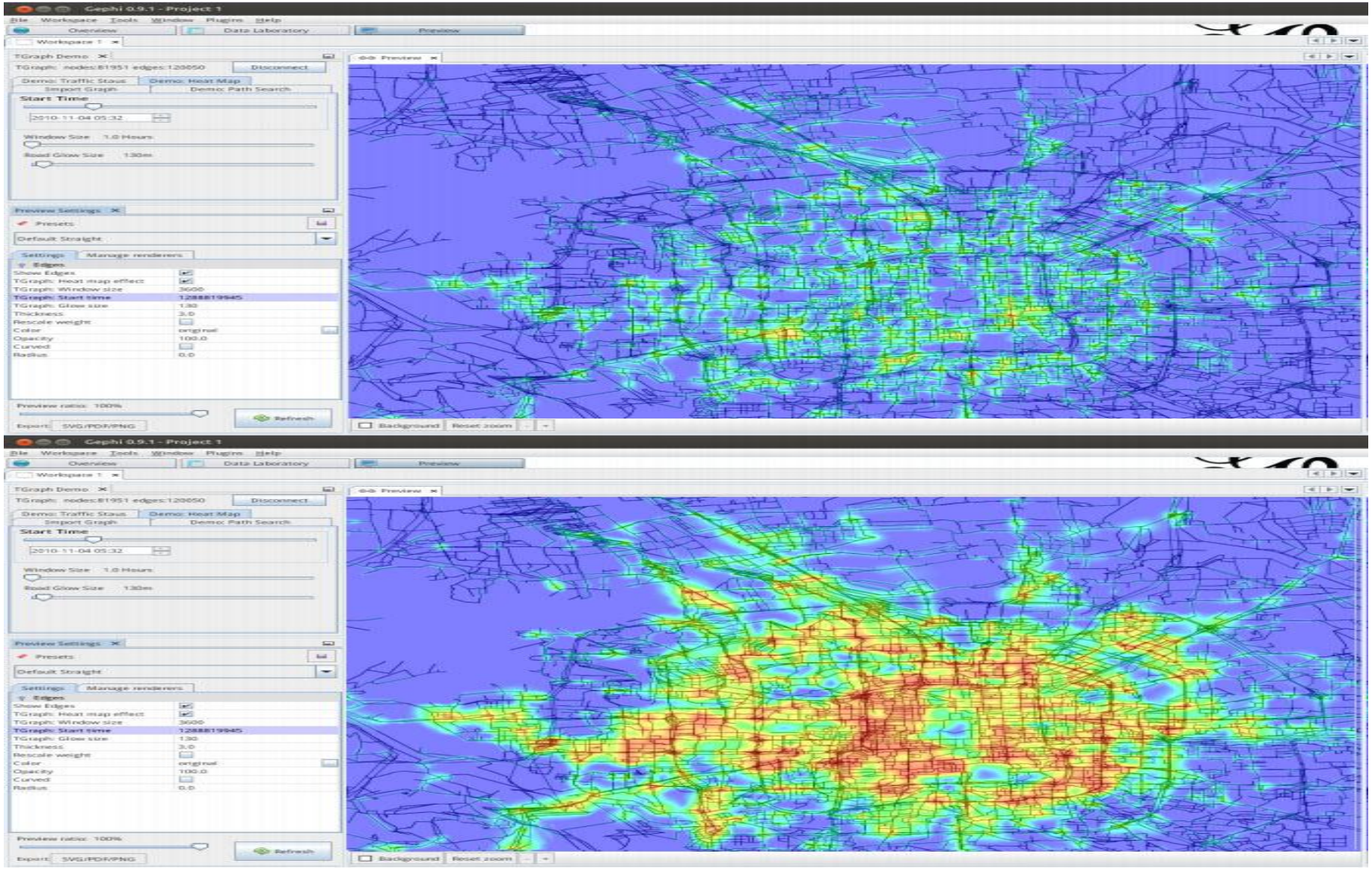


舆情大数据服务：已被国家信息安全管理中心采用，建立Ring微信公众号

研究工作：异质大数据获取、建模与深度分析，异构大数据存储与多模态计算，复杂系统敏捷定制，事件分析与预测。

核心算法： Weiren Yu, Charu C. Aggarwal, Shuai Ma, Haixun Wang: On Anomalous Hotspot Discovery in Graph Streams. ICDM 2013.

如二：TGraph时态图数据管理系统



Haixing Huang, Jinghe Song, Xuelian Lin, Shuai Ma, Jinpeng Huai, TGraph: A Temporal Graph Data Management System (demo**),** **CIKM 2016.**
Xuelian Lin, Shuai Ma, Han Zhang, Tianyu Wo, Jinpeng Huai: One-Pass Error Bounded Trajectory Simplification. **PVLDB** 10(7): 841-852 (2017)

Acknowledgements

Collaborators:

Charu Aggarwal, Sourav S Bhowmick, Yang Cao, Gao Cong, Liang Duan, Wenfei Fan, Kaiyu Feng, Haixing Huang, Renjun Hu, Jinpeng Huai, Jia Li, Jianxin Li, Xuelian Lin, Xudong Liu, Jinghe Song, Haixun Wang, Luoshu Wang, Tianyu Wo…

They are from:

北京航空航天大學 BEIHANG UNIVERSITY

THE UNIVERSITY of EDINBURGH

NANYANG TECHNOLOGICAL UNIVERSITY

THE OHIO STATE UNIVERSITY

IBM

Microsoft Research 微软亚洲研究院

Google Research

facebook

Homepage: http://mashuai.buaa.edu.cn

Email: mashuai@buaa.edu.cn

Address:

Room G1122,
New Main Building,
Beihang University
Beijing, China

Thanks!