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/

北京市大数据科学与脑机智能创新中心

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

- 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

数据科学与计算智能

新型计算技术与系统

认知机理与仿真

数据工程与脑机系统

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

大数据的研究与应用：取得重大突破

- 过去5年大数据的研究，已经产生了重大突破，并在部分领域取得良好的应用
  - 计算基础：大规模云计算、大规模深度学习
  - 感知处理的角度：大规模深度学习，imageNet
  - 知识组织与管理角度：大规模知识图谱

- 基于数据产生知识的问答系统与个人辅助系统
  - Watson DeepQA：智能搜索→知识引擎
  - Apple Siri & Wolfram Alpha

- 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

$$G=F(x)$$

算法

数据

John E Hopcroft Robert Tarjan（1986）

Stephen Cook（1982）

Donald Knuth（1974）

Leslie Valiant（2010）

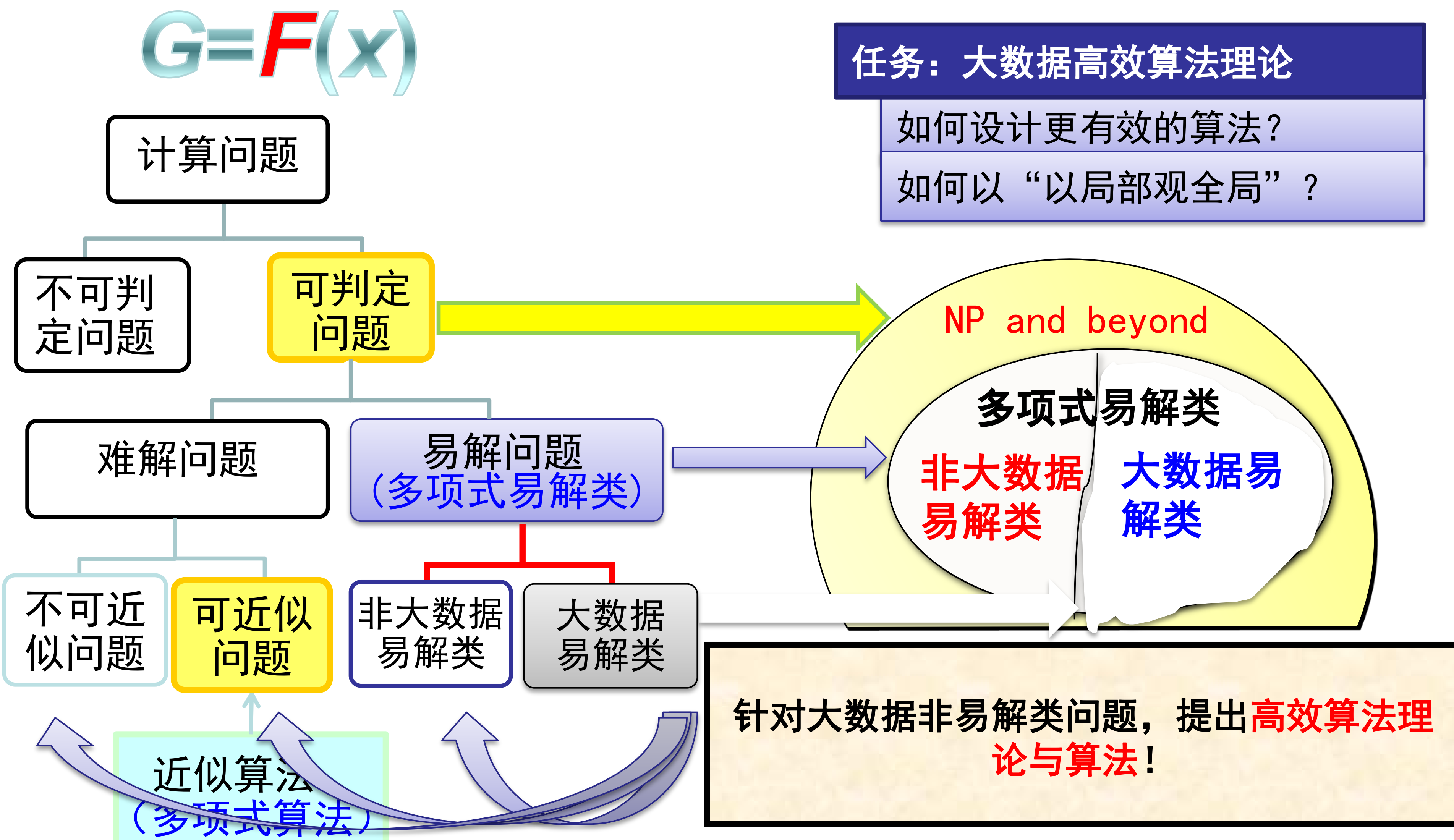
Manuel Blum（1995）

Juris Hartmanis，Richard Edwin Stearns（1993）

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

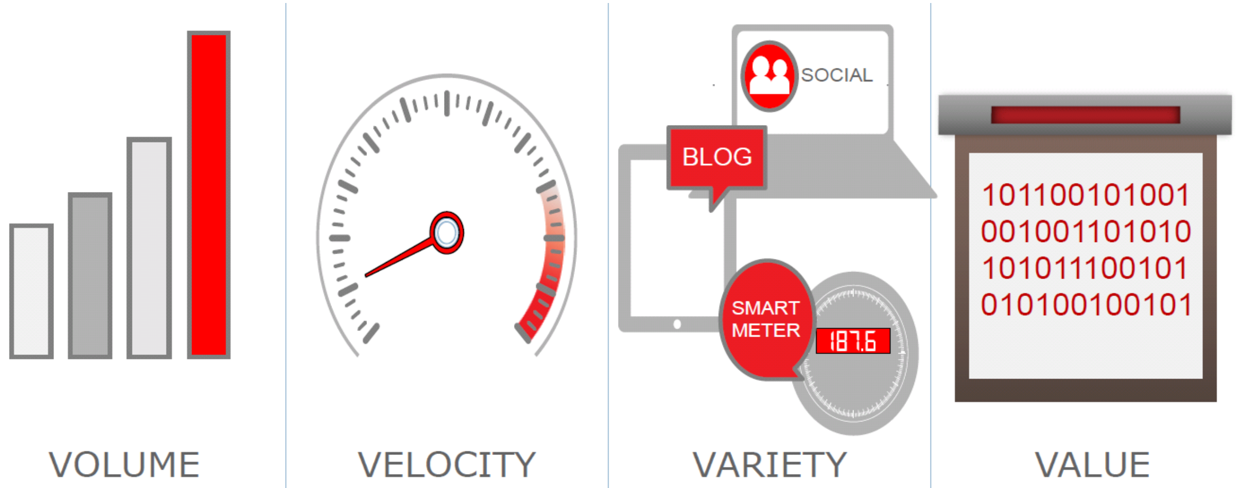
21世纪—大数据时代：计算复杂度与算法理论是否有新的理论问题和新方法？

回答“可计算”问题



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：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”？
NP and beyond
多项式易解类
非大数据易解类
大数据易解类
针对大数据非易解类问题，提出高效算法理论与算法！

Big Graph, e.g., Social Networks



Big volume: a balance between search efficiency and accuracy

Frequent changes: incorporate dynamic and temporal features

Noise & uncertainty: improve data quality, alleviate side effects

Query Techniques for Big Graph Search

$$R = Q(D)$$

Query Approximation Techniques

Main idea: For a class Q of queries with a high computational complexity, find another class Q' of queries that has a lower computational complexity without loss of quality or with a bounded loss of quality.

$$Q(D) \xrightarrow{\text{approximation}} Q'(D)$$

Challenge: balancing accuracy and computational complexity!

(1) E.g., Strong Simulation

Subgraph Isomorphism (NP-Complete)	$\xrightarrow{\text{approximation}}$	Strong Simulation ($O(n^3)$)

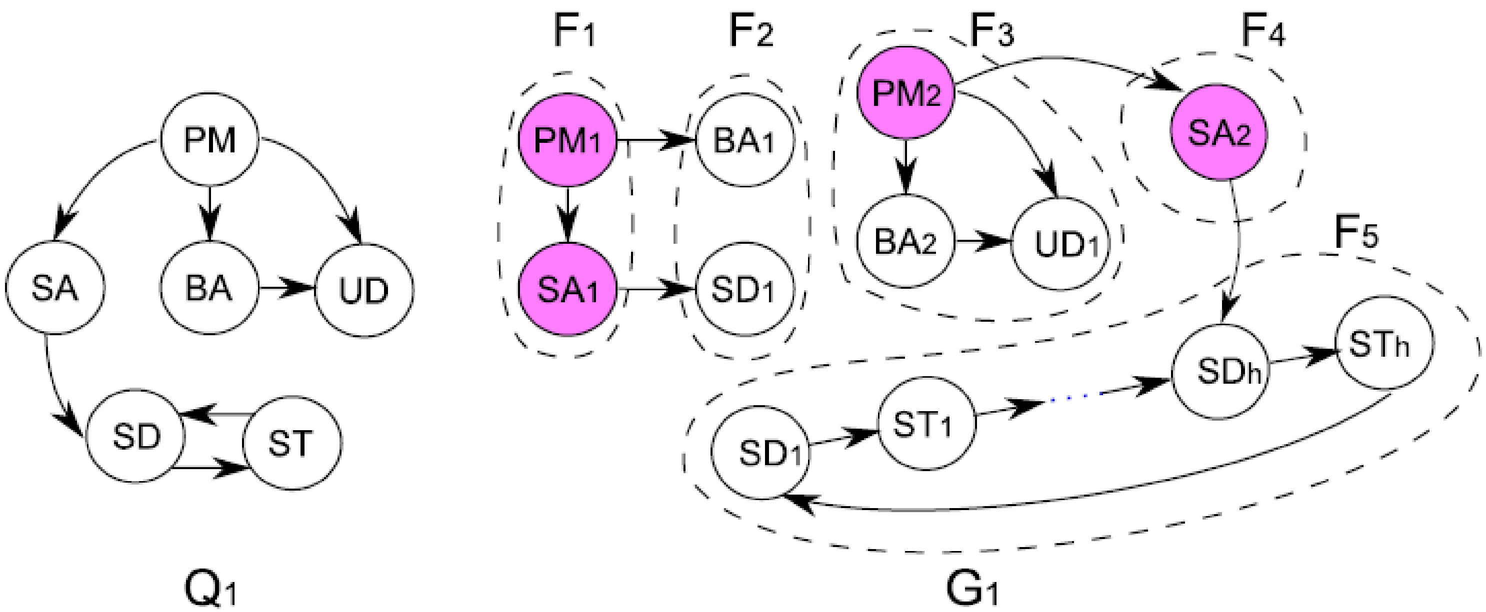
- Subgraph Isomorphism: Pattern graph Q, subgraph G_s of data graph G
  - Q matches G_s if there exists a bijective function f: $V_Q \rightarrow V_{Gs}$ such that
    - ✓ for each node u in Q, u and f(u) have the same label
    - ✓ An edge (u, u') in Q if and only if (f(u), f(u')) is an edge in G_s
  - Q matches G, via subgraph isomorphsim, if there is such a subraph Gs

Goodness : Keep exact structure topology between Q and G_s

Badness : NP-complete; may return exponential many matched subgraphs; In certain scenarios, too restrictive to find matches

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.**

(1) E.g., Strong Simulation



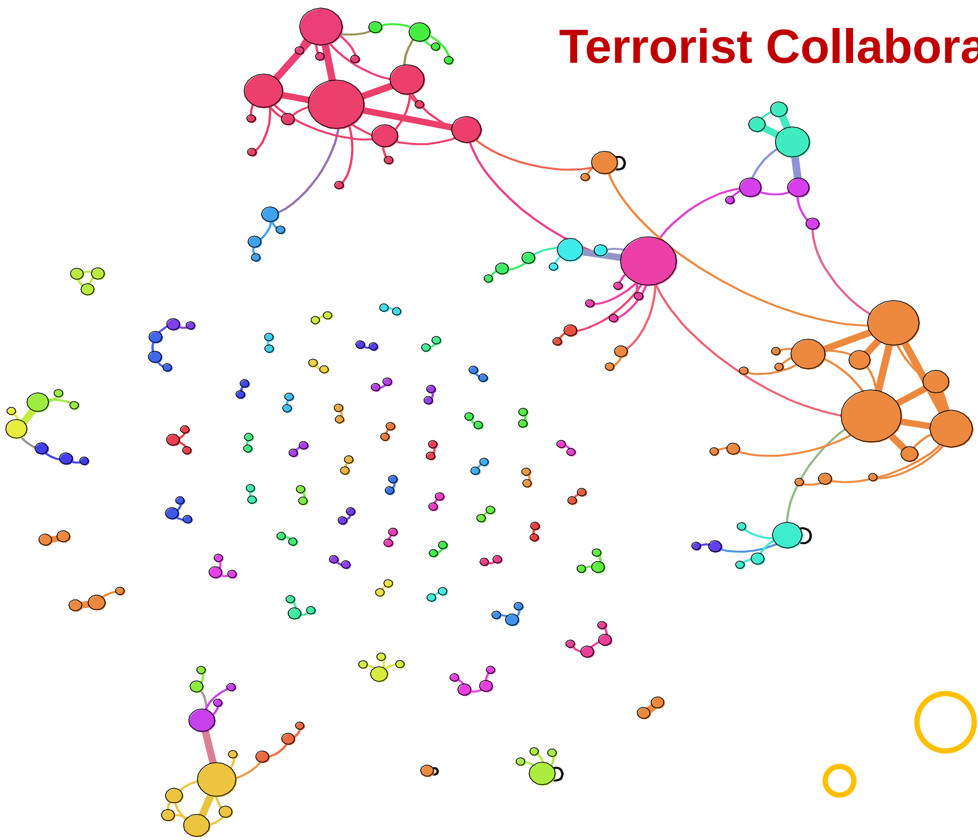
Set up a team to develop a new software product

Strong simulation returns F3, F4 and F5;
Subgraph isomorphism returns empty!

Subgraph isomorphism is too strict for emerging applications!

(1) E.g., Strong Simulation

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.

(1) E.g., Strong Simulation

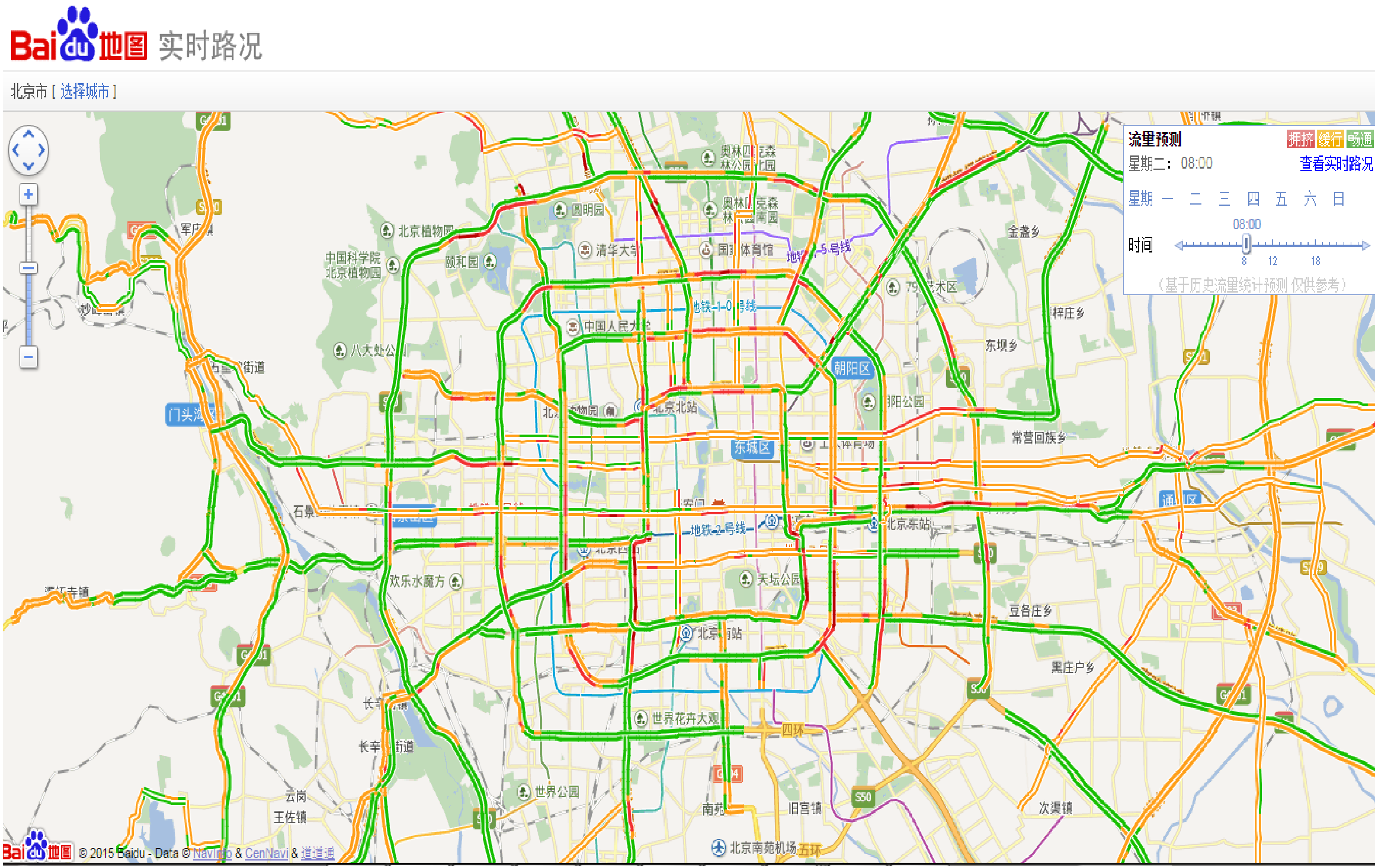
Subgraph Isomorphism (NP-Complete) › Strong Simulation › Dual Simulation › Graph Simulation ($O(n^3)$)

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
×	✓	×
×	×	×
✓	✓	×
✓	×	✓

Preserve 70-80% subgraph isomorphism & 100 times faster!

(2) E.g., Temporal Dense Subgraphs



(2) E.g., Temporal Dense Subgraphs

- Filter-and-Verification methods:

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

95% are filtered

- Data Driven Query Approximation methods:
  - Choose k (a small constant, e.g., 10 or 15)

10^4	10^5	10^6	…	10^8
k	k	k		k

- Experimental results (with the state of the art solution [Bogdanov et al. 2011])

	Accuracy	Efficiency
BEIJING DATA	100.28%	2980 times faster
SYNTHETIC DATA	99.84%	1,079 times faster

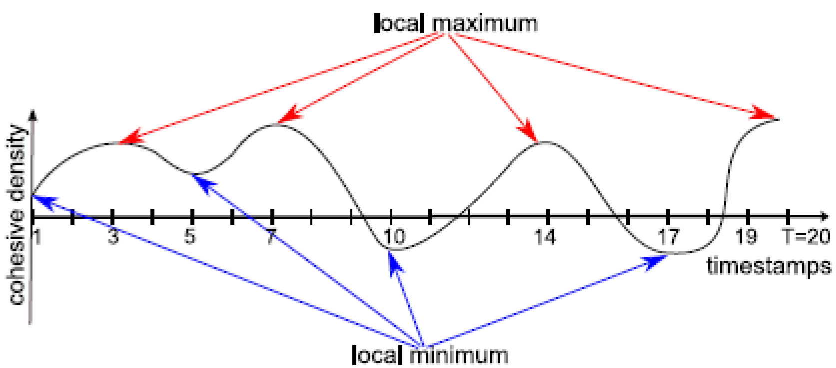
P. Bogdanov, M. Mongiov, and A. K. Singh. Mining heavy subgraphs in time-evolving networks. In ICDM, 2011.
Haixing Huang, Jinghe Song, Xuelian Lin, Shuai Ma, Jinpeng Huai, TGraph: A Temporal Graph Data Management System (demo), **CIKM 2016.**
Shuai Ma, Renjun Hu, Luoshu Wang, Xuelian Lin, Jinpeng Huai,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

(2) E.g., Temporal Dense Subgraphs




In **evolutionary** biology, **convergent evolution** is the process whereby organisms not closely related (not monophyletic), independently **evolve** similar traits as a result of having to adapt to similar environments or ecological niches.

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.* □

Data Techniques for Big Graph Search

$$R = Q(D)$$

Data Approximation Techniques

Main idea: For a class Q of queries on data D, transform D to smaller data D' that can be processed efficiently without loss of quality or with a bounded loss of quality.

$$Q(D) \xrightarrow{\text{approximation}} Q(D')$$

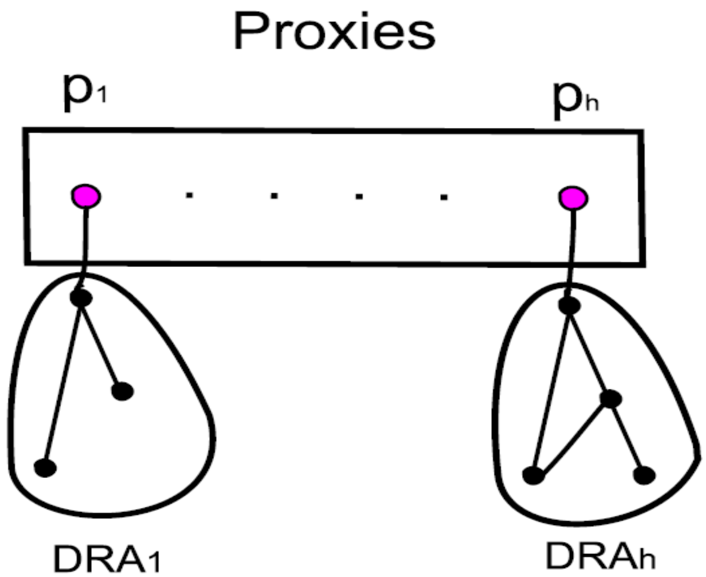
Pareto principle: for many events, roughly 80% of the effects come from 20% of the causes

$$D = \text{HARD}(D) + \cancel{\text{SOFT}(D)}$$

Challenge: **balancing accuracy and computational complexity!**

(1) E.g., Shortest Paths/Distances

- For weighted undirected graphs, we propose a notion of "proxies"
- Each proxy represents the nodes in its DRA (non-overlapping for all proxies)
- Proxies can be computed in O(n) time



Key property: Given nodes u,v in G, proxies u_p,v_p, then:

(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)

On Real-life road and social networks, graphs are reduced by 1/3!
A light-weight general data reduction technique for shortest paths/distances!

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)**.

(2) E.g., Network Link Prediction

Link Prediction

- A network with n nodes，$O(n^2)$ possible links
- CPU speeds: xGHz/s, and assume that a single machine cycle could deal with a node pair.

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

Most link prediction algorithms only predict a subset of the possible links, not all possible links, such as [Dashun et al. 2011, Chungmok et al. 2014].

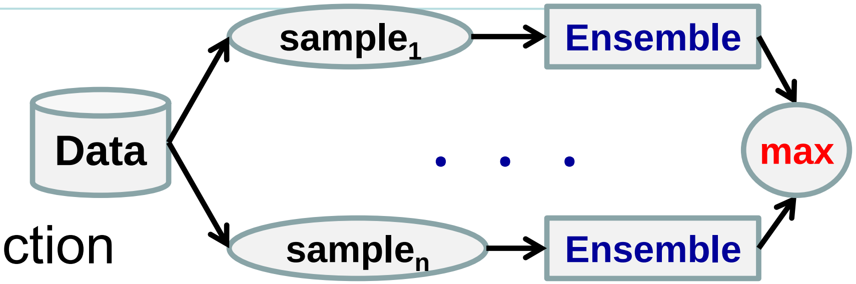
Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, Albert-László Barabási: Human mobility, social ties, and link prediction. **KDD 2011**.

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**.

(2) E.g., Network Link Prediction

Direct Non-negative Matrix Factorization

- Low efficiency
- The sparser the data, the worse the prediction



Framework

Data approximation technique (Ensemble Enabled Sampling)

- Sampling must assure a coverage on $O(n^2)$ possible links
- Link prediction characteristics (triangles)
- Ensemble: the predicted value of a link is the maximum among all ensembles

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

Small data	Accuracy	Big data	Efficiency
YouTube	+18%	Friendster	31 times faster
Wikipedia	+16%	Twitter	21 times faster

Improves both accuracy and efficiency!

Liang Duan, Charu Aggarwal, Shuai Ma, Renjun Hu, and Jinpeng Huai, Scaling up Link Prediction with Ensembles, **WSDM 2016 - Big Data Algorithms Session**.

Other Query and Data Techniques

- **Distributed algorithms:** $Q(D) \rightarrow Q(D_1)$, $Q(D_i)$, $Q(D_n)$
- **Incremental Computation:**

$$Q(D + \Delta) \xrightarrow{\text{Incremental computation}} Q(D) + Q(\Delta)$$

Known results

- **Data Compression:** $Q(D) \xrightarrow{\text{compression}} Q(D')$
- **Data Partition:** $Q(D) \xrightarrow{\text{partitioning}} Q(D_1) + \cdots + Q(D_n)$

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:

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

Email: mashuai@buaa.edu.cn

Address:

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

Thanks!