ArchSummit
全球架构师峰会 2017

主办方 Geekbang 极客邦科技 InfoQ

大规模异构网络数据融合

张宇韬
清华大学数据科学研究院

SPEAKER
INTRODUCE

张宇韬

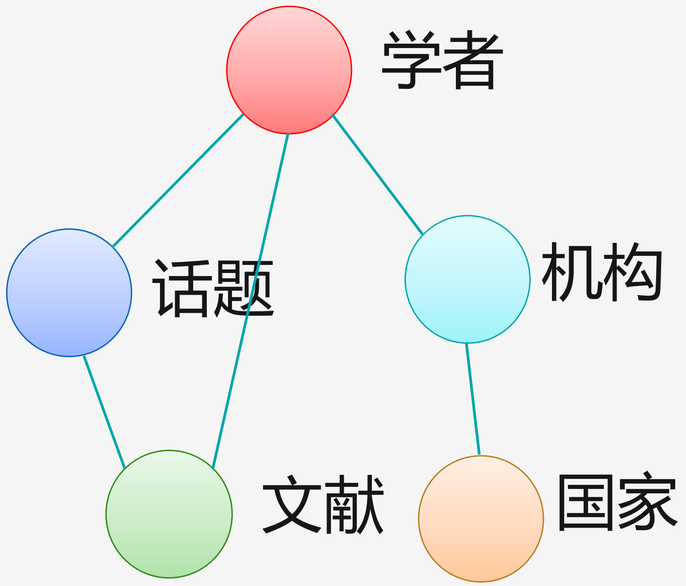
清华大学 科技大数据中心 首席研究员

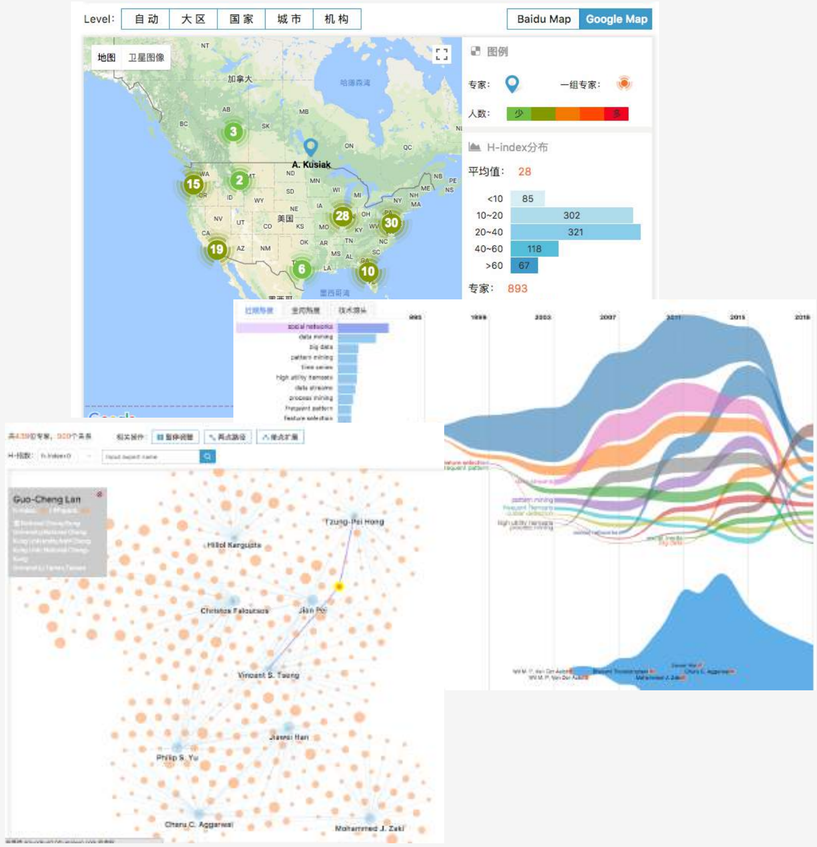
张宇韬，清华大学数据科学研究院科技大数据研究中心首席研究员。研究方向为异构数据融合及知识图谱构建。在KDD、CIKM、VAST等数据挖掘领域国际重要会议上发表多篇论文。作为技术负责人参与研发学术网络分析挖掘系统AMiner，集成上亿的学者、机构、科技文献、专利数据，提供针对科技数据的搜索及可视化分析功能，拥有数百万用户访问量。曾获得吴文俊人工智能科技进步一等奖，IJCAI数据竞赛第二名。

TABLE OF
CONTENTS 大纲

- 异构数据融合与知识图谱构建
- 异构数据融合中的机器学习方法
  - 度量学习
  - 表征学习
- 图嵌入学习与图卷积网络
- 案例：重名实体排岐
- 基于科技知识图谱的智能服务

AMiner科技知识图谱

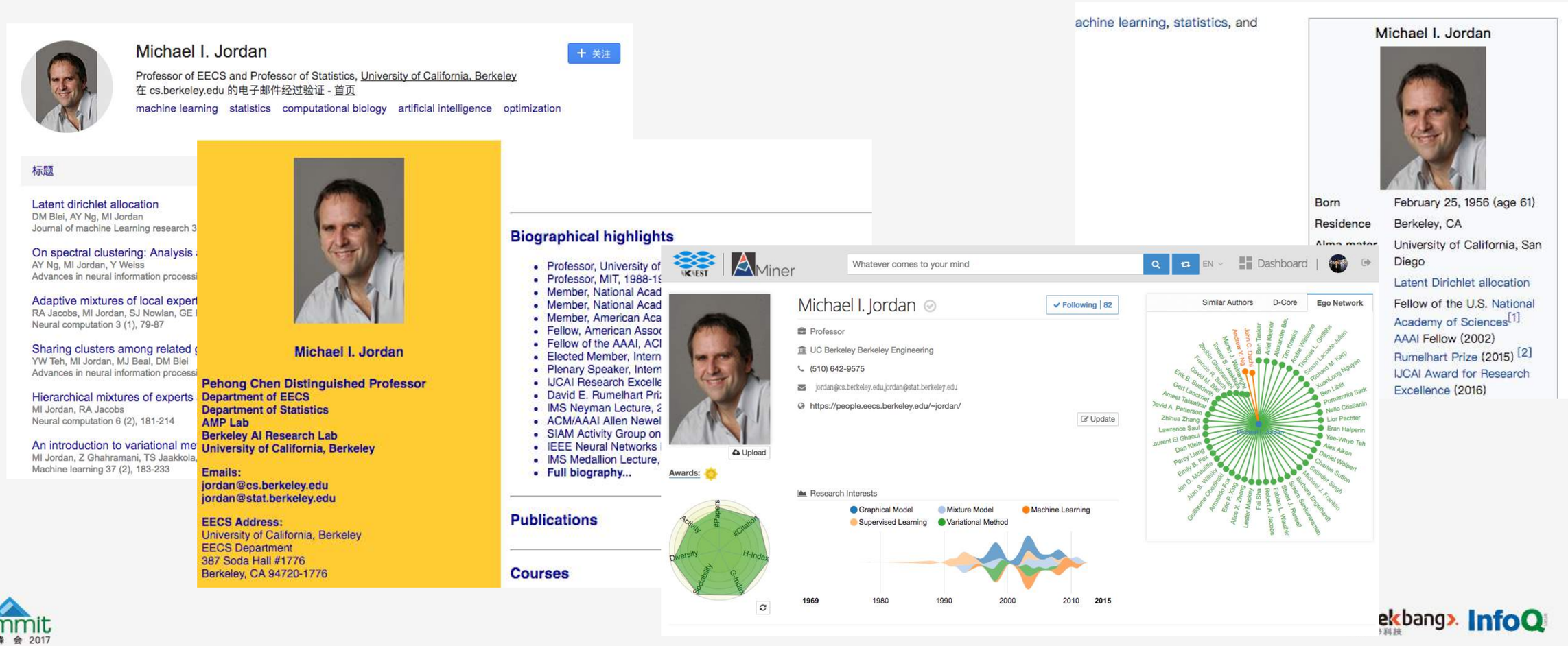




抽取 → 融合 → 挖掘

异构数据融合 (1)

- 相关信息散落在不同的数据源中



异构数据融合 (1)

- 相关信息散落在不同的数据源中

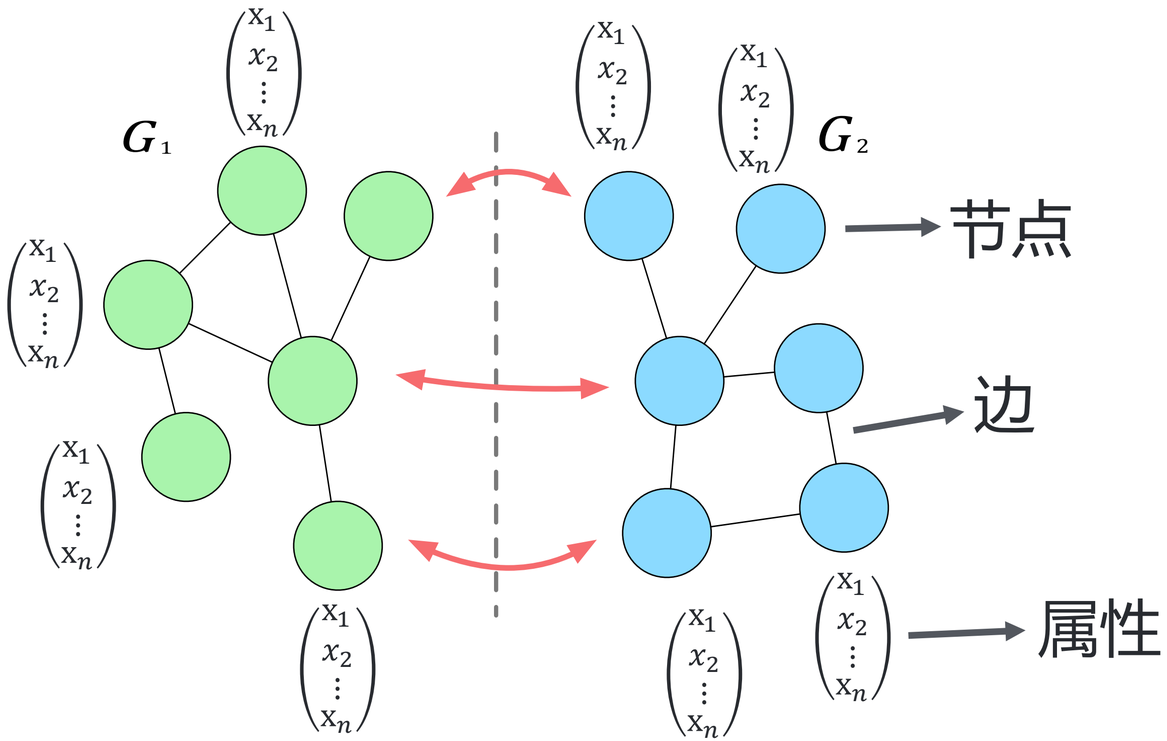
异构数据融合 (2)

- 不同数据源提供了对于同一实体不同视角的描述
- 如何回答：
  - 在NIPS上发表过一作论的应届博士生？
  - 中国2017获得发明专利最多的机构是那些？
  - 机器学习领域引用数最多的女性学者是谁？

异构数据融合 (3)

- 不同数据源间缺乏对于同一实体的统一识别
  - 自然语言的二义性
  - 同名异义(hyponym)
  - 异名同义(synonym)

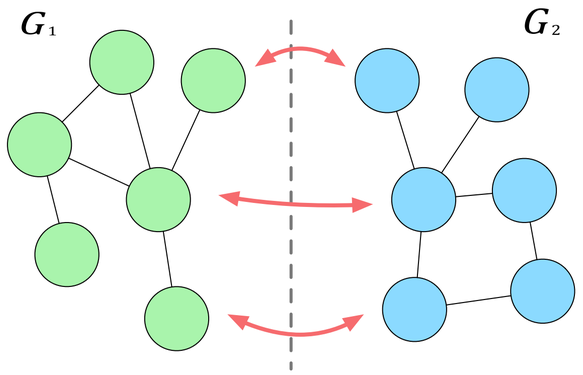
网络数据融合



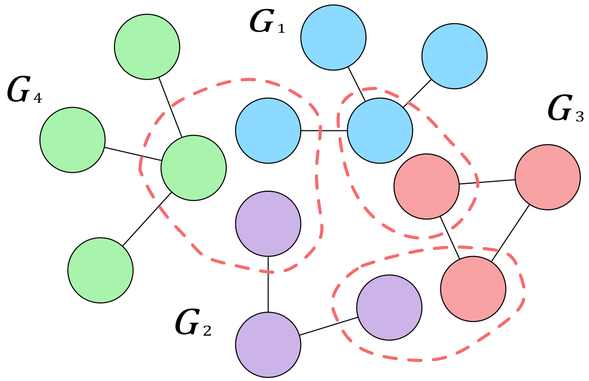
- 社交账户匹配、蛋白质网络(PPI)对齐、重名实体排岐、知识库实体链接、关系数据库集成、视觉图像匹配……

网络数据融合的三种场景

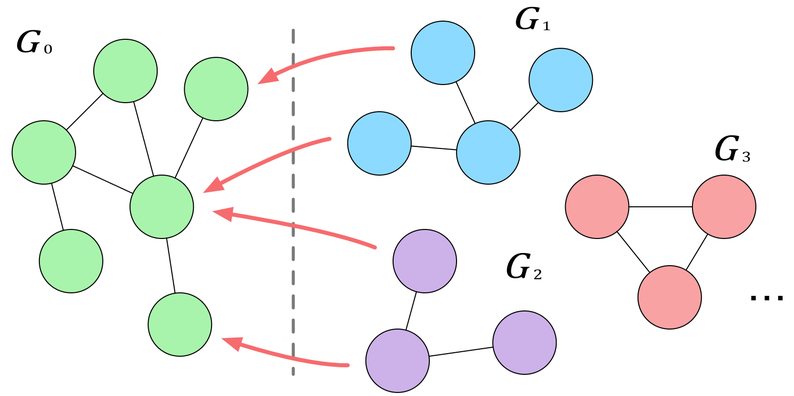
匹配问题 (Matching)



聚类问题 (Clustering)



分配问题 (Assignment)



核心问题：度量节点间的相似度

- 传统方法：定义相似度度量算法
  - Euclidean Distance, Jaccard Similarity, Edit Distance…
- 机器学习方法
  - 度量学习 (Metric Learning)
  - 表征学习 (Representation Learning)

度量学习 (Metric Learning)

- 输入：节点配对 (x_i^1, x_i^2), 相似度$y_i \in R$或标注$y_i \in \{0,1\}$
- 目标：学习一个函数$f(\cdot,\cdot)$, 使得$f(x_i^1, x_i^2) \sim y_i$
- 马氏距离(Mahalanobis Distances):

$$d_M(x_1, x_2) = \sqrt{(x_1 - x_2)^T M (x_1 - x_2)}$$

表征学习 (Representation Learning)

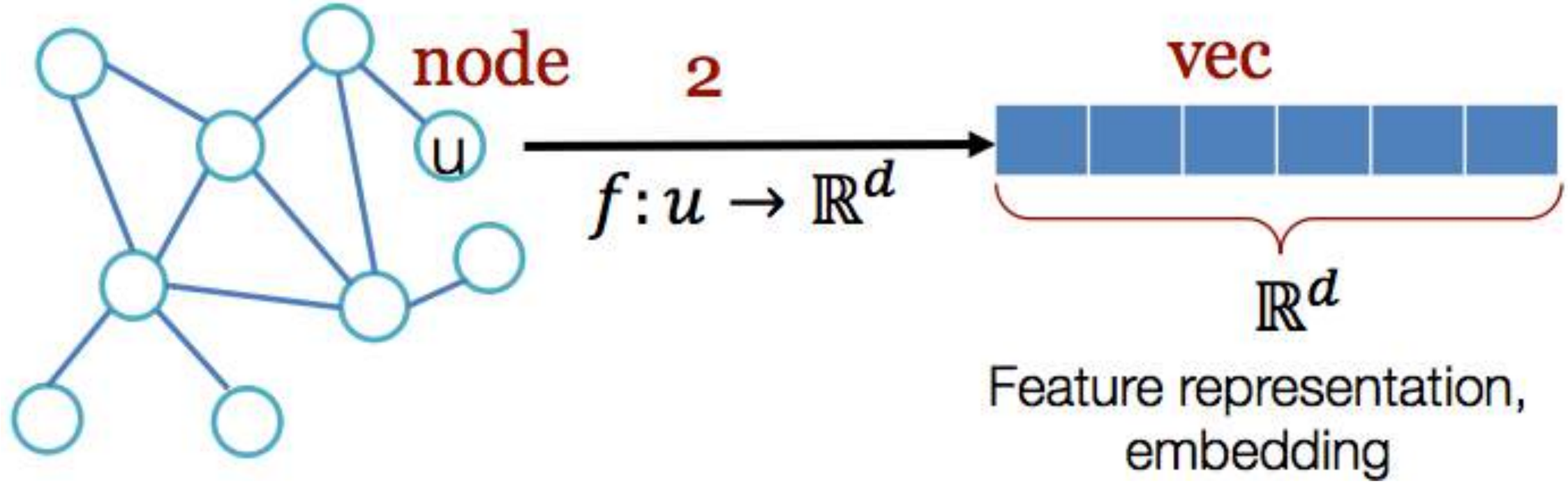
- Embedding： $\mathrm{f}: X \to Y$ 即一个映射函数，将数据从原特征空间X投影到表征空间Y
  - f为单射函数，即任意$\mathrm{x} \in X$有唯一的$\mathrm{y} \in Y$与之对应
  - Structure preserving性质，例如

$$d(x_1, x_2) < d(x_1, x_3) \to d(y_1, y_2) < d(y_1, y_3)$$

表征学习在实际应用中的优势

- 度量学习：
  - 直接预测两点间的相似度
- 表征学习
  - 将节点投影到低维连续空间，可容易计算相似度
- Embedding可作为实体的压缩表示
- Embedding可作为特征被用于其他数据挖掘任务

网络嵌入学习 (Network Embedding)



[Grover & Leskovec KDD'16]

图嵌入的难点

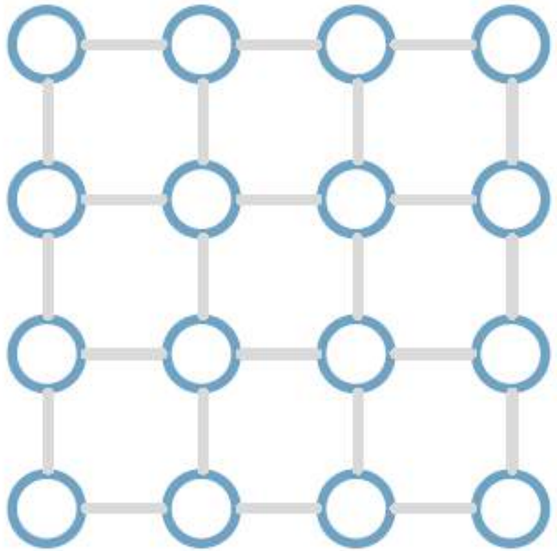
- 文本是线性序列 (word2vec, RNN)
- 图片是固定大小的二维网格 (CNN)
- 网络：
  - 节点间的连接是随机的
  - 有着复杂的子结构

图嵌入的难点

- 文本是线性序列 (word2vec, RNN)
- 图片是固定大小的二维网格 (CNN)
- 网络：
  - 节点间的连接是随机的
  - 有着复杂的子结构

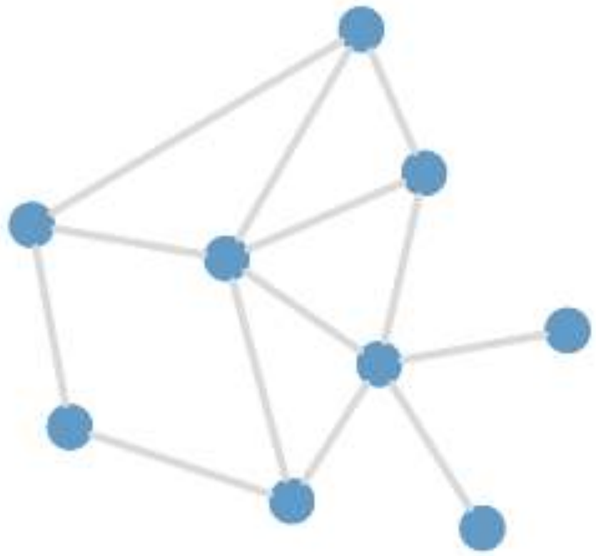
图嵌入的难点

- 文本是线性序列 (word2vec, RNN)
- 图片是固定大小的二维网格 (CNN)
- 网络：
  - 节点间的连接是随机的
  - 有着复杂的子结构



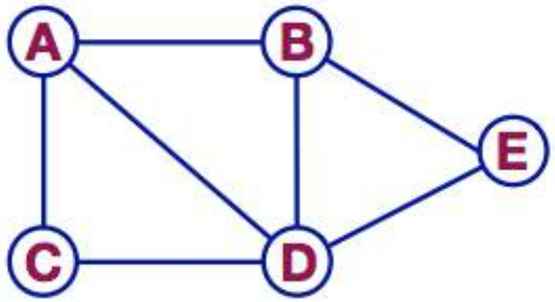
图嵌入的难点

- 文本是线性序列 (word2vec, RNN)
- 图片是固定大小的二维网格 (CNN)
- 网络：
  - 节点间的连接是随机的
  - 有着复杂的子结构



图嵌入的难点

Graph: $G = (\mathcal{V}, \mathcal{E})$



邻接矩阵 A

	A	B	C	D	E
A	0	1	1	1	0
B	1	0	0	1	1
C	1	0	0	1	0
D	1	1	1	0	1
E	0	1	0	1	0

特征向量 X

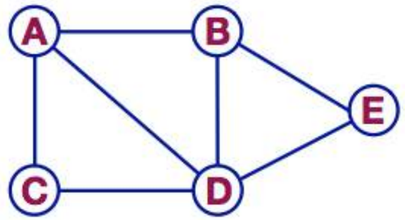
	Feat	
A	1	0
B	0	0
C	0	1
D	1	1
E	1	0

[Thomas Kipf]

图嵌入的难点

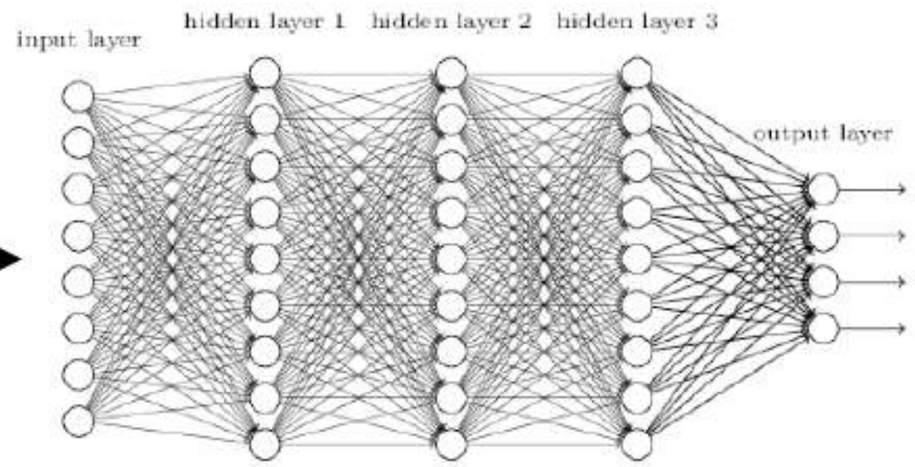
- 将邻接矩阵和特征向量直接作为输入构建三层神经网络

Graph: $G = (\mathcal{V}, \mathcal{E})$



邻接矩阵 A　　特征向量 X

	A	B	C	D	E	Feat	
A	0	1	1	1	0	1	0
B	1	0	0	1	1	0	0
C	1	0	0	1	0	0	1
D	1	1	1	0	1	1	1
E	0	1	0	1	0	1	0

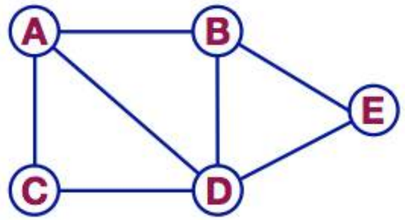


[Thomas Kipf]

图嵌入的难点

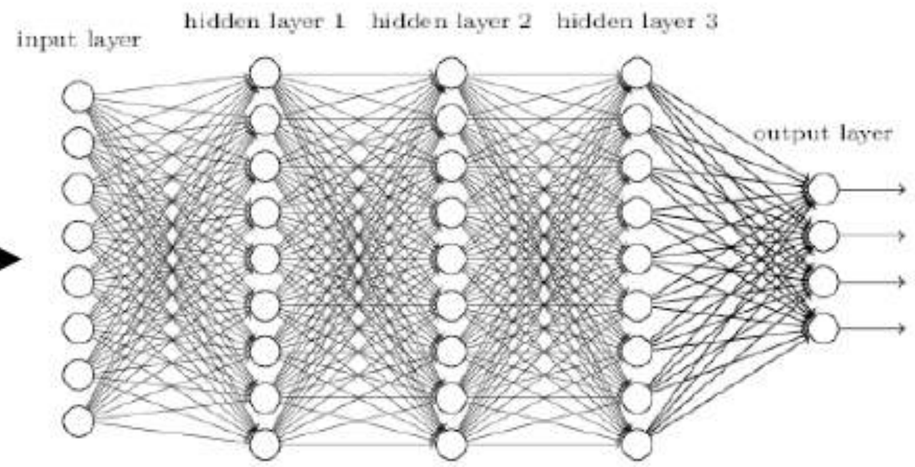
- 将邻接矩阵和特征向量直接作为输入构建三层神经网络

Graph: $G = (\mathcal{V}, \mathcal{E})$



邻接矩阵 A　　特征向量 X

	A	B	C	D	E	Feat	
A	0	1	1	1	0	1	0
B	1	0	0	1	1	0	0
C	1	0	0	1	0	0	1
D	1	1	1	0	1	1	1
E	0	1	0	1	0	1	0



[Thomas Kipf]

问题：

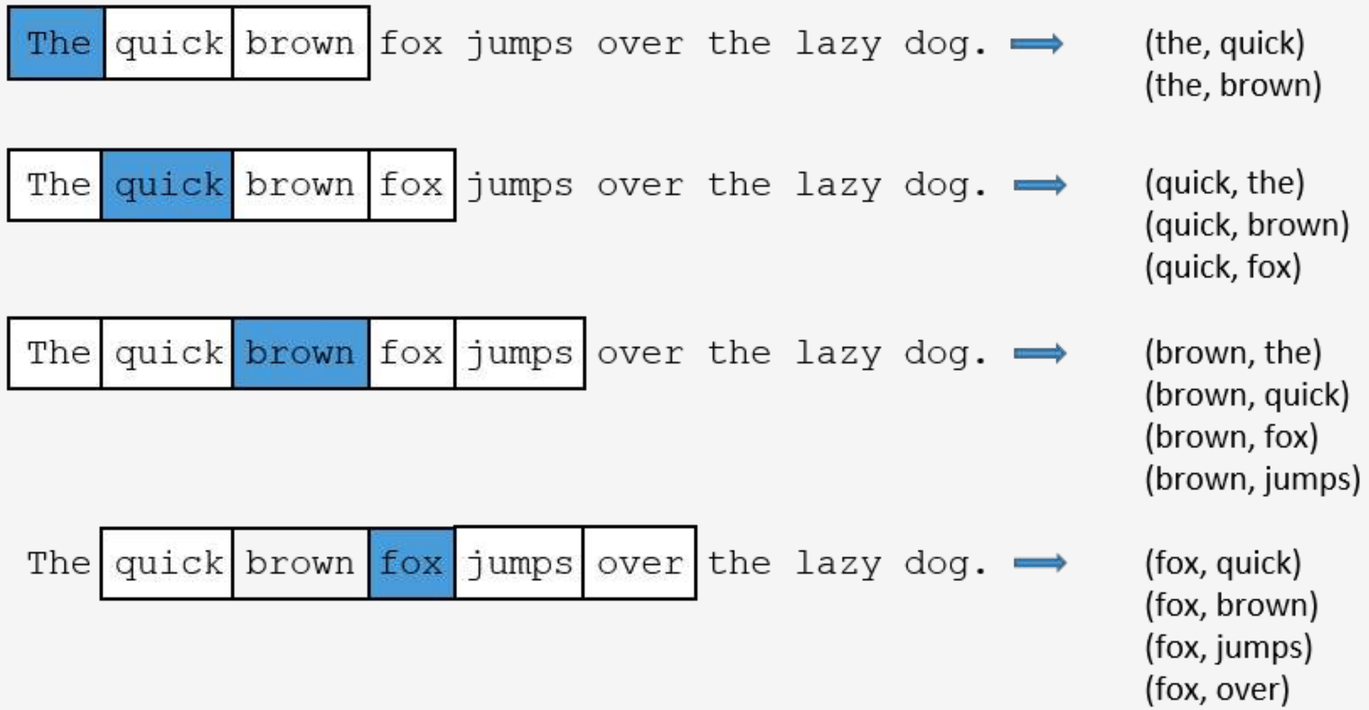
- 参数数量巨大
- 当图结构变化时需要重新训练

常用的图嵌入方法

- 矩阵分解 (Unsupervised, Transductive, Pure Structure)
  - 谱方法
- 图的“线性化” (Unsupervised, Transductive, Pure Structure)
  - 使用RandomWalk等方法将图转化为类似句子的点序列
  - DeepWalk, node2vec, LINE ….
- 图卷积网络 (Semi-Supervised, Inductive, Structure & Attributes)

图的“线性化”

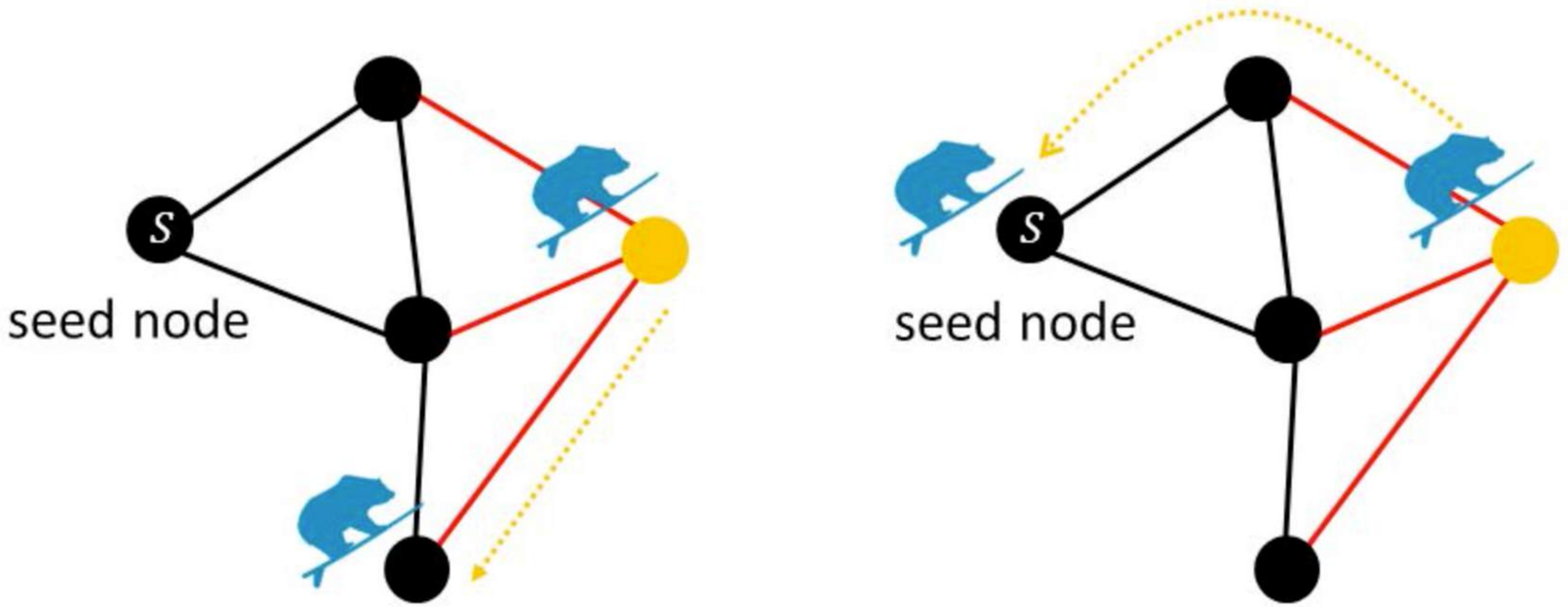
- 词向量模型 Word2vec



http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

图的“线性化”

- 根据随机游走（以概率p跳回源点）采样节点序列



[Shin et al. SIGMOD’15]

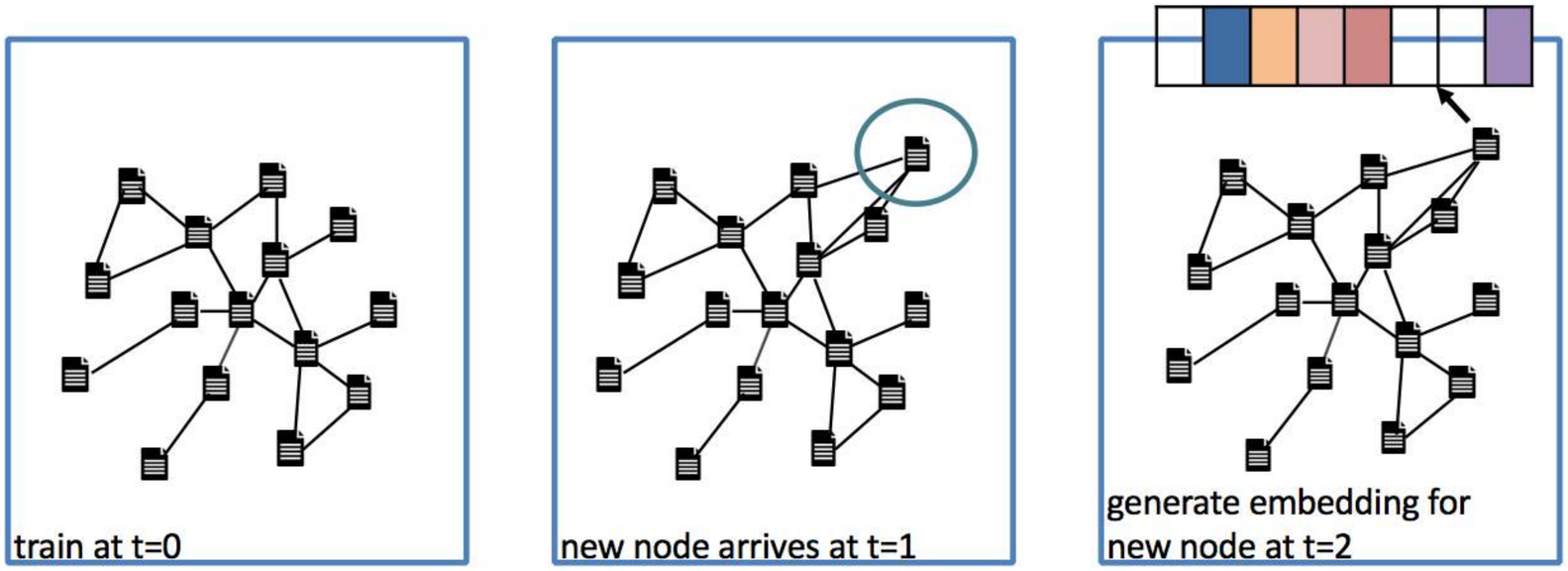
Transductive VS Inductive

- Transductive Learning:

  当网络变化时需要重新训练

- Inductive Learning:

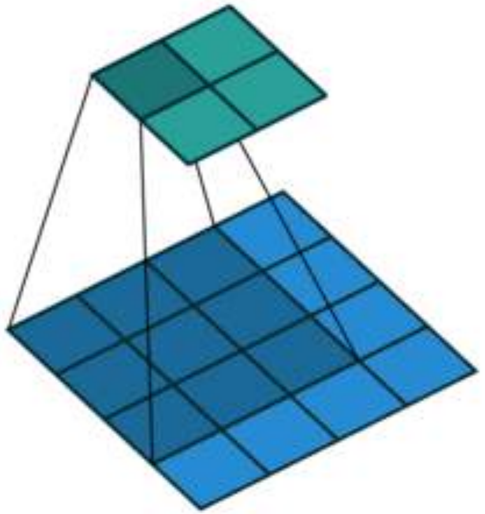
  可预测训练时未见过的新节点的embedding

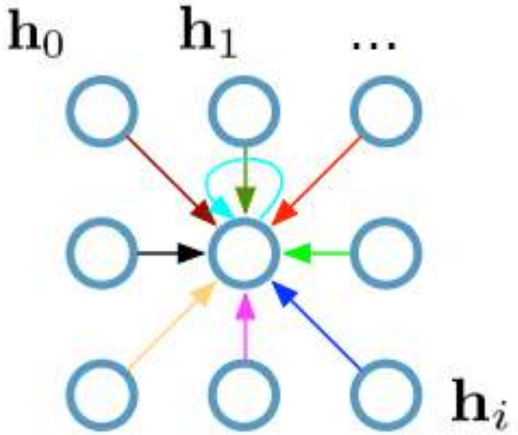


[Jure Leskovec]

二维网格上的卷积网络

- 3x3的卷积神经网



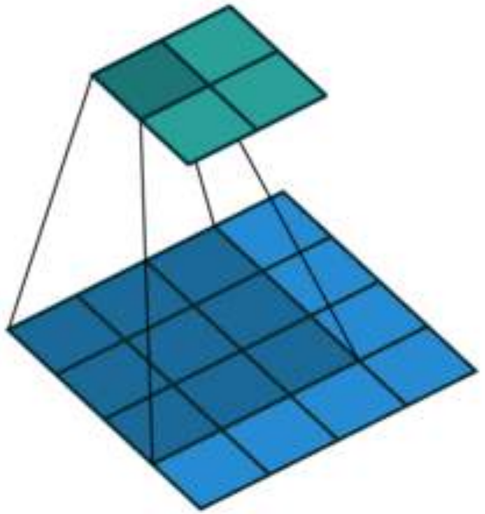


感受域
Receptive field

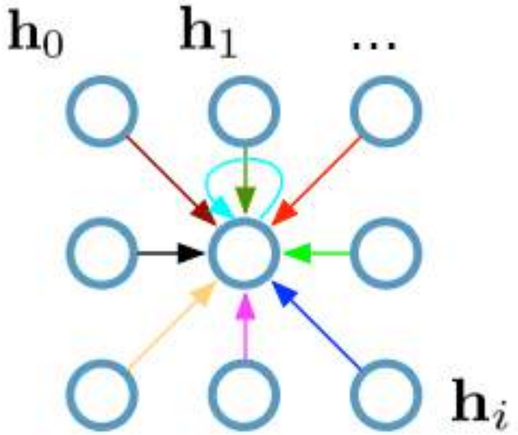
[Thomas Kipf]

二维网格上的卷积网络

- 3x3的卷积神经网



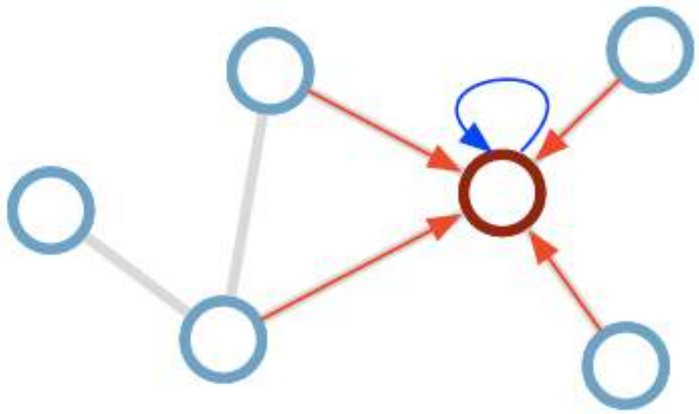
图上的消息传递



$$\mathbf{h}_4^{(l+1)} = \sigma\left(\mathbf{W}_0^{(l)}\mathbf{h}_0^{(l)} + \mathbf{W}_1^{(l)}\mathbf{h}_1^{(l)} + \cdots + \mathbf{W}_8^{(l)}\mathbf{h}_8^{(l)}\right)$$

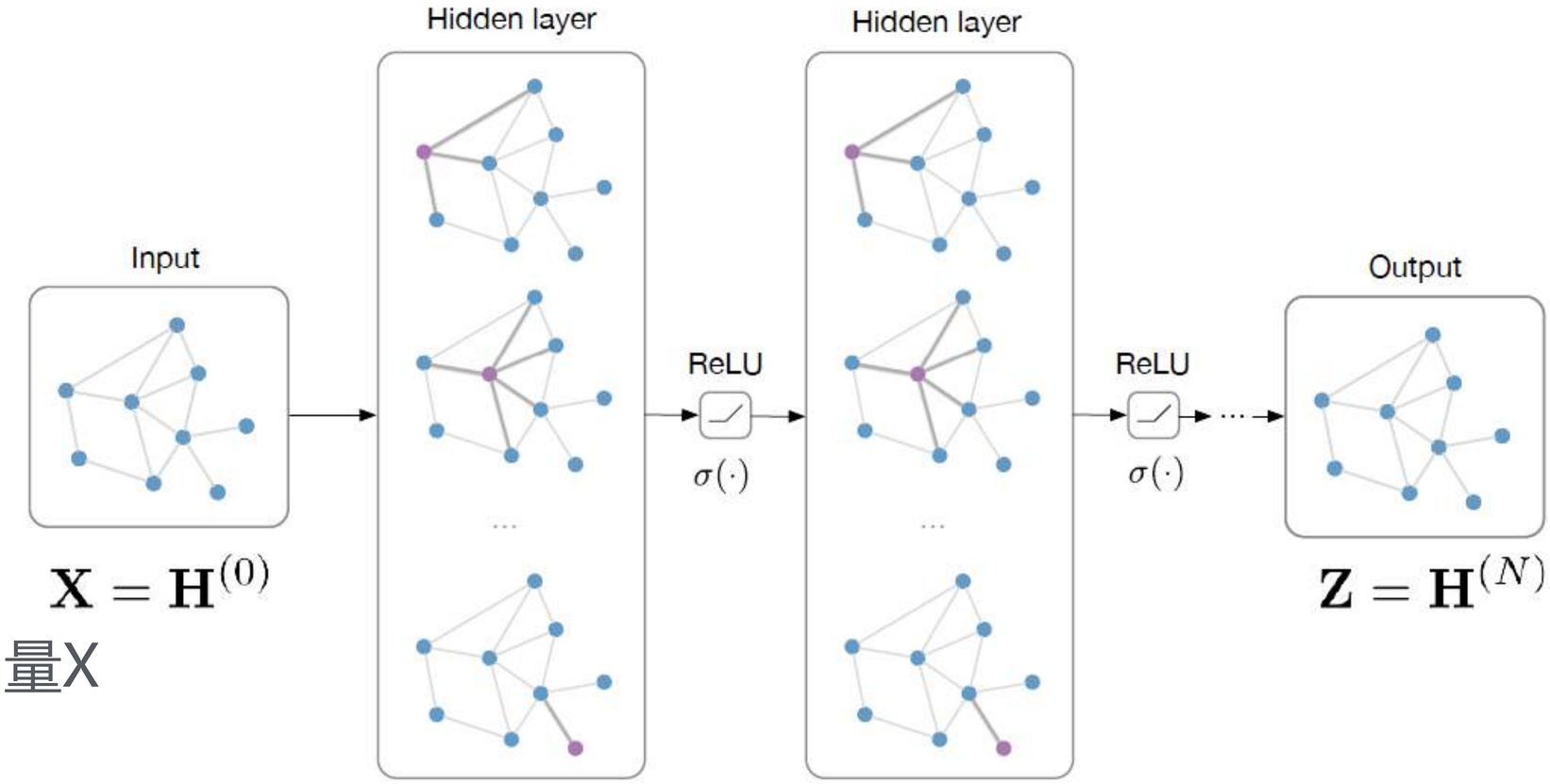
图卷积网络

- 拓展到广义的图上，定义图上的卷积操作



$$\mathbf{h}_i^{(l+1)} = \sigma\left(\mathbf{h}_i^{(l)}\mathbf{W}_0^{(l)} + \sum_{j\in\mathcal{N}_i}\frac{1}{c_{ij}}\mathbf{h}_j^{(l)}\mathbf{W}_1^{(l)}\right)$$

图卷积网络



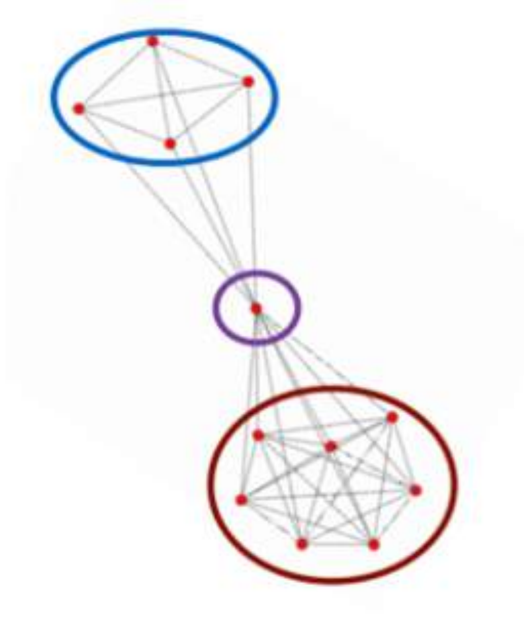
节点特征向量X

$$\mathbf{H}^{(l+1)} = \sigma\left(\hat{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right)$$

[Kipf & Welling, ICLR 2017]

与图同构问题的联系

- Weisfeiler-Lehman算法



Algorithm 1: WL-1 algorithm (Weisfeiler & Lehmann, 1968)

Input: Initial node coloring $(h_1^{(0)}, h_2^{(0)}, ..., h_N^{(0)})$

Output: Final node coloring $(h_1^{(T)}, h_2^{(T)}, ..., h_N^{(T)})$

$t \leftarrow 0$;

repeat

  for $v_i \in \mathcal{V}$ **do**

    $h_i^{(t+1)} \leftarrow \text{hash}\left(\sum_{j \in \mathcal{N}_i} h_j^{(t)}\right)$;

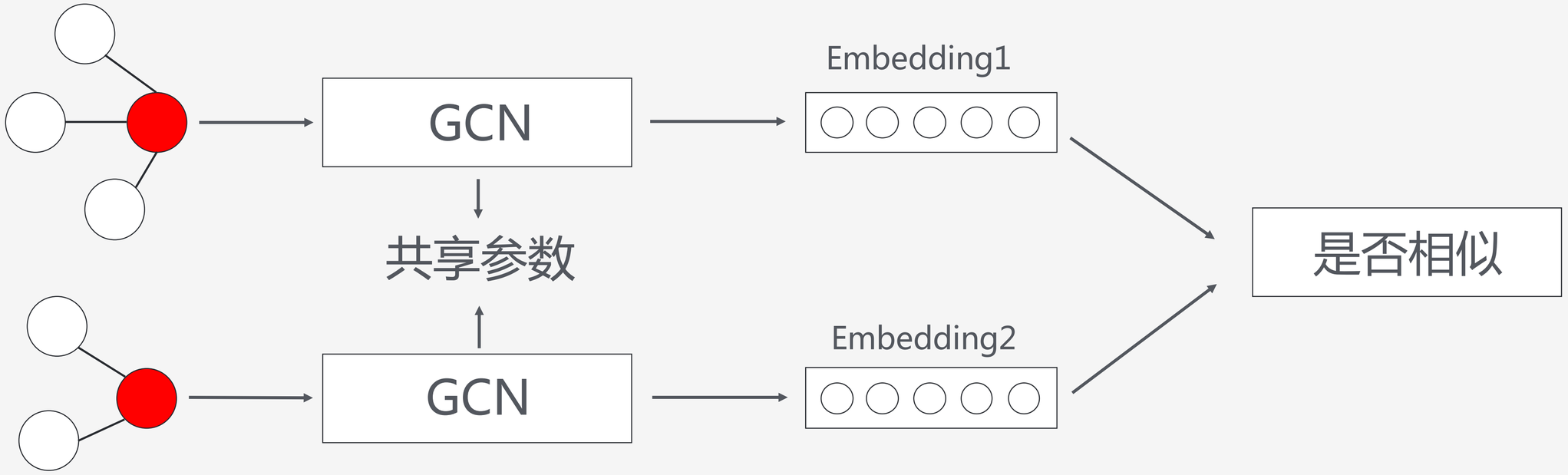
  $t \leftarrow t + 1$;

until *stable node coloring is reached*;

[Thomas Kipf]

网络训练

- 优化目标：表示同一实体节点的embedding尽可能相近
- Siamese网络



案例: 重名实体排岐

Name	Affiliation
Jing Zhang	Shanghai Jiao Tong Univ.
	Yunnan Univ.
	Tsinghua Univ.
	Alabama Univ.
	Univ. of California, Davis
	Carnegie Mellon University
	Henan Institute of Education

Jing Zhang

List of publications from the DBLP Bibliography Server - FAQ

Coauthor Index - Ask others: ACM DL - ACM Guide - CiteSeer - CSB - Google

2007

83 EE Jing Zhang, Guizhong Liu: Hyperspectral images lossless compression by a novel three-dimensional wavelet coding

82 EE Jing Peng, Dongqing Yang, Changjie Tang, Jing Zhang, Jianjun Hu: CACS: A Novel Classification Algorithm Base

81 EE Jing Zhang, Xi Chen, Ming Li: Computing Exact p-Value for Structured Motif. CPM 2007: 162-172

80 EE Jing Zhang, Jie Tang, Juan-Zi Li: Expert Finding in a Social Network. DASFAA 2007: 1066-1069

79 EE Guojun Chen, Jing Zhang, Xiaoli Xu, Yuan Yin: Real-Time Visualization of Tire Tracks in Dynamic Terrain with LO

78 EE Jing Zhang, Hai Huang: Federate Job Mapping Strategy in Grid-Based Virtual Wargame Collaborative Environmer

77 EE Maria Wimmer, Michael Goul, Jing Zhang: Minitrack: E-Government Information and Knowledge Management. H

76 EE Kai Kang, Jing Zhang, Baoshan Xu: Optimizing the Selection of Partners in Collaborative Operation Networks. IC

75 EE Lingshuang Shao, Jing Zhang, Yong Wei, Junfeng Zhao, Bing Xie, Hong Mei: Personalized QoS Prediction forWe

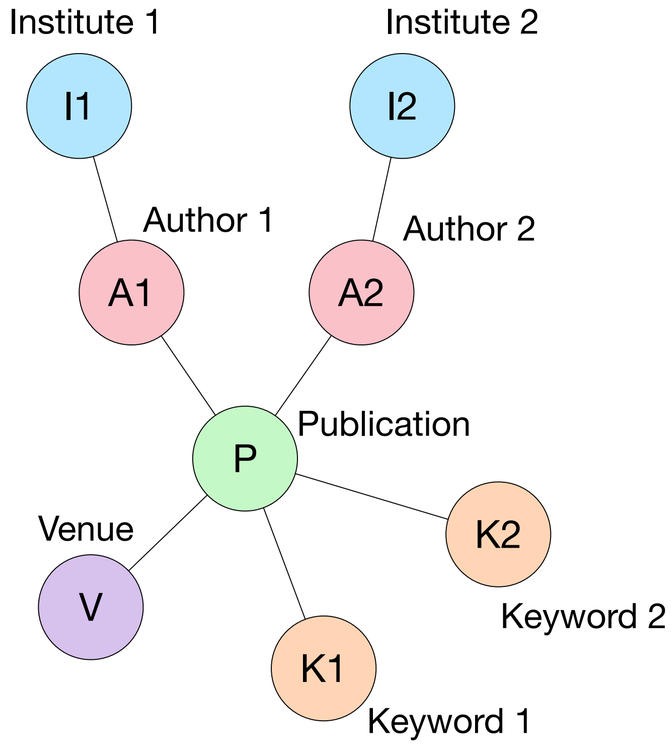
74 EE Benyong Liu, Jing Zhang, Xiaowei Chen: Adaptive Training of a Kernel-Based Representative and Discriminative

73 EE Jilong Wang, Jing Zhang: Federation Based Solution for Peer-to-Peer Network Management. International Confer

72 EE Jing Zhang, Fanhuai Shi, Jianhua Wang, Yuncai Liu: 3D Motion Segmentation from Straight-Line Optical Flow. M

问题形式化

- 每篇文献可表示成一个异构子图
  - 包含文献、作者、机构、会议、关键词等节点
- 需要匹配同类型且表示同一实体的节点
  - 关键在于度量节点相似度



两个应用场景

- 知识图谱初始建立
  - 根据Embedding将表示同一实体的节点聚类在一起 (Clustering)
- 增量数据对齐将Embedding扩展到新来的子图 (Inductive Learning)
  - 将新的节点分配给知识库中的已有节点 (Assignment)

知识图谱初始建立

- 问题一：缺乏人工标注训练样本
- 解决方案：弱监督学习
  - 基于规则生成正例
    - 例如：将同名且名字不常见的作者认为是同一个人
  - 随机采样负例

知识图谱初始建立

- 问题二：聚类个数未知
- 解决方案：使用贝叶斯信息准则(BIC)估计聚类个数

增量数据更新与维护

- 根据模型预测新节点的embedding
- 将新进入的子图中的节点与知识图谱对齐
  - 可能需要新建节点
- 构建二分类模型
  - 输入： 子图节点embedding，知识图谱中候选节点的embedding
  - 输出： 是否为同一实体

增量数据更新与维护

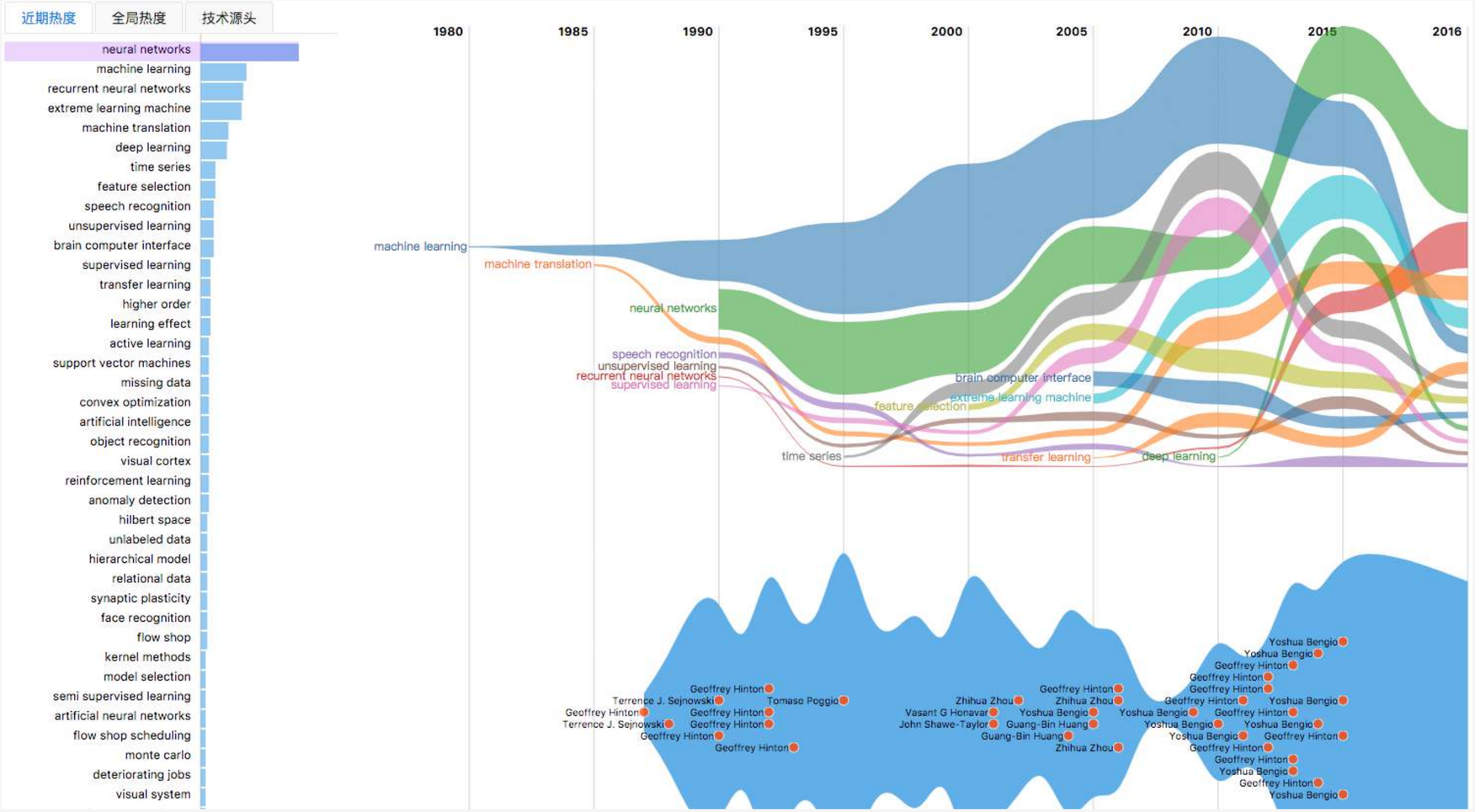
- 候选配对过滤 (Blocking)
  - 使用粗略的快速算法（例如基于规则），对候选配对空间进行过滤
  - 降低候选匹配个数，避免O(N^2)次相似度计算

TABLE OF

CONTENTS 大纲

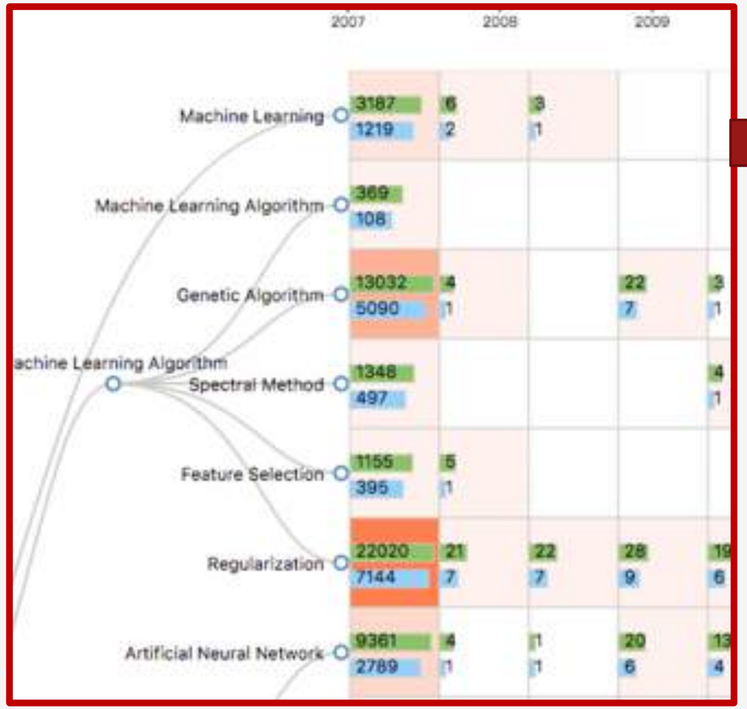
- 异构数据融合与知识图谱构建
- 异构数据融合中的机器学习方法
  - 度量学习
  - 表征学习
- 图嵌入学习与图卷积网络
- 案例：重名实体排岐
- 基于科技知识图谱的智能服务

技术趋势分析


近期热度
全局热度
技术源头
neural networks
machine learning
recurrent neural networks
extreme learning machine
machine translation
deep learning
time series
feature selection
speech recognition
unsupervised learning
brain computer interface
supervised learning
transfer learning
higher order
learning effect
active learning
support vector machines
missing data
convex optimization
artificial intelligence
object recognition
visual cortex
reinforcement learning
anomaly detection
hilbert space
unlabeled data
hierarchical model
relational data
synaptic plasticity
face recognition
flow shop
kernel methods
model selection
semi supervised learning
artificial neural networks
flow shop scheduling
monte carlo
deteriorating jobs
visual system
1980
1985
1990
1995
2000
2005
2010
2015
2016
machine learning
machine translation
neural networks
speech recognition
unsupervised learning
recurrent neural networks
supervised learning
feature selection
brain computer interface
extreme learning machine
time series
transfer learning
deep learning
Geoffrey Hinton
Terrence J. Sejnowski
Tomaso Poggio
Zhihua Zhou
Vasant G Honavar
John Shawe-Taylor
Guang-Bin Huang
Yoshua Bengio

基于地理位置的专家发现

学科交叉创新分析



THANK YOU

如有需求，欢迎至 [讲师交流会议室] 与我们的讲师进一步交流

ArchSummit
全球架构师峰会 2017