图神经网络在交叉学科领域的应用研究

魏忠钰 副教授

复旦大学 大数据学院

2020年7月3日
狗熊会

http://www.sdspeople.fudan.edu.cn/zywei/

网络(图)，无处不在

- **金融市场中的实体网络**
  - 公司法人、自然人
    - 股票持有关系
- **社交网络平台中的用户关系网络**
  - 用户
    - 互相关注、相关转发
- **生物医学领域的实体网络**
  - 基因，蛋白质
    - 实体上下位信息，功能表达的聚合
- **新闻文本文档中的词语网络**
  - 词，句子
    - 语义关系

网络中任务

- **节点表示学习**
  - 如何更好的给一个节点学习一个向量表示，使相近节点的表示相似
- **节点分类**
  - 用户性别分类、蛋白质功能分类等
- **关系分类**
  - 社会网络中的好友关系分类、基因关系分类等
- **社区发现**
  - 社会网络中的群体发现、基因网络中的聚合表达等

图神经网络的基本部件

- **给定输入的基本图表示 $G = (V, E)$**
  - V 是一个节点结合， E 作为节点之间的关系

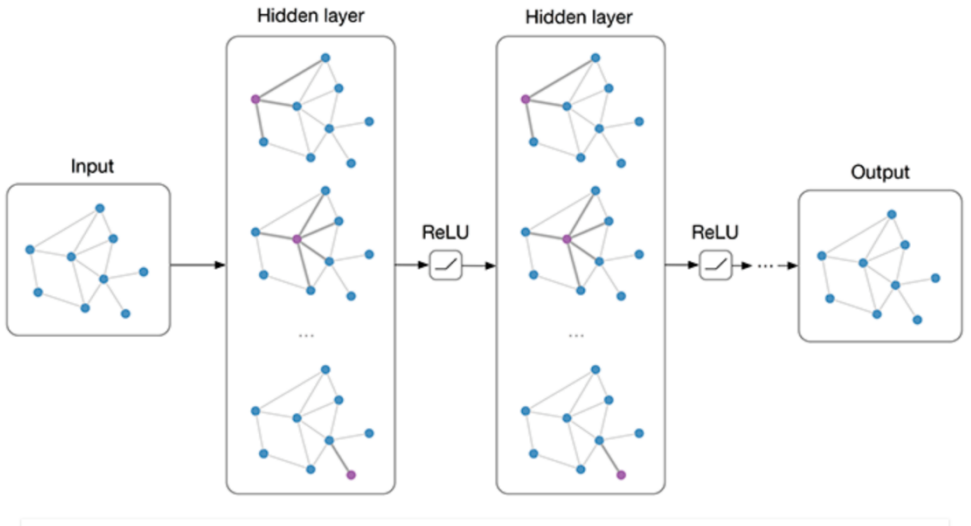
- **节点表示更新公式**

$$f(H^{(l)}, A) = \sigma\left(AH^{(l)}W^{(l)}\right)$$

  - 节点初始化：随机初始化或者使用手工特征
  - 邻接矩阵构建：记录节点之间的关系
  - 节点更新：邻居节点表示的加权平均

- **网络输出**

$$\mathbf{O} = G(\mathbf{H}, \mathbf{X}_N)$$



Kipf and Welling, Semi-Supervised Classification with Graph Convolutional Networks, ICLR 2017

图神经网络的训练

- **无监督模式**
  - 利用图结构特征，使节点的表示尽量符合图中的结构信息
  - 如：邻居节点的表示更相近

$$E(H) = \sum_{i,j} w_{ij} \left\| \frac{h_i}{\sqrt{d_i}} - \frac{h_j}{\sqrt{d_j}} \right\|_2^2$$

- **有监督模式**

$$loss = \sum_{i=1}^{p} (\mathbf{t}_i - \mathbf{o}_i)$$

  - t 为 节点的目标标签，o为网络输出标签，p为节点个数
- **半监督模式**
  - 使用部分节点的标签信息

Kipf and Welling, Semi-Supervised Classification with Graph Convolutional Networks, ICLR 2017

目录

- **图神经网络的背景介绍**
- **计算金融**
  - 面向股票预测的金融实体表示学习 （CIKM’ 2018）
  - 引入外部新闻资讯的金融事件预测 (CIKM’ 2019)
- **量化政治**
  - 结合国会议员关系的议案投票结果预测 (IJCAI’ 2020)
- **社会媒体用户画像**
  - 联合用户言论和关系的用户画像 (在研项目)
- **总结**

合作学生

陈滢美
2019届硕士研究生

杨依莹
2019届硕士研究生

杨雨樵
2019届本科生

宁上毅
2018级硕士研究生

林晓强
2020届本科生

林耿
2017级本科生

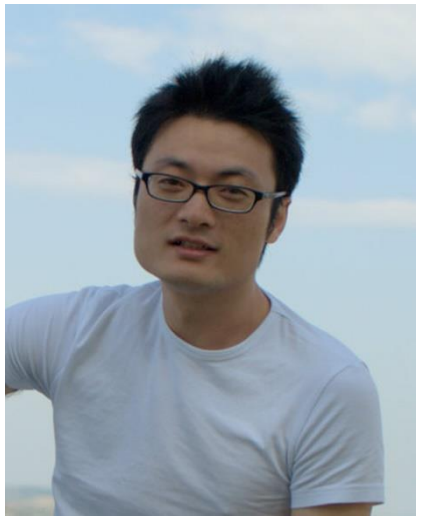
合作学者

蒋昌建
复旦大学
国务学院 教授

吴力波
复旦大学
经济学院 教授

黄萱菁
复旦大学
计算机学院 教授

周葆华
复旦大学
新闻学院 教授



黄增峰
复旦大学
大数据学院 研究员

金融市场中的网络

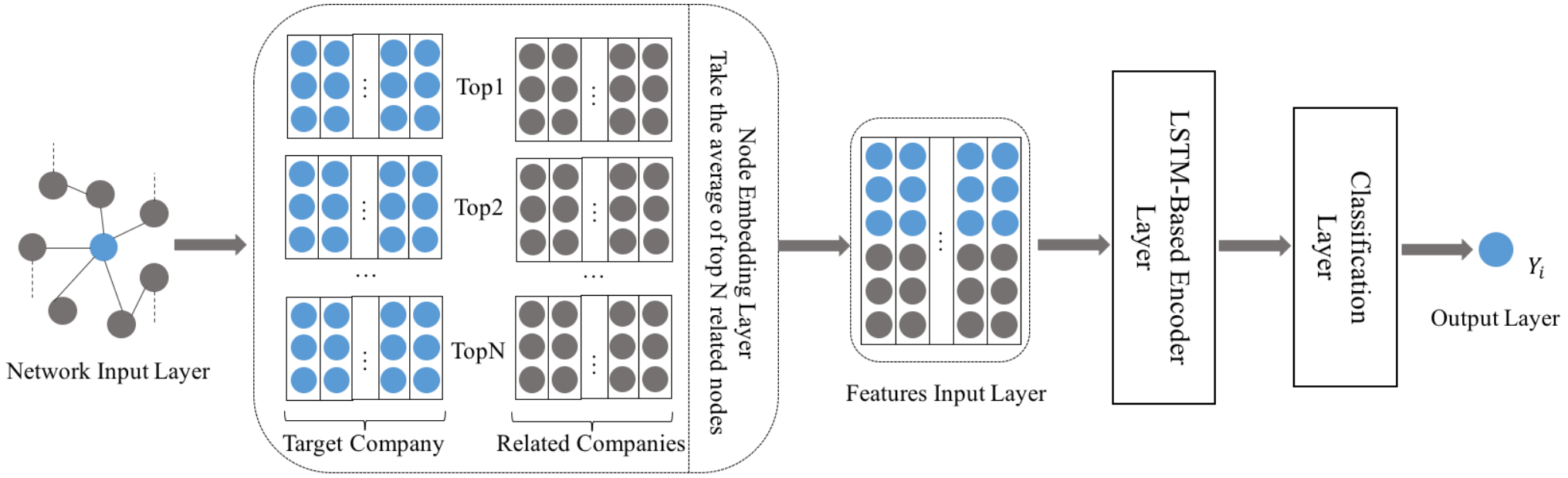
- **金融市场中的“单位”是相互依存的，彼此相互关联，形成相应的网络**
- **金融实体：**
  - 公司、机构、金融行业从业者
- **实体关系：**
  - 金融行为：如基金间的共同重仓持股关系、公司相互持股关系等
  - 社会关系：如基金经理间的校友关系、同城关系、年龄相仿等

✓ 金融网络中的实体表示学习

金融网络中的实体表示学习

- 借助实体关系，为金融实体学习一个低维度的稠密向量
- 基于金融实体的表示，进行金融实体间的相关度计算
  - 相关公司推荐
  - 关联波动预测
  - 关联交易挖掘
  - 股票价格预测

结合公司网络关系的股票价格预测



公司表示学习

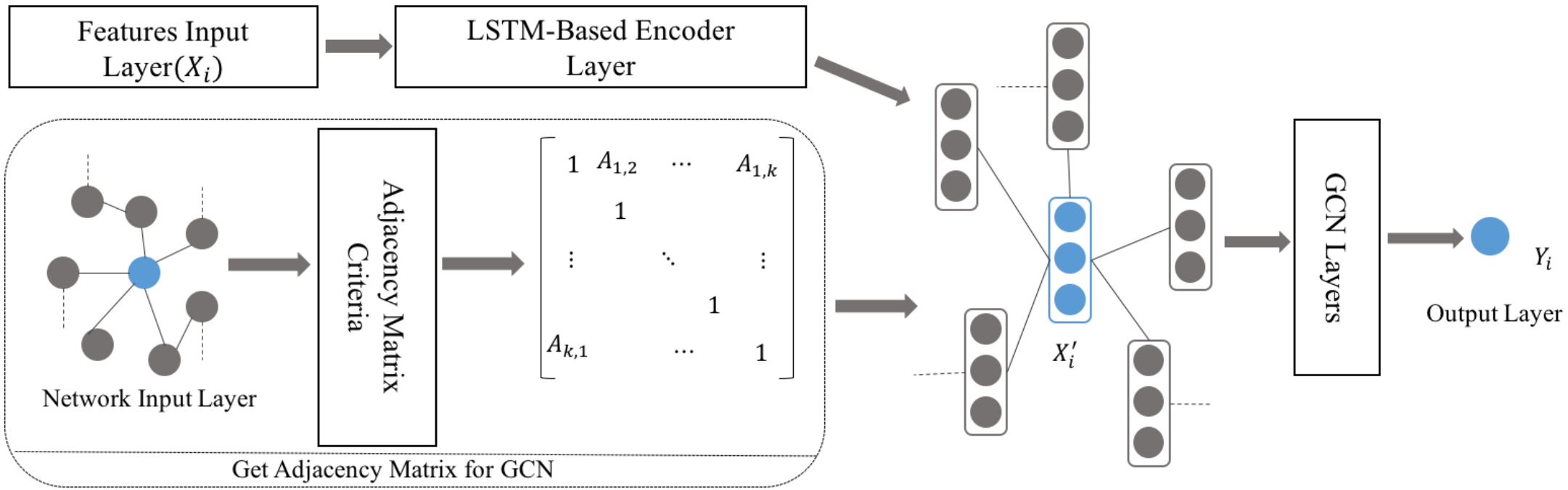
结合邻居特征
混合特征表示

基于循环神经网络
的历史信息建模和目标预测

- 使用公司表示，计算公司相关度，选择邻居节点

Incorporating Corporation Relationship via Graph Convolutional Neural Networks for Stock Price Prediction, CIKM‘18

基于图卷积模型的公司表示学习



- 使用图卷积模型，利用公司网络信息进行公司的表示学习
- 不限公司个数

金融实体网络构建

- 数据来源：万得数据库
- 实体：上市公司、持股机构
- 关系：持股关系（上市公司的十大持股机构）

- 截至2017年4月29日的数据
  - 20, 836 个网络节点 （公司+机构）

实验设置 - 股票价格预测

- **数据来源：Tushare**
- **对象：沪深三百上市公司**
- **预测标签：股票涨跌**
  - 如果开盘价格高于上一个交易日的开盘价，则标记为涨，否则为跌
- **任务设定：利用历史信息进行当前交易日涨跌的预测**
  - 七个工作日股价信息作为历史信息
  - 某一日的特征表示 <开盘价，高点，低点，收盘价，交易量>

	训练	测试
测试样例数	31,066	13,315
上涨天数的比例	0.528	0.501
时间区间	29/04/2017 – 13/10/2017	16/10/2017 – 31/12/2017

实验设置 – 对比方法

- Majority： 始终猜数据中最多的那一类（跌）
- LR： 单个公司的离散特征，使用多层感知机做预测
- LSTM： 采用循环神经网络进行特征学习

- 使用邻居信息进行特征扩展：
  - Deepwalk + LSTM
  - Node2vec + LSTM
  - LINE + LSTM
  - GCN + LSTM: 使用图神经网络进行节点表示学习

实验结果

- 使用邻居节点的特征可以帮助目标实体进行更好的股价预测
- 使用图神经网络进行公司表示学习可以得到更好的预测效果

方法	准确率	特征表示	邻居信息	分类器
majority	50.10%	-	-	-
LR	52.07%	离散	无	回归
LSTM	53.17%	循环神经网络	无	多层感知机
DeepWalk+LSTM	56.93%	循环神经网络	是	多层感知机
node2vec+LSTM	56.61%	循环神经网络	是	多层感知机
LINE+LSTM	57.00%	循环神经网络	是	多层感知机
GCN+LSTM	**57.98%**	循环神经网络	是	多层感知机

事件

- **事件是特定时间、地点下的一个状态变化**
  - 经典的事件抽取和分类任务，ACE 2005
  - 话题检测和跟踪
  - 样例：XX 在去年**过世**了， 小明**打了**小红
- **三元组（或者多元组）的事件表示**
  - <XX，过世， NULL>，<小明，打了，小红>

金融事件

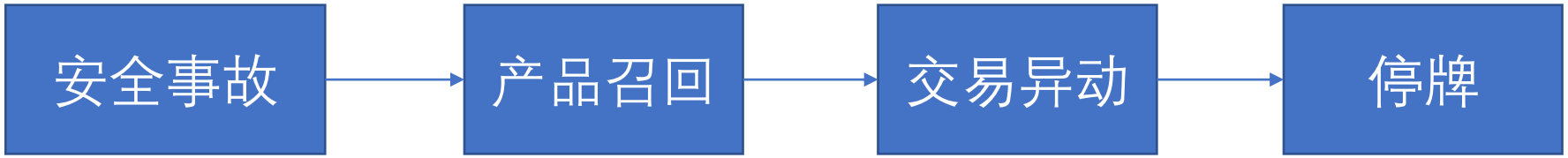
- **对金融市场产生影响的重要事件**
  - 例：***停牌***了近半个月的科大讯飞(002230.SZ)宣布***复牌***
  - 例：***河北雄安新区***，设立于在17年4月，但京津冀协同战略、打造北京疏解集中承载地、国家主席深入河北考察调研等新闻早已频繁出现

- **公告事件： 金融实体往往是其中的角色扮演者**
- **开放域事件： 自然事件，政治事件**

金融事件链条

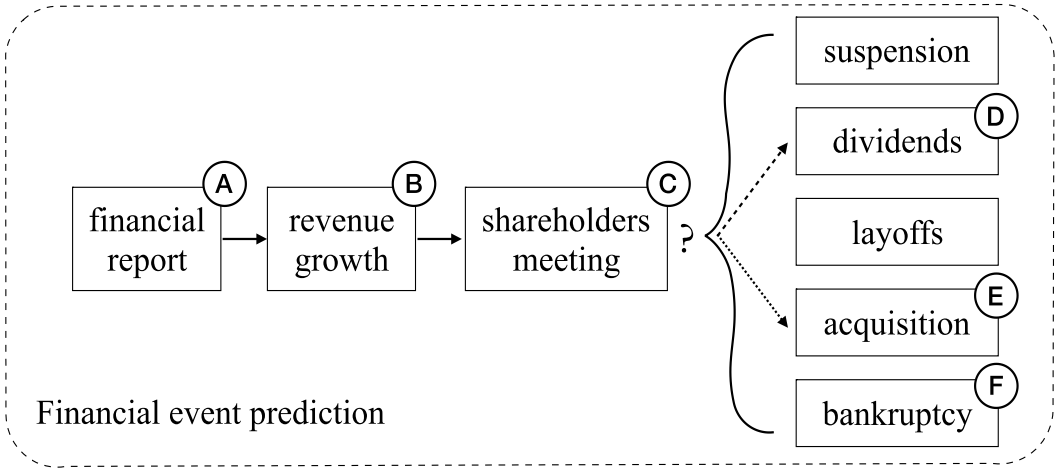
- **由同一时间轴上顺次发生、互有关联的多个事件形成的序列。**

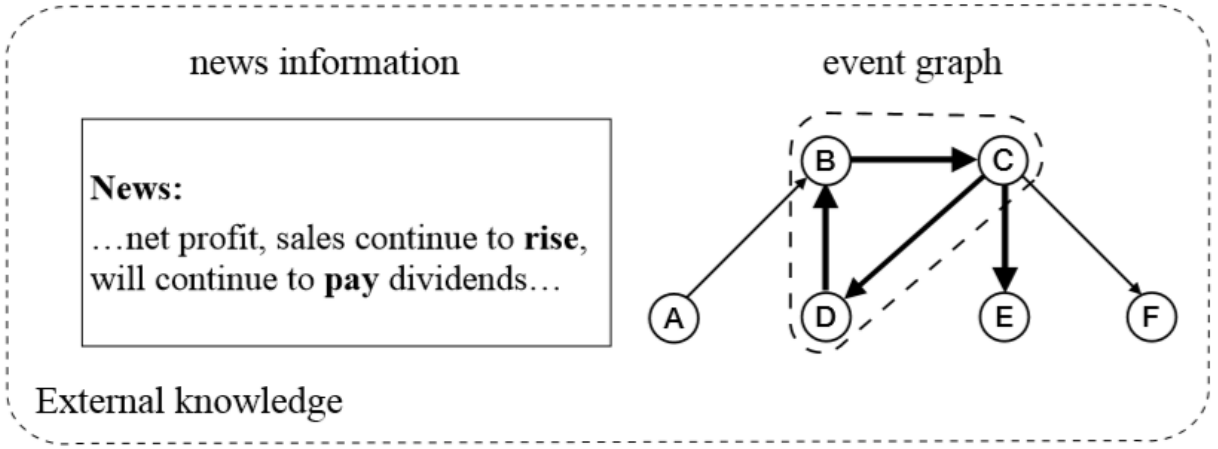
- 例：三星的Note 7 手机爆炸事件



金融公告事件预测

- **基于特定公司的历史公告事件链条，预测下一个公告事件**
  - 金融公告事件的发生受到**链条上历史事件**的影响
  - 金融公告事件的发生受到**开放域事件**的影响





金融领域的事件建模的相关工作

- **事件表示学习**
  - 基于频率统计进行事件表示学习 （Ding et al., EMNLP 2014 ）
  - 使用层次化神经网络，进行事件的表示学习（ Ding et al., IJCAI 2015）
- **事件预测**
  - 使用图神经网络进行开放域事件链条的预测（ Duan et al., IJCAI 2018）

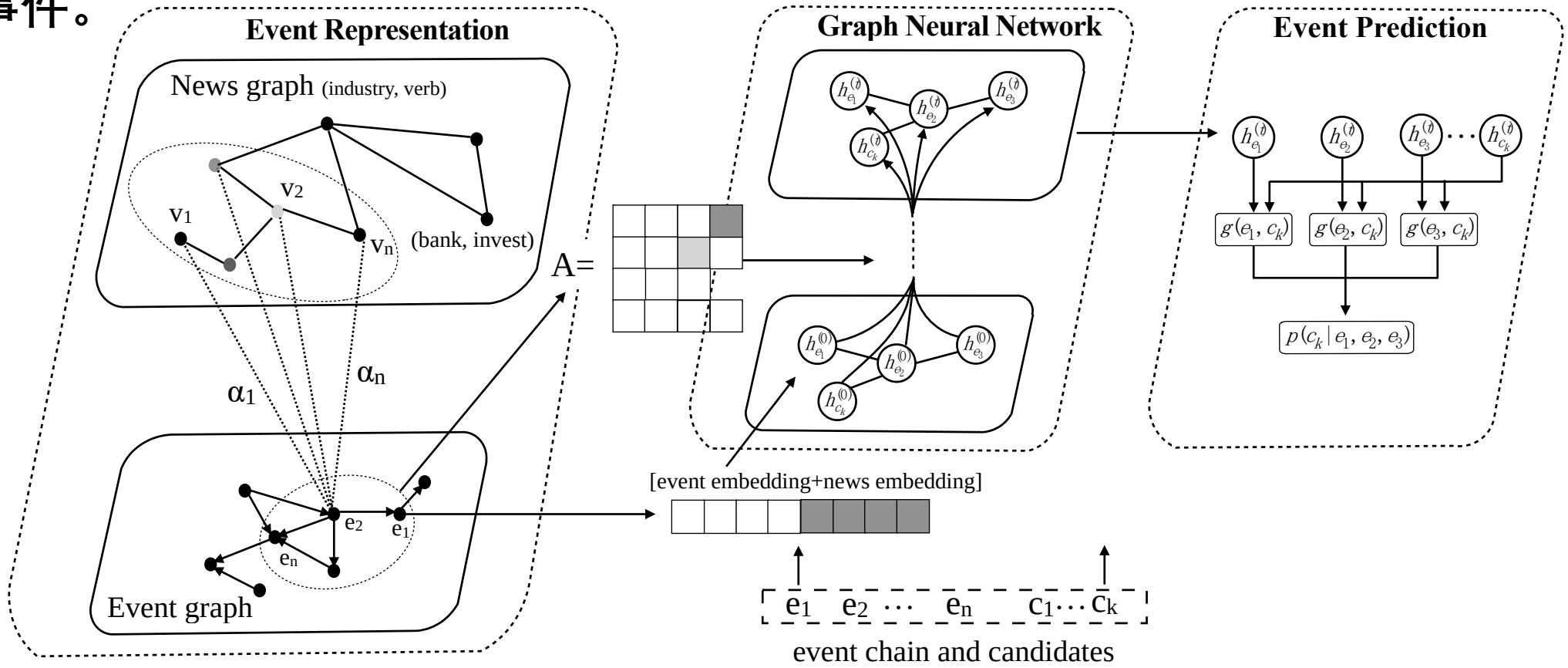
✓ **本研究的创新点**

✓ **以金融市场的公告事件为预测对象**

✓ **区分公告事件和开放域事件**

结合外部知识的金融公告事件预测框架

- 给定公告事件链条 $e_1, e_2, \ldots, e_n$ 和候选事件列表 $c_1,\ldots,c_k$ ， 输出概率最高的候选事件。



结合开放域事件的公告事件表示

基于事件图的公告事件表示更新

金融公告事件预测

结合开放域事件的金融公告事件表示

- **金融公告事件初始化**
  - $\mathrm{h_e}$ 表示金融事件e的初始化表示
- **相关开放域事件表示学习**
  - 利用时间窗口信息选取相关开放域事件集合
  - 使用图注意力机制混合所有相关开放域事件，得到表示$\mathrm{h_v}$
- **结合相关开放域事件的金融公告事件表示**
  - 通过对$\mathrm{h_e}$ 和$\mathrm{h_v}$的拼接获得金融公告事件的表示
  - $h_e^{(0)} = \mathrm{h_e} \oplus \mathrm{h_v}$

基于推理图的公告事件表示更新

- **构建推理图**
  - 节点：包含历史金融公告事件节点和候选事件节点
  - 邻接矩阵：获取金融公告事件图中的对应边信息
- **节点初始化表示**
  - $h^{(0)} = \left[h_{e_1}^{(0)}, \ldots, h_{e_n}^{(0)}, h_{c_1}^{(0)}, \ldots, h_{c_k}^{(0)}\right]$
- **基于门机制图神经网络的节点表示更新**

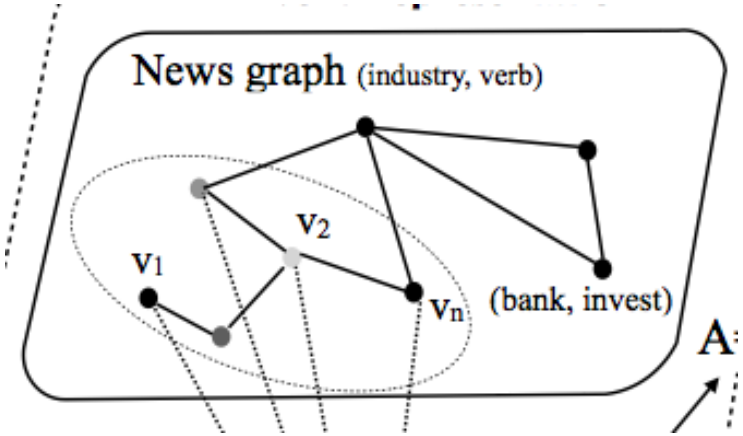
$$
\begin{aligned}
a^{(t)} &= \mathbf{A}^{\top} h^{(t-1)} + b \\
z^{(t)} &= \sigma(W_z a^{(t)} + U_z h^{(t-1)}) \\
r^{(t)} &= \sigma(W_r a^{(t)} + U_r h^{(t-1)}) \\
c^{(t)} &= tanh(W_c a^{(t)} + U_c(r^{(t)} \odot h^{(t-1)}) \\
h^{(t)} &= (1 - z^{(t)}) \odot h^{(t-1)} + z^{(t)} \odot c^{(t)}
\end{aligned}
$$

金融公告事件预测

- **给定公告事件链条 $e_1, e_2, \ldots, e_n$ 和候选事件列表 $c_1,\ldots,c_k$ ，输出概率最高的候选事件。**
- **对每一个候选事件，综合考虑每一个历史事件，计算其发生概率。**
  - $s_j = p(c_j|e_1, e_2, \ldots, e_n) = \frac{1}{n}\sum_{i=1}^{n} s_{ij}$
  - 给定一个事件对，$h_{e_i}^{(t)}$ 和$h_{c_j}^{(t)}$，两个事件的相关性计算公式，欧氏距离 $s_{ij} = g(h_{e_i}^{(t)}, h_{c_j}^{(t)})$
- **训练目标函数：**
  - $L(\Theta) = \sum_{i=1}^{n}\sum_{j=1}^{k}(max(0, m - s_{iy} + s_{ik})) + \frac{\lambda}{2} \| \Theta \|^2$

金融事件图构造

- **开放域事件图构造**
  - 获取与公司相关的新闻，进行事件抽取
  - 节点：简化的二元组表示，金融实体和动作
  - 边：同一天发生的事件节点之间有连边



- **公告事件图构造**
  - 从公司公告中获取相关事件
  - 节点：金融公告事件
  - 边：同一家公司顺次发生的公告事件之间或有一条边

公告事件集合

- 2013 至 2017年的1,271,130个公告事件，360个事件类型，25个事件类别
- 3,556 个公司
- 平均每家公司357个事件
- 平均事件间隔为6.52天

Event type	Number	Percent
Restructuring suspension	93,455	6.90%
Disclosure suspension	83,999	6.20%
Trading irregularity	53,723	3.97%
Notice of shareholders meeting	46,804	3.46%
Block trading	42,094	3.11%
Dividend announcement	27,573	2.04%
Equity pledge announcement	25,296	1.87%
Releasing shares announcement	16,819	1.24%
Disclosure of annually report	14,460	1.07%
Shareholders reduce shareholding	14,295	1.06%

开放域事件语料集合

- 169 万 外部新闻
- 28 行业 （鉴于公司稀疏，在实际操作中，我们针对行业进行事件二元组构建）
- 30,042 事件类型（动词）
- 200,864 (行业, 动词) 对

实验设置

- **针对某个上市公司的所有公告事件，通过划窗方式构建事件链条**
  - 窗口为5，给定前四个事件，预测第五个事件发生的概率

	训练	验证	测试
年份	2013-2016	2017前半年	2017后半年
个数	1,687,366	327,437	239,095

- 评价指标： Top-3 的预测准确率

对比模型

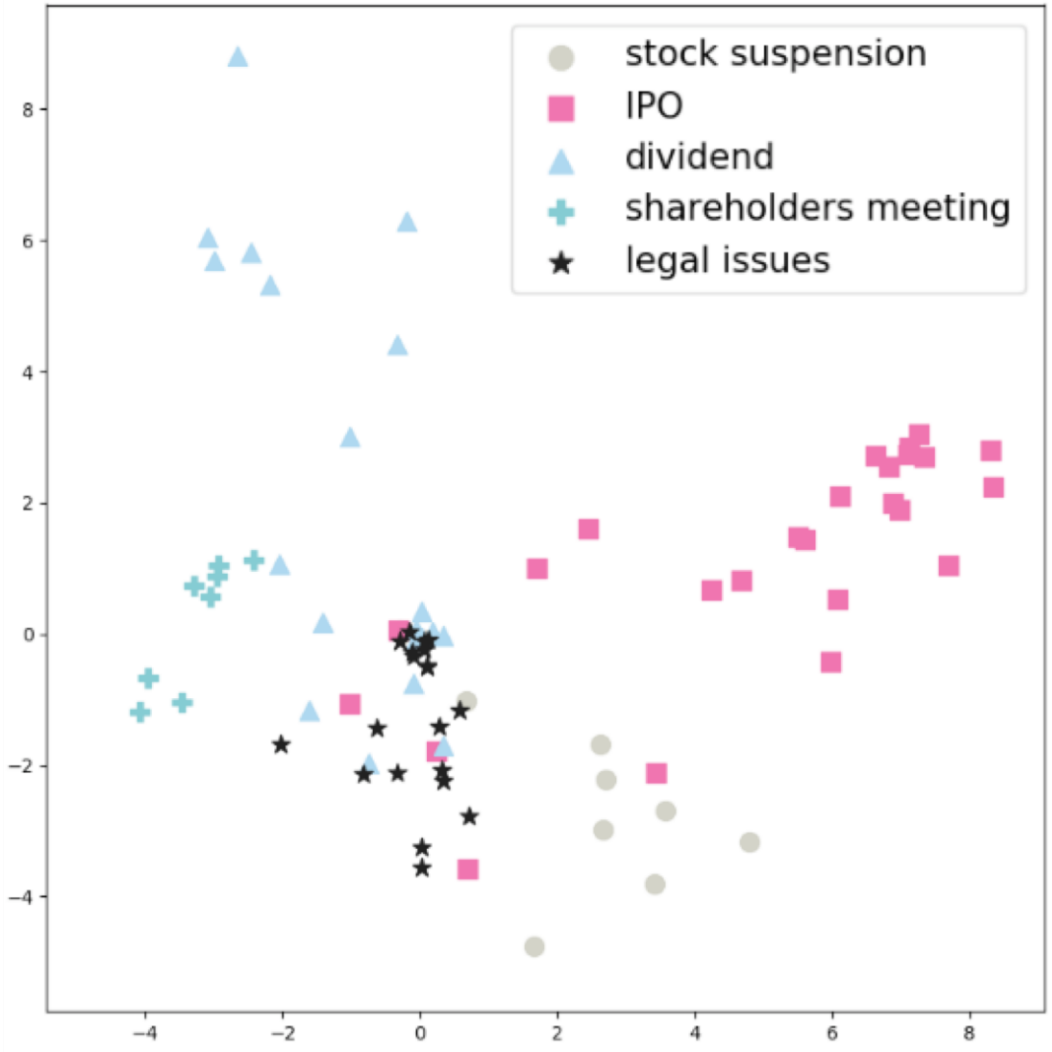
- **PMI [Chambers and Jurafsky, 2008]:** 基于互信息的事件预测模型
- **Word2Vec [Mikolov *et al.*, 2013b]:** 使用公告事件的字表示进行事件建模
- **DeepWalk [Perozzi *et al.*, 2014]:** 使用deepwalk进行事件表示学习
- **LSTM [Wang *et al.*, 2017]:** 对历史事件链条进行综合考虑

- **GGNN:** 我们的图卷积模型.
- **GGNN+News (mean):** 结合开放域事件的图模型，平均加权
- **GGNN+News (attention):** 结合开放域事件的图模型，注意力加权

实验结果

Methods	Acc.(%)	Pre.(%)	Re.(%)	F1(%)
PMI[1]	59.79	88.16	57.43	66.44
DeepWalk [10]	78.93	86.26	79.99	81.24
Word2Vec [8]	80.35	86.11	81.42	82.12
LSTM [11]	86.75	90.97	87.93	88.47
GGNN	88.12	90.65	88.73	89.15
GGNN+News (mean)	89.99	92.25	90.73	91.08
GGNN+News (attention)	**90.55**	92.57	91.13	**91.53**

事件表示的可视化分析



Two-dimensional PCA projection of 128-dimensional event embeddings.

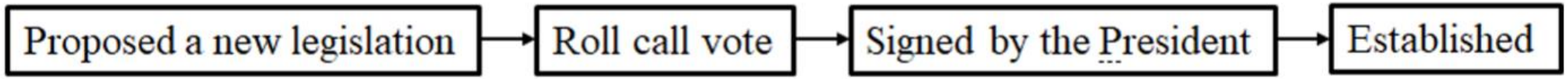
目录

- **图神经网络的背景介绍**
- **计算金融**
  - 面向股票预测的金融实体表示学习 （CIKM’ 2018）
  - 引入外部新闻资讯的金融事件预测 (CIKM’ 2019)
- **量化政治**
  - 结合国会议员关系的议案投票结果预测 (IJCAI’ 2020)
- **社会媒体用户画像**
  - 联合用户言论和关系的用户画像 (在研项目)
- **总结**

美国国会的立法流程

- 议员：议事机关中的公务人员
- 议案：由议员向国家议事机关提出的议事原案
- 投票：议员针对议案进行立场表达

- 美国国会的立法流程

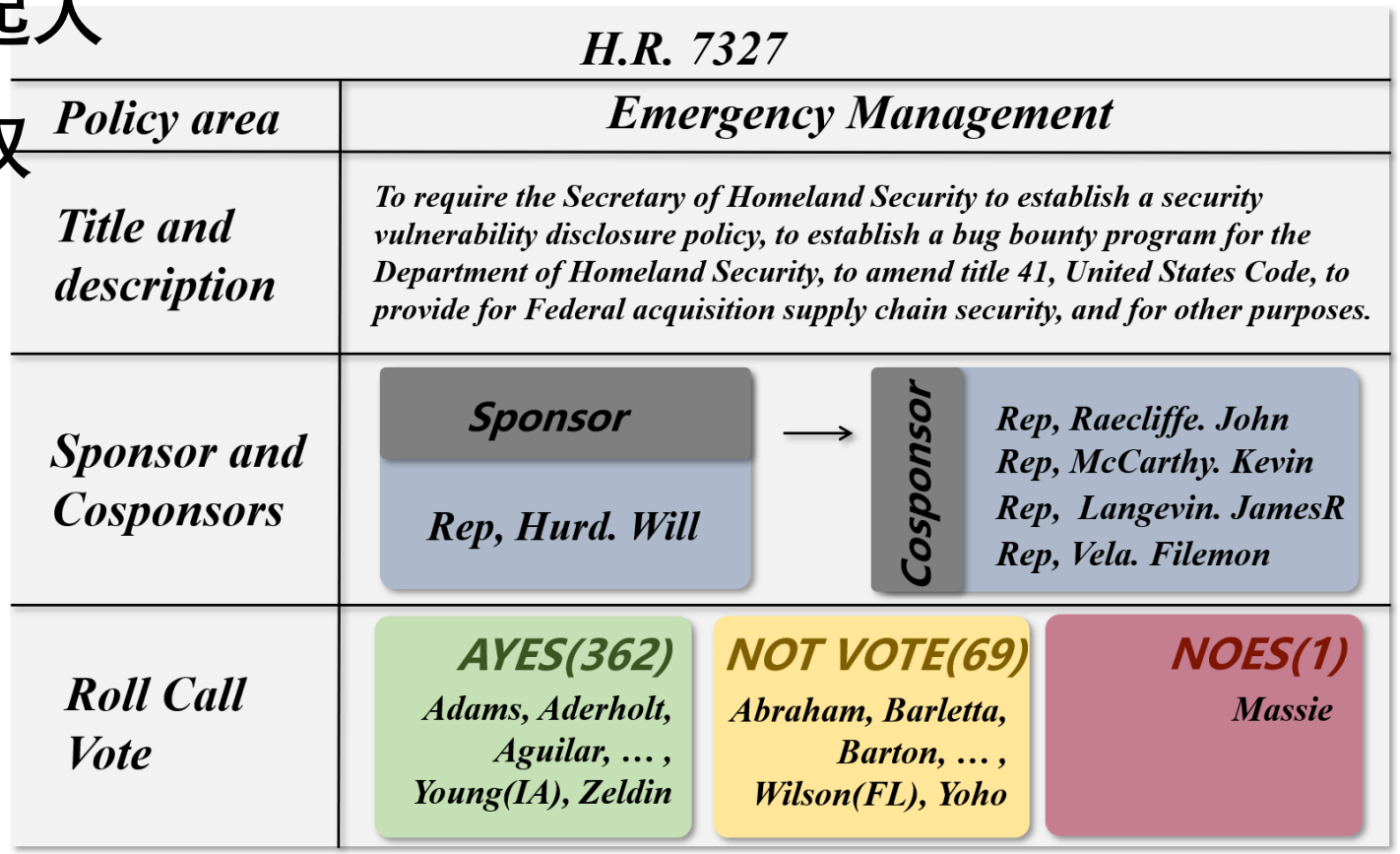


议员信息

- ID：议员的唯一标识
- 党派：民主党、共和党等
- 州：议员所属的州
- 国会工作年限：参议院、众议院的服务时间

议事原案投票样例

- 政策领域：议案所属的类目，如，环境、外交等
- 议案文本：标题以及描述
- 发起议员：主发起和联合发起人
- 投票结果：赞成，反对，弃权



国会议员的投票结果预测

- **给定一个议事原案，判定议员的投票立场（支持，反对，弃权）**
- **Ideal point model [Clinton et al., 2004]**
  - 通过历史的投票纪录学习议员表示（投票结果相同的议员更相近）
  - 将议案和议员映射到相同空间，进行立场判断
- **[Gerrish and Blei., 2011] 在Ideal point model上扩充议案文本信息**

- **缺少对议员之间关系的建模**

本研究贡献

- 我们引入议员的关系网络进行议员的表示学习
- 我们使用议案的文本信息，进行议案的表示学习
- 我们采用一个三元的损失函数进行议员和议案的联合表示学习
- 我们构建了第一个公开的议员投票结果预测语料集

Joint Representation Learning of Legislator and Legislation for Roll Call Prediction, IJCAI’20.

任务定义

- **给定一个议事原案和国会议员，预测每一个议员的投票结果**
  - 议员 $M = \{m_1, m_2, \dots, m_k\}$
    - ID: $m_i(ID)$
    - Party: $m_i(p)$
    - State: $m_i(s)$
  - 议案 $L = \{l_1, l_2, \dots, l_n\}$
    - 议案描述: $l_i(\mathrm{d})$
    - 发起人和联合发起人网络: $l_i(\mathrm{s})$
  - 投票结果 $\mathrm{R} = \{r(m_i, l_j) | 1 \le i \le k, 1 \le j \le n\}$
    - $r(m_i, l_j)$ 是议员i在议案j上的投票结果
    - 投票结果有三个标签：赞成，反对，弃权

基于图神经网络的议员表示学习

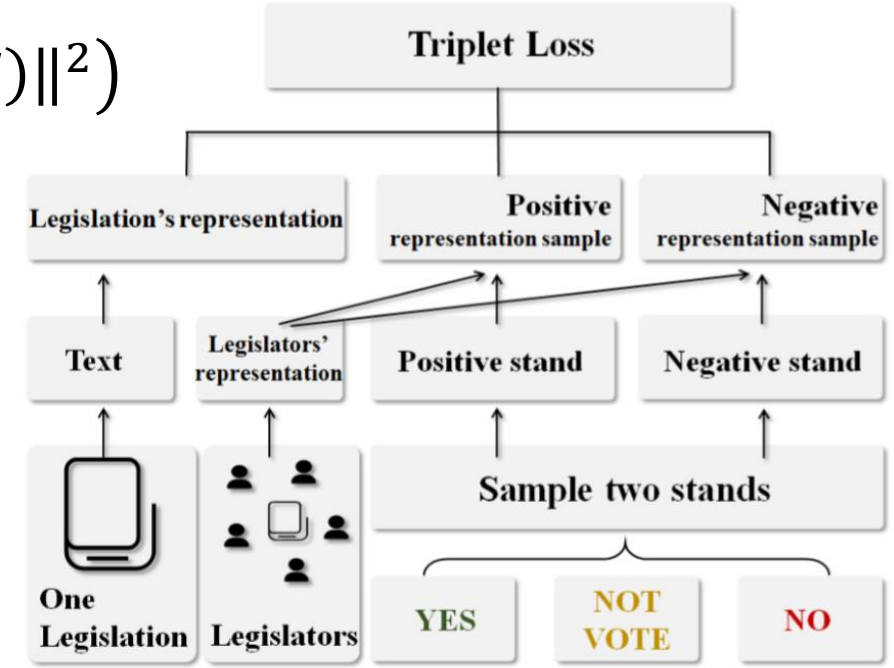
- **议员表示初始化**
  - $X_{initial}(i) = m_i(ID) \oplus m_i(p) \oplus m_i(s)$
- **议员网络构建**
  - 基于发起-联合发起的关系构建关系矩阵A
  - a_{ij}表示议员$\mathrm{m_i}$和议员$\mathrm{m_j}$共同发起过的议案个数
- **基于图卷积网络的议员表示更新**
  - $Z = f(X, A) = AReLU\left(AX_{input}W^0\right)W^1$

基于循环神经网络的议案表示学习

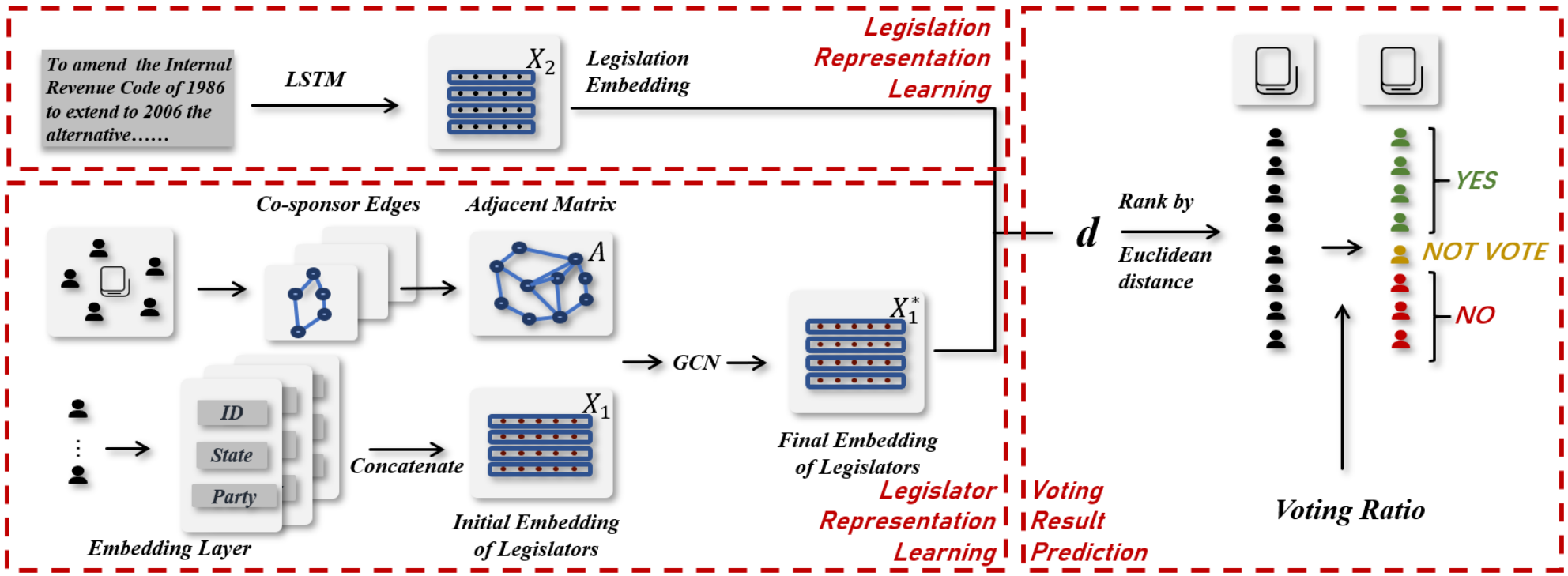
- **我们拼接了议案标题和描述作为议案文本信息，使用循环神经网络进行议案的表示学习**
  - $X_{bill}(j) = LSTM(l_j(d))$
  - $l_j(d))$ 是议案l_j的文本信息

基于三元损失函数的联合表示学习

- **训练过程**
  - 采样一个（$m_i^+, m_k^-, l(j)$）三元组
  - 满足r(m_i^+, $l(j)$) < r(m_k^-, $l(j)$) , 给定 YES < Not Vote < NO
- **三元损失函数**
  - $L = \max\left(\varepsilon, \|\mathrm{X}_m(i)^+ - \mathrm{X}_l(j)\|^2 - \|\mathrm{X}_m(k)^- - \mathrm{X}_l(j)\|^2\right)$

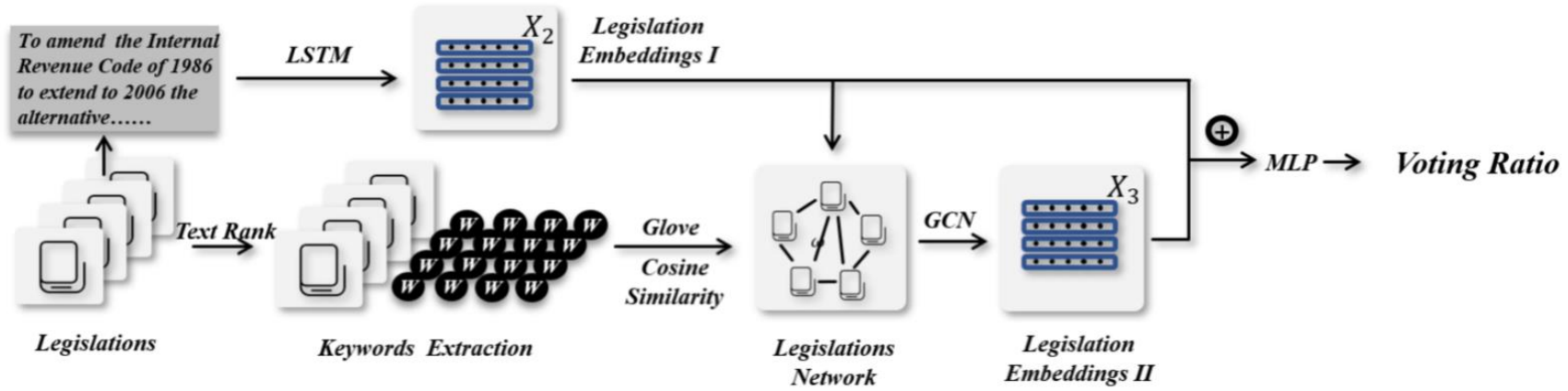


结合议员关系和议案文本信息投票结果预测



- 议员投票结果预测
  - 计算议员与议案的欧氏距离
  - 给定**投票结果比例（三个标签）**，从欧式距离从小到大进行标签分配

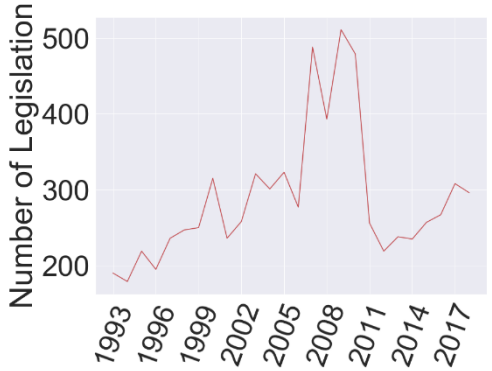
投票结果比例预测

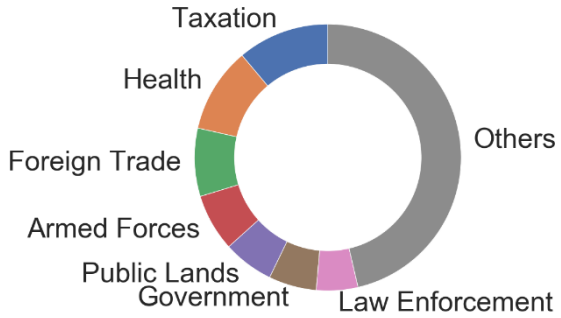


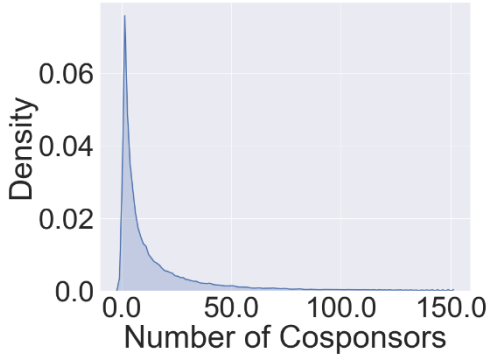
- **我们使用语义图神经网络进行议案的表示学习，并进行投票结果比例的预测**
  - 语义图的节点是单词，边是在同一个议案中的共现关系
  - 预测结果是一个三维向量，采用MSE作为损失函数

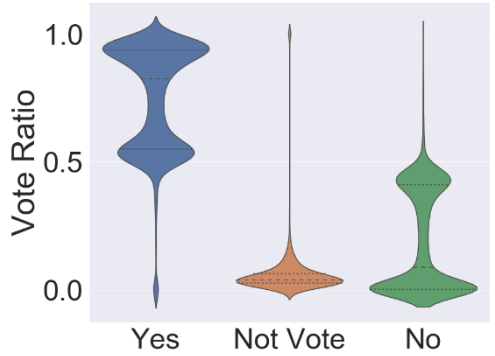
语料集介绍

- **数据来源：美国国会网 https://www.congress.gov**
  - 时间分布：1993年到2018年
  - 议员数目：2,347
  - 议题数目：215,857
  - 投票结果数：2,234,082









实验设置

- **实验数据集划分**
  - 5年为窗口，以滑窗的方式构建实验数据集，前4年为训练集，第5年为测试集
  - 使用前四年的议员关系构建议员网络网络
  - 从1993 - 2018，一共包含22个实验集

- **评测指标**
  - 所有的实验集合的平均准确率

对比方法

- **Clinton:** 基于贝叶斯方法的模型，将议员和议案投影到同一空间， ideal point model
- **Gerrish:** 基于 ideal point model，将文本信息考虑到推断过程
- **LSTM + Party:** 使用党派特征作为议员表示，使用LSTM进行议案表示学习，两者串联进行投票结果预测
- **LSTM + deepwalk:** 使用deepwalk为议员进行表示学习，使用LSTM进行议案表示学习，两者串联进行投票结果预测
- **LSTM + node2vec:** 使用deepwalk为议员进行表示学习，使用LSTM进行议案表示学习，两者串联进行投票结果预测
- **LSTM + GCN:** 使用GCN为议员进行表示学习，使用LSTM进行议案表示学习，两者串联进行投票结果预测
- **LSTM + GCN + triplet loss:** 我们提出的模型，自动预测议案投票结果比例
- **LSTM + GCN + triplet loss (GT):** 我们提出的模型，使用真实的投票结果比例

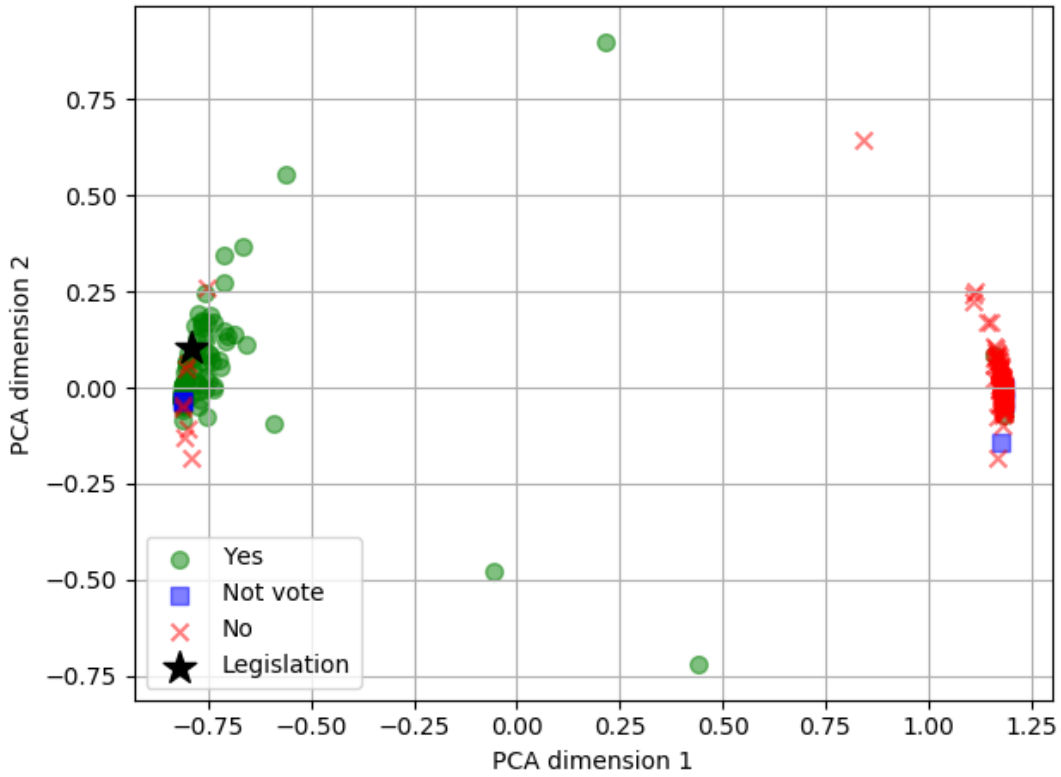
实验结果

Model	Acc
clinton	68.10
gerrish	75.30
LSTM + deepwalk	75.88
LSTM + node2vec	75.89
LSTM + Party	76.26
LSTM + GCN	76.75
LSTM + GCN + triplet loss	**78.09**
LSTM + GCN + triplet loss (GT)	81.86

- 党派信息是很强的议员特征，与投票结果强相关
- 采用图神经网路和三元损失函数可以获得最高的预测准确率

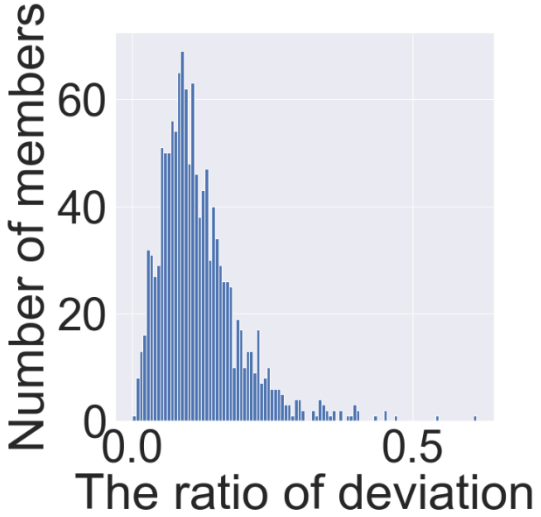
议员和议案表示学习结果案例

- 针对议员和议案的表示学习结果，我们采用PCA进行2维的特征提取
- 投赞成票的议员和议案有明显的聚集效果
- 投反对票的议员有聚集效果

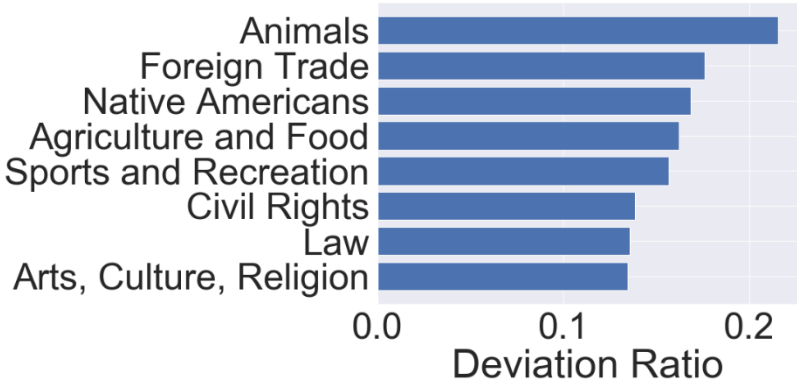


党派立场与投票结果相关性分析

- **党派内部一致性**
  - 投票结果与党派立场一致的同党派议员个数
  - 党派立场：党派大多数议员的投票立场
  - 88.42% （民主党：87.05% VS 共和党：89.11%）



- **偏离度分析**
  - 议员偏离度：议员的投票中与党派立场不一致的比例
    - 绝大部分议员投票与党派立场保持一致
  - 议案偏离度：议案中投票偏离党派立场的议员比例
    - 动物、外贸、农业、体育的偏离度高



高偏离度议员行为预测

- 针对偏离度最高的5%议员，我们观察不同模型的预测准确度
- 我们提出的模型优势更加明显

Model	Acc
LSTM + deepwalk	62.37
LSTM + node2vec	62.61
LSTM + Party	62.85
LSTM + GCN	62.70
LSTM + GCN + triplet loss	**65.93**
LSTM + GCN + triplet loss (GT)	67.68

目录

- **图神经网络的背景介绍**
- **计算金融**
  - 面向股票预测的金融实体表示学习 （CIKM’ 2018）
  - 引入外部新闻资讯的金融事件预测 (CIKM’ 2019)
- **量化政治**
  - 结合国会议员关系的议案投票结果预测 (IJCAI’ 2020)
- **社会媒体用户画像**
  - 联合用户言论和关系的用户画像 (在研项目)
- **总结**

社交网络中的用户画像

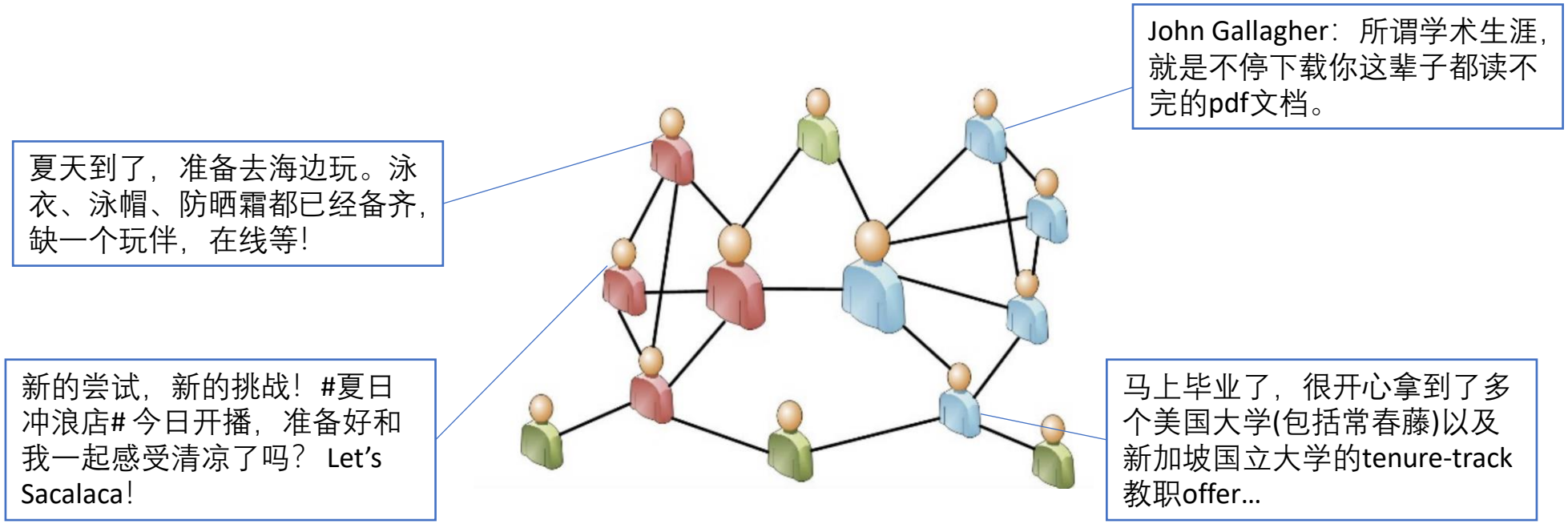
- **对于给定用户，针对不同的属性进行相关标签的判断**
  - 属性可以包括，性别、年龄、教育水平、地域、职业、工资收入、用户偏好、社群归属等等。

Convertlab 用户画像与精准经销

BlueMC 社群画像

社交网络中异质图特性

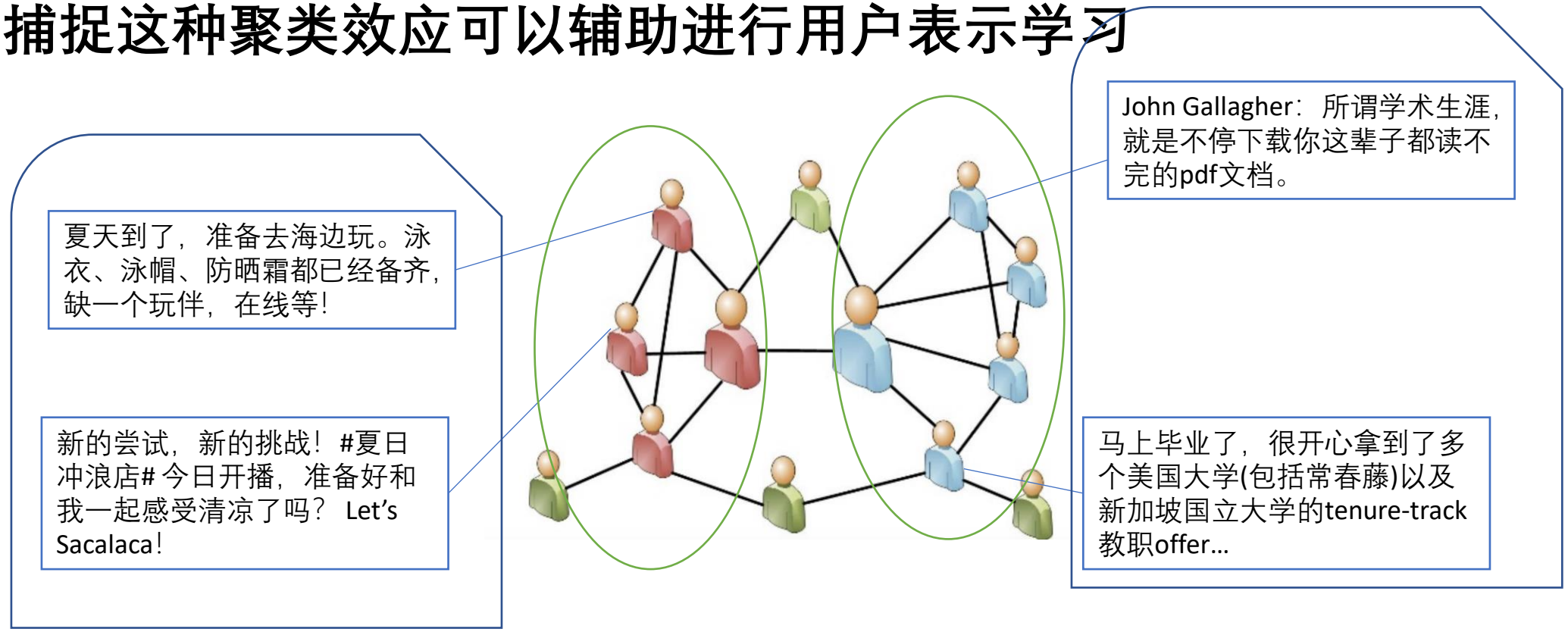
- 用户通过关注关系形成网络
- 词语通过语义关联形成网络
- 用户通过发言与词语相关联



社交网络中的聚类效应

- 用户网络中包含了**群组**特性
- 文本网络中体现了**话题**特性

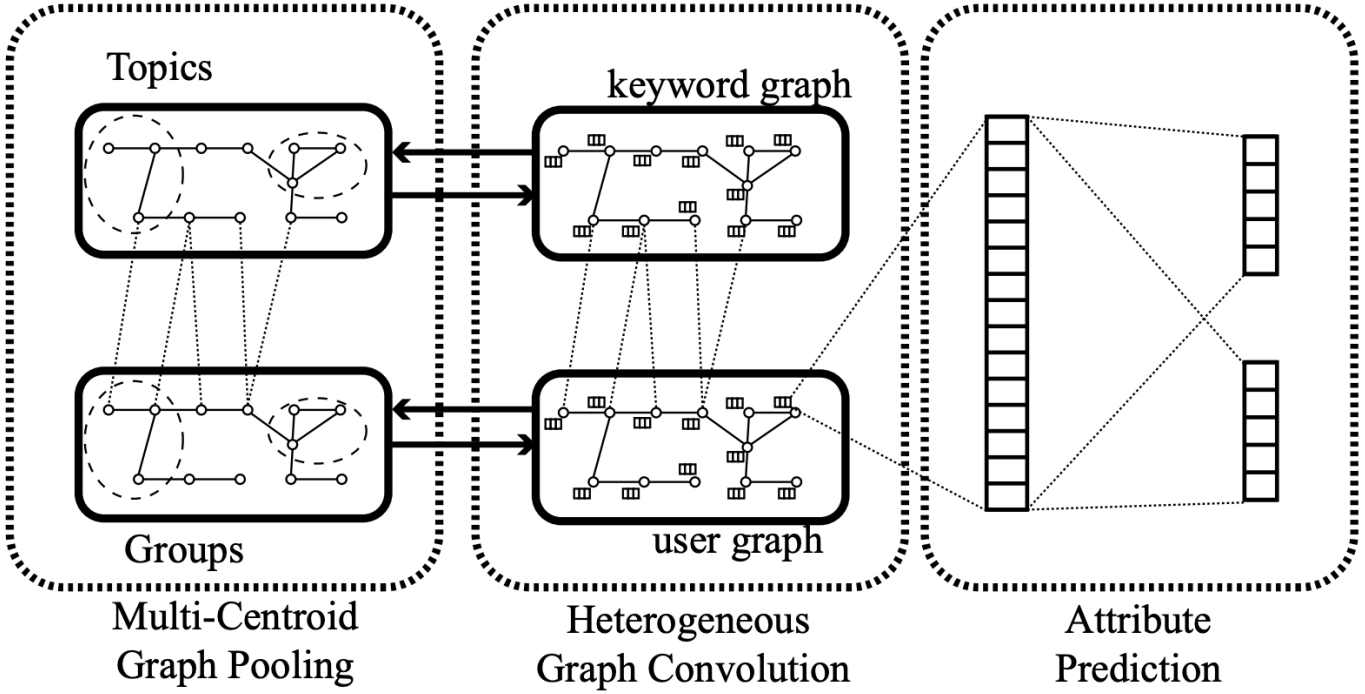
✓ 捕捉这种聚类效应可以辅助进行用户表示学习



捕捉社交媒体聚类特性的用户画像模型

- **异质图网络构建：**
  - 用户网络：用户的关注关系
  - 语义网络：词语的共现关系

- **多中心的图池化方法：**
  - 用户网络形成**多个群组**
  - 语义网络形成**多个话题**



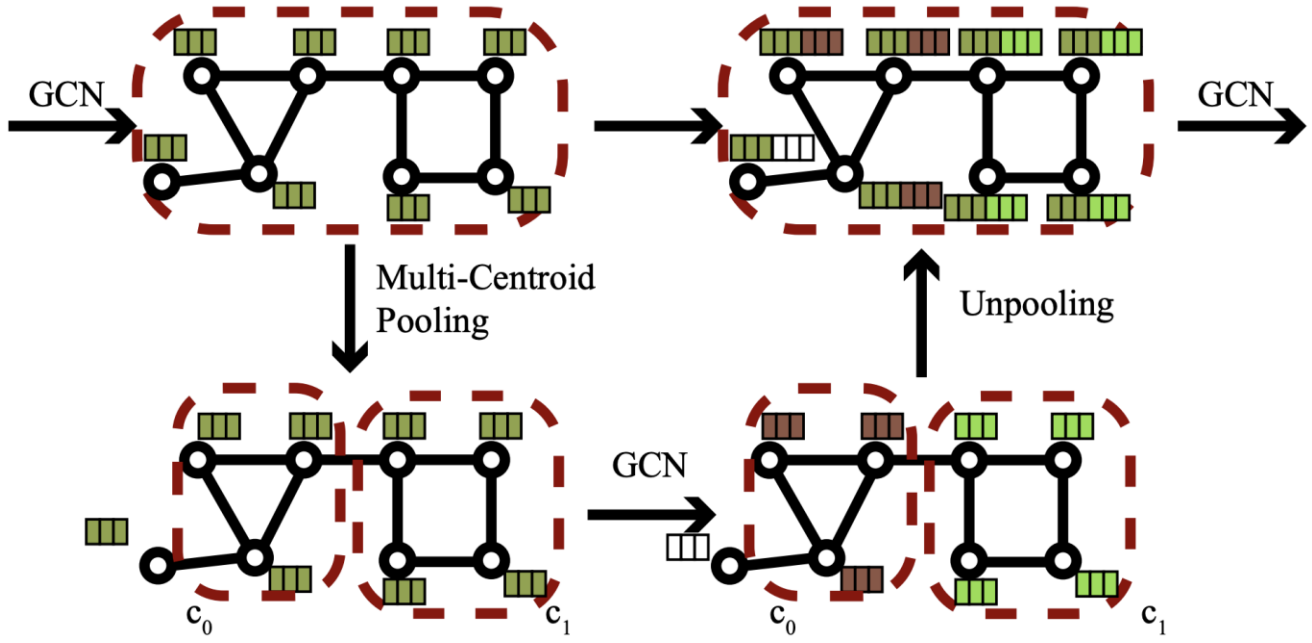
Capturing Community Characteristics for User Modeling via Heterogeneous Multi-Centroid Graph Pooling, 投稿中

异质网络中的GCN模型

- 同时在用户网络和文本网络进行表示学习
  - $X^{n+1} = \sigma\left(\left(\hat{A}X^{n} + \lambda_1 CY\right)W_1\right)$
  - $Y^{n+1} = \sigma\left(\left(\hat{B}Y^{n} + \lambda_2 C'X\right)W_2\right)$
  - 其中X, Y分别为用户和文本的节点表示,
  - A, B, C分别为用户关系、文本关系、用户到文本的邻接矩阵

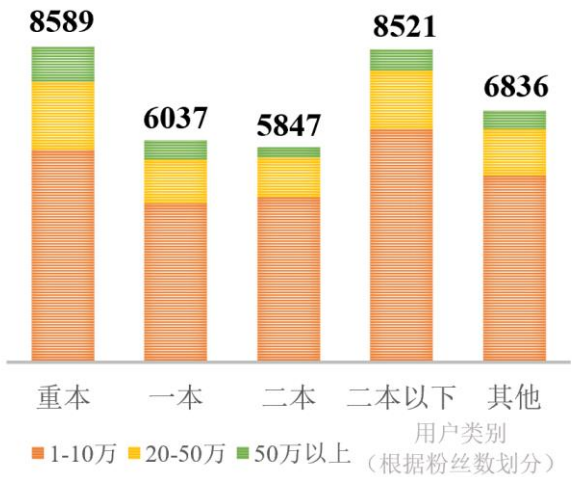
面向节点聚类的图池化操作

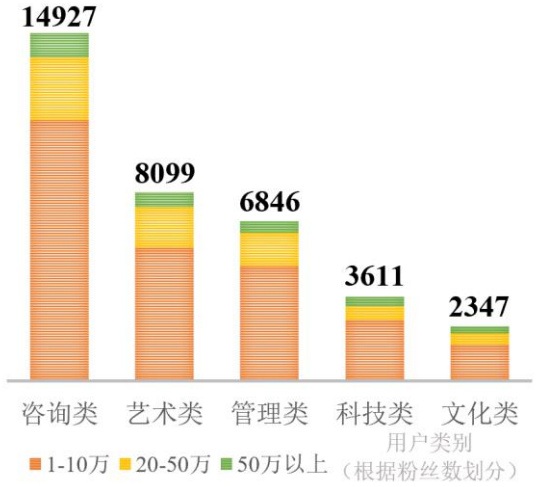
- **类别捕捉：**维护k个聚类中心，选择距离聚类中心最近的点形成类
- **类内信息传递：**在同类节点中进行信息交互和表示更新



数据集构建

- **用户数：35,830**
  - 包含职业和教育信息
- **微博总数：520万**
- **属性信息：职业、教育属性信息**
  - 教育划分：按用户所填学校对应的高考平均分数线划分（2015年的录取分数）
  - 职业划分：使用词表示进行K-means聚类，人工识别职业分类
- **关系网络：关注关系**
- **发言特征：原创微博和评论**





实验设置

- **数据集合划分**
  - 将用户划分成不重叠的三个部分。

训练集	验证集	测试集
23,734	5,984	5,984

- **评价指标**
  - 在教育、工作五分类任务上的分类准确率

实验数据集总体描述

- **微博用户画像语料集 Weibo**
  - 图中节点为用户和词语，预测用户属性标签
- **电影评论数据集 IMDB**
  - 图中节点为电影与演员，预测电影类型
- **论文引用关系数据集 DBLP**
  - 图中节点为论文与作者，预测论文类型

数据集	节点数	边数	类别数	特征数
IMDB	10,621	143K	3	1,232
DBLP	18,385	509K	4	334
Weibo	45,830	1,095K	5	32

对比方法

- **GCN:** Kipf & Welling, 2017
- **HAN:** 使用注意力机制的的异质网络. Wang et al. 2019

- **GCN+Multi-Pooling:** 使用用户网络GCN和池化聚类
- **HGCN:** 使用异质网络GCN
- **HGCN+Multi-Pooling:** 使用异质网络GCN和池化方法

实验结果

- 我们的模型在所有的语料集中都达到了最佳性能
- 引入语义的异质网络GCN能够提升节点属性预测的性能
- 引入池化操作能够进一步提升节点属性预测的性能

模型	IMDB	DBLP	微博
GCN	55.4%	88.2%	43.3%
HAN	55.4%	90.2%	42.4%
GCN+Multi-Pooling	56.8%	89.5%	44.2%
HGCN	58.2%	90.2	45.9%
HGCN+Multi-Pooling	**59.6%**	**91.0%**	**46.9%**

总结

- 引入图神经网络可以更好的使用应用领域的背景知识
- 对于领域的了解，构建任务，是计算背景的第一道门槛
- 对于模型结果的解读，需要目标领域专家的参与
- 特定场景任务需求可以推动计算模型的创新

报告相关论文

- **计算金融**
  - Yingmei Chen and Zhongyu Wei*, *Incorporating Corporation Relationship via Graph Convolutional Neural Networks for Stock Price Prediction*, CIKM'18
  - Yiying Yang, Zhongyu Wei*, Qin Chen and Libo Wu, *Using External Knowledge for Financial Event Prediction based on Graph Neural Networks*, CIKM'19
  - 数据：http://www.sdspeople.fudan.edu.cn/zywei/data/financial-event-prediction-fudanU.zip
- **量化政治**
  - Yuqiao Yang^, Xiaoqiang Lin^, Geng Lin, Zengfeng Huang, Changjian Jiang* and Zhongyu Wei*, *Joint Representation Learning of Legislator and Legislation for Roll Call Prediction*, IJCAI'20.
  - 数据：http://www.sdspeople.fudan.edu.cn/zywei/data/fudan-USRollCall.zip
- **用户画像**
  - Shangyi Ning, Qin Chen, Zengfeng Huang, Weijian Sun and Zhongyu Wei, *Capturing Community Characteristics for User Modeling via Heterogeneous Multi-Centroid Graph Pooling*, 投稿中

参考文献

- Zhang Z, Cui P, Zhu W. Deep learning on graphs: A survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2020.
- Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020.
- Kinderkhedia M. Learning Representations of Graph Data--A Survey[J]. arXiv preprint arXiv:1906.02989, 2019.
- Yang C, Xiao Y, Zhang Y, et al. Heterogeneous Network Representation Learning: Survey, Benchmark, Evaluation, and Beyond[J]. arXiv preprint arXiv:2004.00216, 2020.
- Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. ICLR 2017.
- Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.
- Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014: 701-710.
- Clinton J, Jackman S, Rivers D. The statistical analysis of roll call data[J]. American Political Science Review, 2004: 355-370.
- Gerrish S M, Blei D M. Predicting legislative roll calls from text[C]//Proceedings of the 28th International Conference on Machine Learning, ICML 2011. 2011.
- Chambers N, Jurafsky D. Unsupervised learning of narrative event chains[C]//Proceedings of ACL-08: HLT. 2008: 789-797.
- Wang X, Ji H, Shi C, et al. Heterogeneous graph attention network[C]//The World Wide Web Conference. 2019: 2022-2032.
- Gao H, Ji S. Graph u-nets[J]. arXiv preprint arXiv:1905.05178, 2019.

图神经网络在交叉学科领域的应用研究

魏忠钰 副教授

复旦大学 大数据学院

2020年7月3日
狗熊会

http://www.sdspeople.fudan.edu.cn/zywei/