社交网络挖掘期末汇报

汇报人：陈乐偲 姜俊哲 于淼芃

目录

CONTENTS

PART ONE

社区发现

社区发现

Community Detection

评价标准：模块度

$$Q = \frac{1}{2m} \sum_{1 \le i,j \le n} \left(\left[A_{ij} - \frac{d_i d_j}{2m} \right] \delta(c_i, c_j) \right)$$

Louvain Algorithm

Louvain on Facebook-combined

基于模块度和贪心思想的社区发现[1]

- 1.初始每个点视作一个社区
- 2.对每个点贪心选择一个模块度增加最大的社区加入
- 3.迭代2，直至无法更新
- 4.对每个社区缩点，回到1

[1] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008, 2008(10): P10008.

Louvain Algorithm

Louvain on Facebook-combined

基于模块度和贪心思想的社区发现

Infomap

Random Walk Based Method

基于随机游走的社区发现

对于随机游走的一段路径，采用分层编码，编码第一种是群组的名字，不同群组的名字编码不一样；第二种是每个群组内部的节点及跳出标志，不同节点的名字编码不一样。但是，不同群组内部的节点的编码可以复用。

对于一个好的社区划分，可以带来更短的平均编码长度

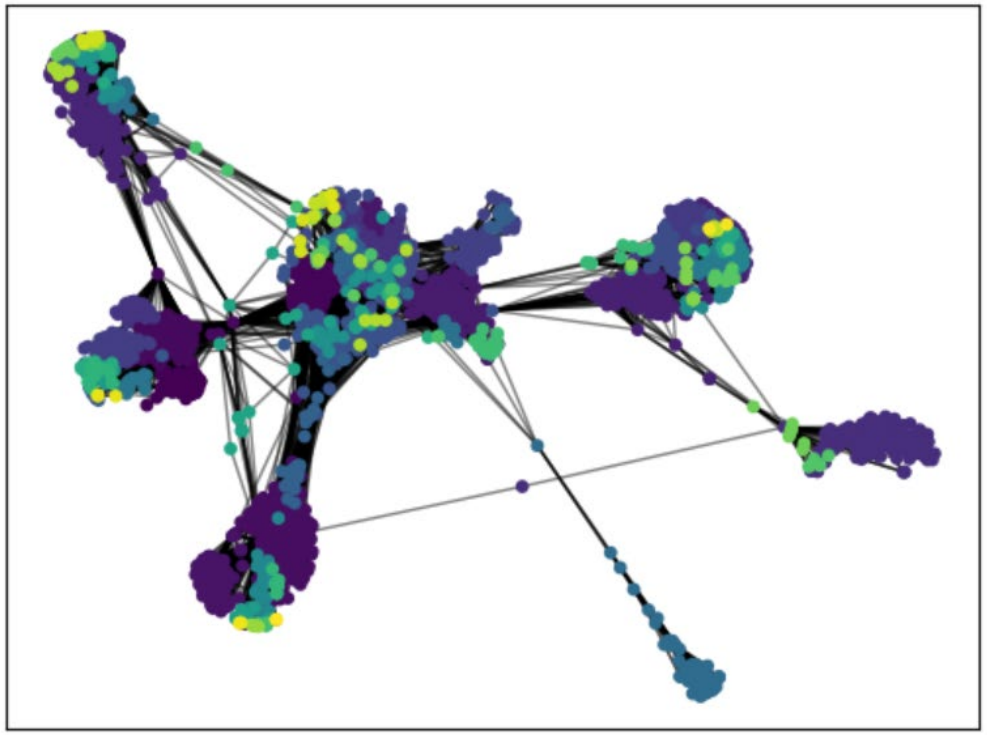
贪心加缩点 优化信息熵

[1] Rosvall M, Axelsson D, Bergstrom C T. The map equation[J]. The European Physical Journal Special Topics, 2009, 178(1): 13-23.

Infomap

Infomap on Facebook-combined

基于随机游走的社区发现

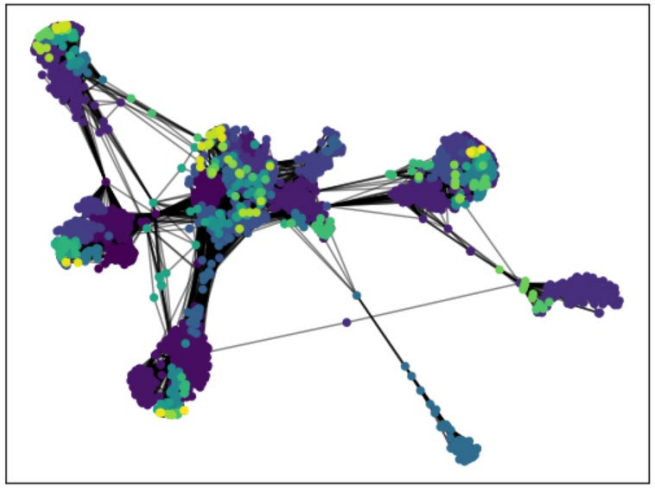


社区发现

Community Detection

Louvain

Infomap



	Louvain	Infomap
社区数	16	78
模块度	0.83	0.81

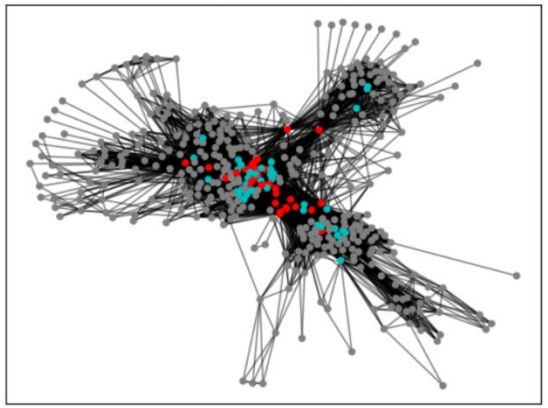
PART TWO

网络分析

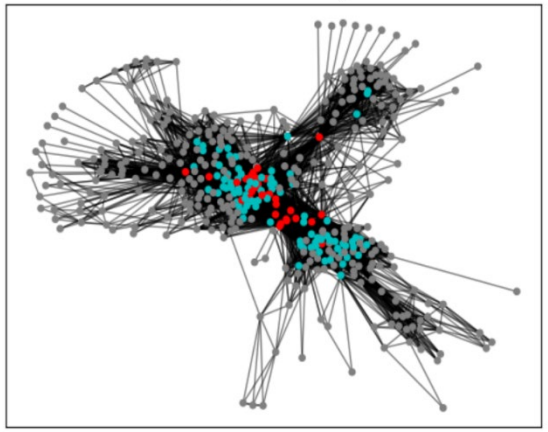
中心性度量

Centerness

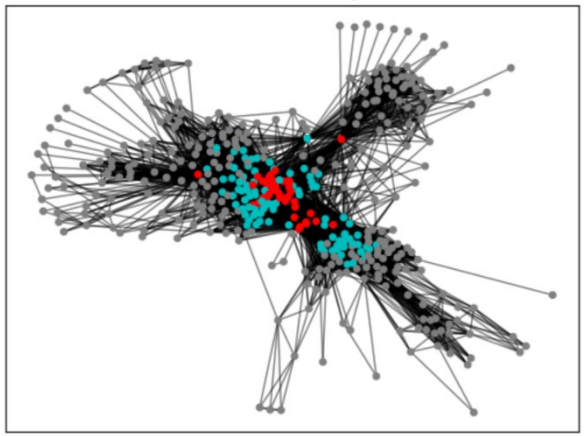
pagerank centrality



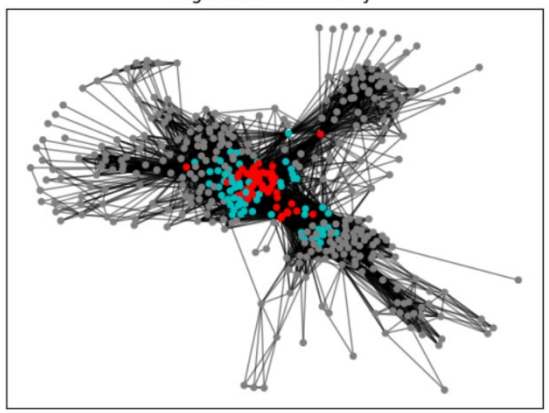
degree centrality



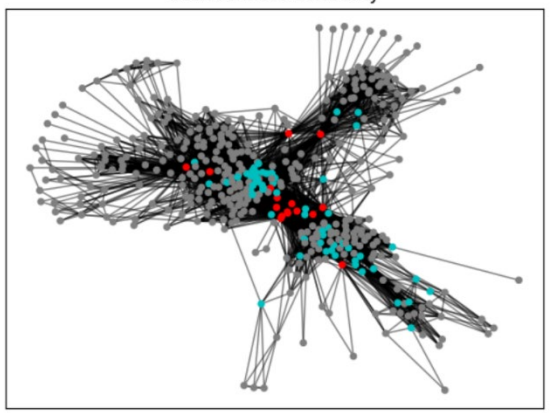
Katz centrality



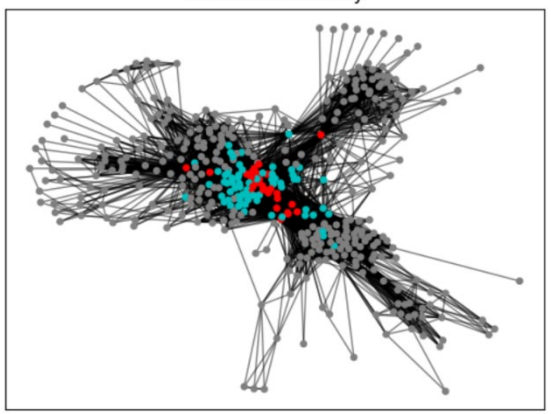
eigenvector centrality



betweenness centrality



closeness centrality



各种属性度量

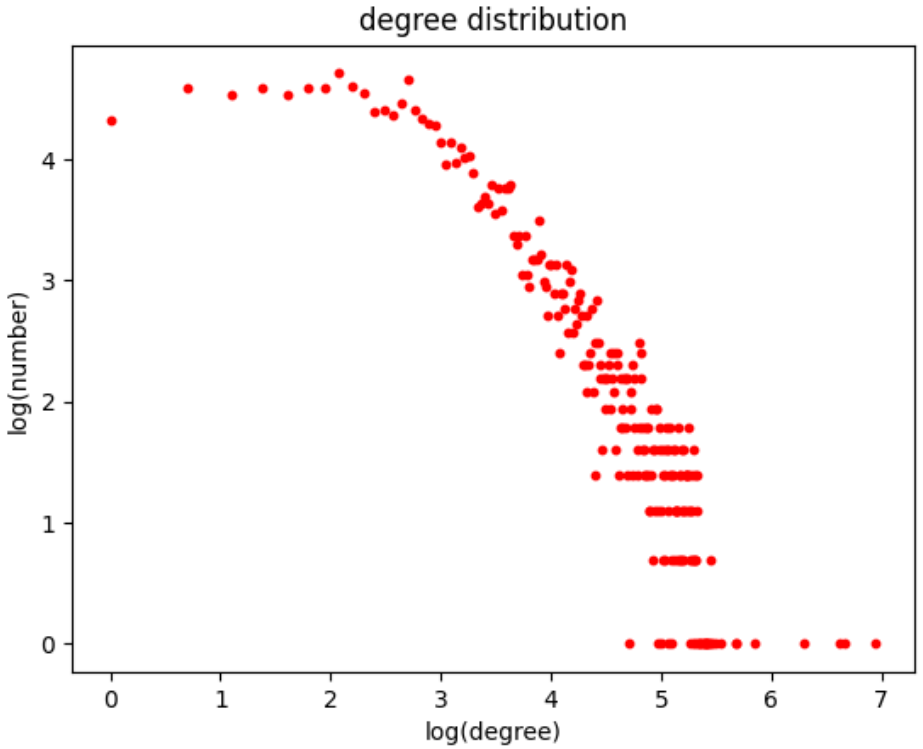
Other Metrics

平均距离、直径及平均局部聚类系数

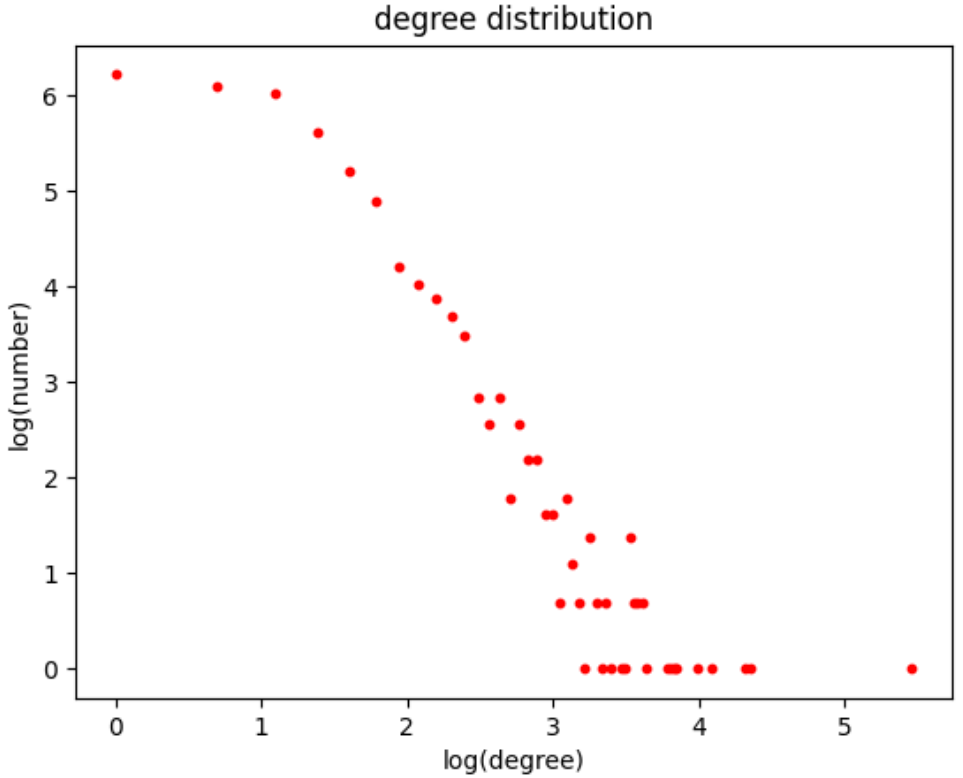
	平均距离	直径	聚类系数
Facebook	3.69	8	0.606
优先链接模型	4.17	9	0.015
小世界模型	9.03	19	0.648
随机图	4.42	9	0.003
完全图	1	1	1

度分布

Degree Distribution



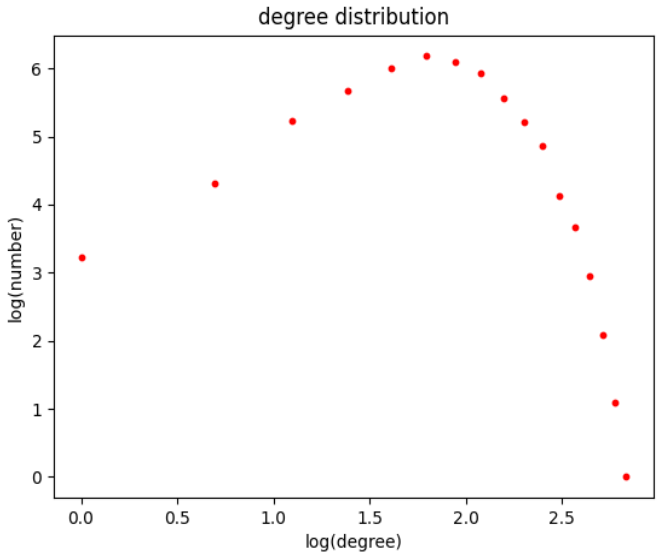
Facebook



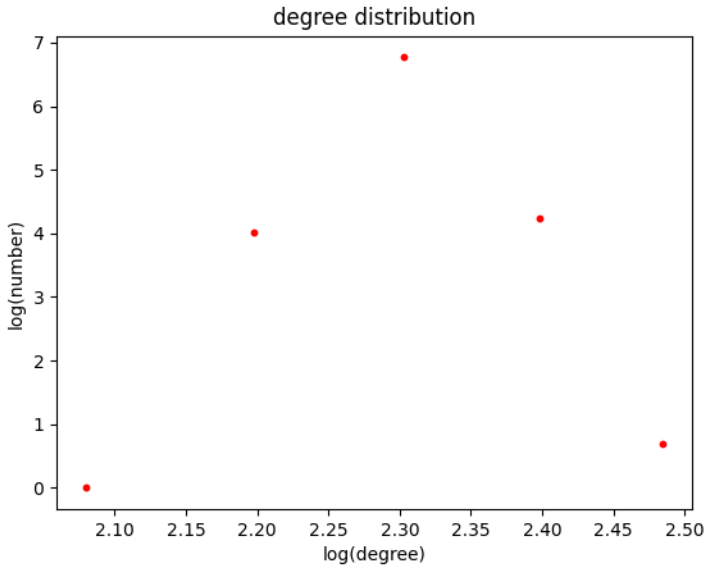
优先链接模型

度分布

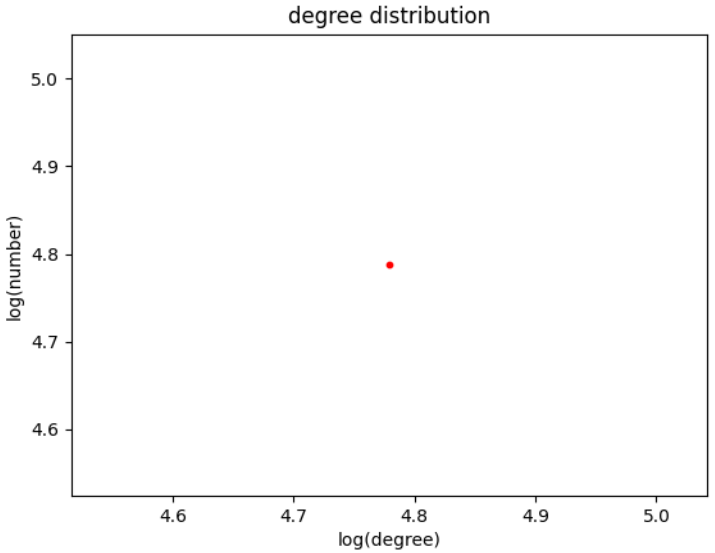
Degree Distribution



随机图



小世界模型



完全图

网络演化

Network Evolution

Ego-facebook + 链接预测

时间戳	平均距离	直径	聚类系数
0	1.95	2	0.58
1	1.92	2	0.29
2	1.88	2	0.29
3	1.86	2	0.32
4	1.83	2	0.35

网络演化

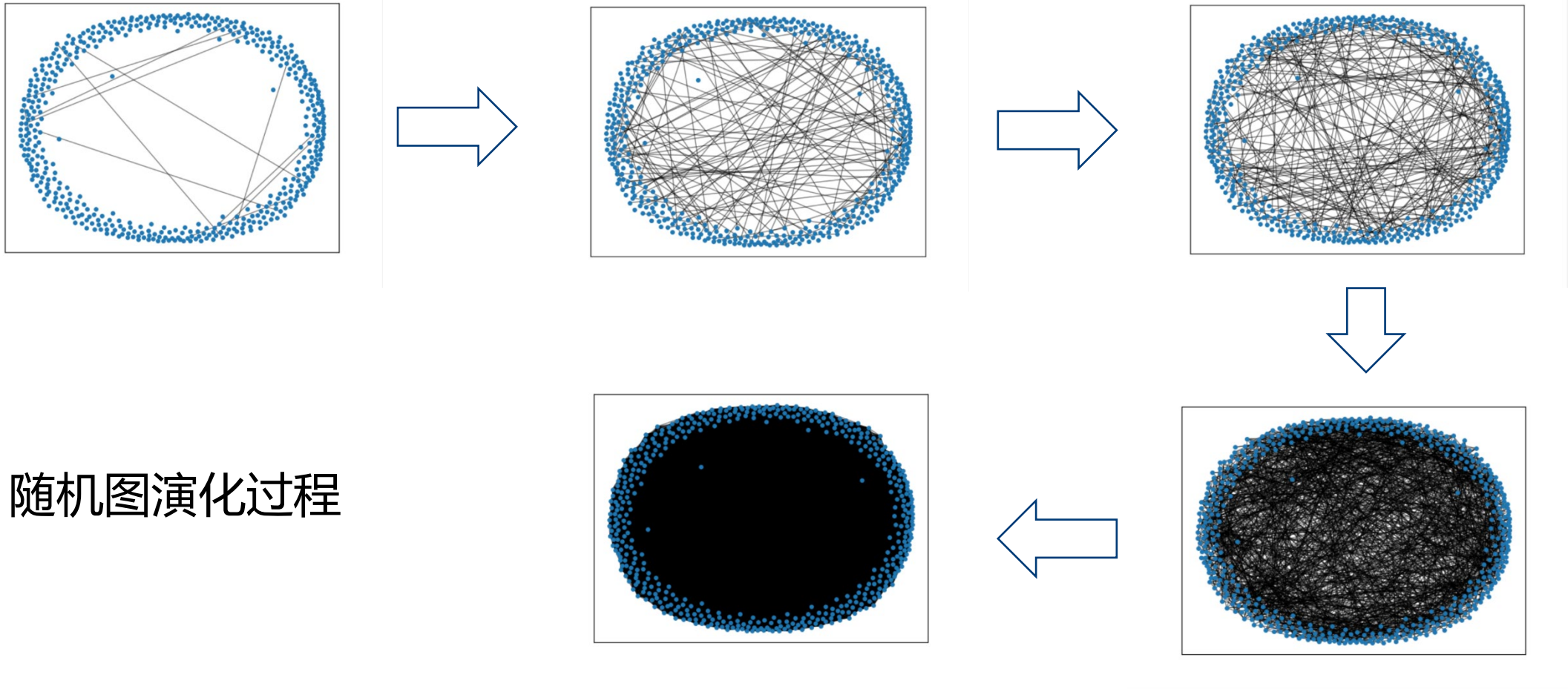
Network Evolution

随机图演化
N=500,取最大连通子图

时间戳	结点数	平均路径	直径	聚类系数	平均度
0	2	1.00	1	0.0	0.04
1	9	2.83	6	0.0	0.56
2	39	6.40	16	0.0	0.96
3	497	4.08	8	0.01	4.92
4	500	1.90	3	0.1	51.09

网络演化

Network Evolution



随机图演化过程

PART THREE

结点分类

结点分类任务介绍

Node classification

数据集: 学术引用网络 Cora（Citeseer Pubmed）

数据集介绍（以Cora为例）：

- 结点：论文（2708）
- 边：学术引用（ 5429）
- 结点特征：论文词袋（1433）
- 结点标签：论文类别（7）

任务介绍：图半监督学习

非学习方法：标签传播（LP）

Label Propagation

原理：相同类别的学术论文更倾向于相互引用

公式：$$\mathbf{Y}' = \alpha \cdot \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}\mathbf{Y} + (1-\alpha)\mathbf{Y},$$

参数设置：α=0.9，niters=30

LP in [1]	LP of ours
68	71.3

[1] Yang Z, Cohen W, Salakhudinov R. Revisiting semi supervised learning with graph embeddings[C]//International conference on machine learning. PMLR, 2016: 40-48.

基于图表示学习的方法：随机游走

Node2Vec

From Word2Vec to Node2Vec[1]
基于NLP中的Word2Vec模型，将随机游走视作上下文

基于Node2Vec获取Embedding表示，并且辅以经典的机器学习模型，如SVM、随机森林等

Node2Vec+ SVM	Node2Vec+ Random Forest	Node2Vec+ Logistic Regression
75.2	69.7	72.7

[1] Grover A, Leskovec J. node2vec: Scalable feature learning for networks[C]//Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016: 855-864.

基于图表示学习的方法：随机游走

Node2Vec

但可以从矩阵分解的角度证明，此类模型存在缺陷：

- 学习得到的Embedding具有旋转不变性 [1]
- 非端到端训练

[1]Levy O, Goldberg Y. Neural word embedding as implicit matrix factorization[J]. Advances in neural information processing systems, 2014, 27: 2177-2185.

图神经网络（GNN）

Graph Neural Network

基于图结构和结点特征，
获取结点的表示向量，用于各类任务

- 结点分类
- 链接预测

图卷积神经网络：GCN

Graph Convolution Network

将卷积运算推广至图结构
利用图信号处理中的傅里叶变换，将傅里叶变换的卷积核作为由切比雪夫多项式定义的可学习参数，可以得到切比雪夫卷积（ChebConv）

$$\mathbf{X}' = \sum_{k=1}^{K} \mathbf{Z}^{(k)} \cdot \mathbf{\Theta}^{(k)}$$

对图拉普拉斯矩阵的特征值等做出合理的近似，得到GCN

$$\mathbf{X}' = \hat{\mathbf{D}}^{-1/2}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-1/2}\mathbf{X}\mathbf{\Theta},$$

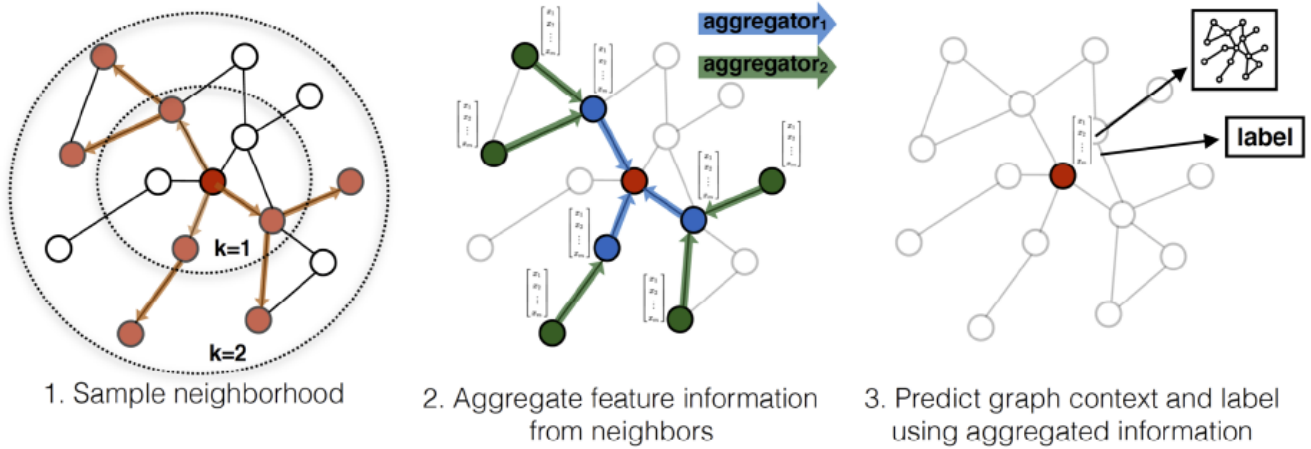
[1] Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering[J]. arXiv preprint arXiv:1606.09375, 2016.

[2] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.

其他图神经网络：GraphSAGE

Sample and Aggregate on Graph

GNN（Graph Neural Network）的本质是消息传递机制（Message Passing） 可以拆分为采样和聚合操作，



1. Sample neighborhood　2. Aggregate feature information from neighbors　3. Predict graph context and label using aggregated information

[1] Hamilton W L, Ying R, Leskovec J. Inductive representation learning on large graphs[J]. arXiv preprint arXiv:1706.02216, 2017.

其他图神经网络：GAT

Graph Attention Network

Attention is all your need!
注意力：为不同的邻居赋予不同的权重

$$\mathbf{x}_i' = \alpha_{i,i}\mathbf{\Theta}\mathbf{x}_i + \sum_{j \in \mathcal{N}(i)} \alpha_{i,j}\mathbf{\Theta}\mathbf{x}_j,$$

权重的计算和Embedding相似，

$$\alpha_{i,j} = \frac{\exp\left(\mathrm{LeakyReLU}\left(\mathbf{a}^\top [\mathbf{\Theta}\mathbf{x}_i \,\|\, \mathbf{\Theta}\mathbf{x}_j]\right)\right)}{\sum_{k \in \mathcal{N}(i)\cup\{i\}} \exp\left(\mathrm{LeakyReLU}\left(\mathbf{a}^\top [\mathbf{\Theta}\mathbf{x}_i \,\|\, \mathbf{\Theta}\mathbf{x}_k]\right)\right)}.$$

[1] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[J]. arXiv preprint arXiv:1710.10903, 2017.

图神经网络：实验结果

Graph Neural Network

GCN	SAGE	GAT
81.1	79.6	81.19

选用GCN作为基础模型：

- 理论推导严谨
- 模型表现相对满意
- 自实现了该模型

使用优化器改进：K-FAC

Kronecker-factored Approximate Curvature

- 基于Fisher信息矩阵的自然梯度的二阶优化
- 对于交叉熵损失，Fisher信息矩阵是Hessian矩阵的期望
- 对于经验Fisher信息矩阵，可以使用基于Kronecker积的近似矩阵分解在低复杂度内计算矩阵
- 在合理的独立性等假设下，经验Fisher信息矩阵是分块对角的，可以在低复杂度内求逆

GCN	GCN + K-FAC
81.1	81.7

[1] Martens J, Grosse R. Optimizing neural networks with kronecker-factored approximate curvature[C]//International conference on machine learning. PMLR, 2015: 2408-2417.

使用优化器改进：SAM

Sharpness Awareness Minimization

Sharpness：泛化误差
SAM:根据PAC贝叶斯泛化误差理论上界设计相应的优化器

将SAM运用于GNN中

GCN	GCN + SAM
81.1	82.1
GAT	**GAT + SAM**
81.9	82.3

[1] Foret P, Kleiner A, Mobahi H, et al. Sharpness-Aware Minimization for Efficiently Improving Generalization[J]. arXiv preprint arXiv:2010.01412, 2020.

更深的网络

Deeper Graph Neural Networks

由于过平滑化效应（OverSmoothness），更深的GNN不一定更好
根据六度空间理论，也不需要过深的GNN

过平滑化理论[1]：根据GNN和动力系统的关系，模型学到的embedding会收敛到某一子空间

GCN2	GCN4	GCN8	GCN16
81.1	62.5	56.4	15.3

[1] Oono K, Suzuki T. On asymptotic behaviors of graph cnns from dynamical systems perspective[J]. 2019.

更深的网络

Deeper Graph Neural Networks

训练更深的网络：

- DropEdge[1]：减缓过平滑化效应，从随机游走的角度证明[2]

GCN2	GCN4	GCN8
81.1	62.5	56.4
GCN2+DropEdge	**GCN4+DropEdge**	**GCN8+DropEdge**
82.4	69.5	65.2

[1] Rong Y, Huang W, Xu T, et al. Dropedge: Towards deep graph convolutional networks on node classification[J]. arXiv preprint arXiv:1907.10903, 2019.

[2] Lovász L. Random walks on graphs: A survey[J]. Combinatorics, Paul erdos is eighty, 1993, 2(1): 1-46.

更深的网络

Deeper Graph Neural Networks

训练更深的网络：

- DropEdge：减缓过平滑化效应，
- 残差连接（Jump Knowledge Network [1]）

GCN4	GCN8
62.5	56.4
JK+DropEdge	**JK+DropEdge**
82.9	82.8

JumpKnowledge、DropEdge、DropOut相辅相成
使用JumpKnowledge+DropEdge可以令DropOut概率为0.9

[1] Xu K, Li C, Tian Y, et al. Representation learning on graphs with jumping knowledge networks[C]//International Conference on Machine Learning. PMLR, 2018: 5453-5462.

图神经网络的本质

Correct and Smooth

本质类似一种迭代法[1]
Correct：减小训练集误差
Smooth：平滑Embedding（GNN的初始化trick）

在简单的MLP上加入Correct and Smooth可以大幅提升性能
使用Correct and Smooth作为后处理手段优化GCN的结果

MLP	MLP+Correct and Smooth
51.8	72.9

GCN	GCN+SAM	GCN+SAM+Correct and Smooth
81.1	82.1	83.0

[1] Huang Q, He H, Singh A, et al. Combining Label Propagation and Simple Models Out-performs Graph Neural Networks[J]. arXiv preprint arXiv:2010.13993, 2020.

实验结果

Experiments

Method	Detail	Accuracy
Deeper GCN	GCN4+DropEdge+ JKNet + SAM	82.9
GCN with post process	GCN2+SAM+ Correct and Smooth	83

PART FOUR

链接预测

链接预测

Link Prediction

基于Embedding
转化为二分类问题，给定任意两个点对，判断边的有无（0/1二分类）
真实图的稀疏性：使用负采样技术保证正负样本的均衡

图自编码器（GAE）

Graph Auto Encoder

思想：重构原图

编码器：GCN等
解码器：简单的点积编码器等

[1] Kipf T N, Welling M. Variational graph auto-encoders[J]. arXiv preprint arXiv:1611.07308, 2016.

变分自编码器（VGAE)

Variational Graph Auto Encoder

将自编码器看作含隐变量的参数推断
使用EM算法进行参数推断，并使用神经网络拟合隐变量的分布等
可以看作GAE加入KL散度作为正则项

[1] Kipf T N, Welling M. Variational graph auto-encoders[J]. arXiv preprint arXiv:1611.07308, 2016.

球面变分自编码器（S-VGAE)

Sphere - VGAE

解决一般VGAE使用正态分布的KL崩塌问题，
使用vMF分布[1,2]（von Mises-Fisher Distribution）替代正态分布
通过控制凝聚度超参数κ使得KL散度具有正下界

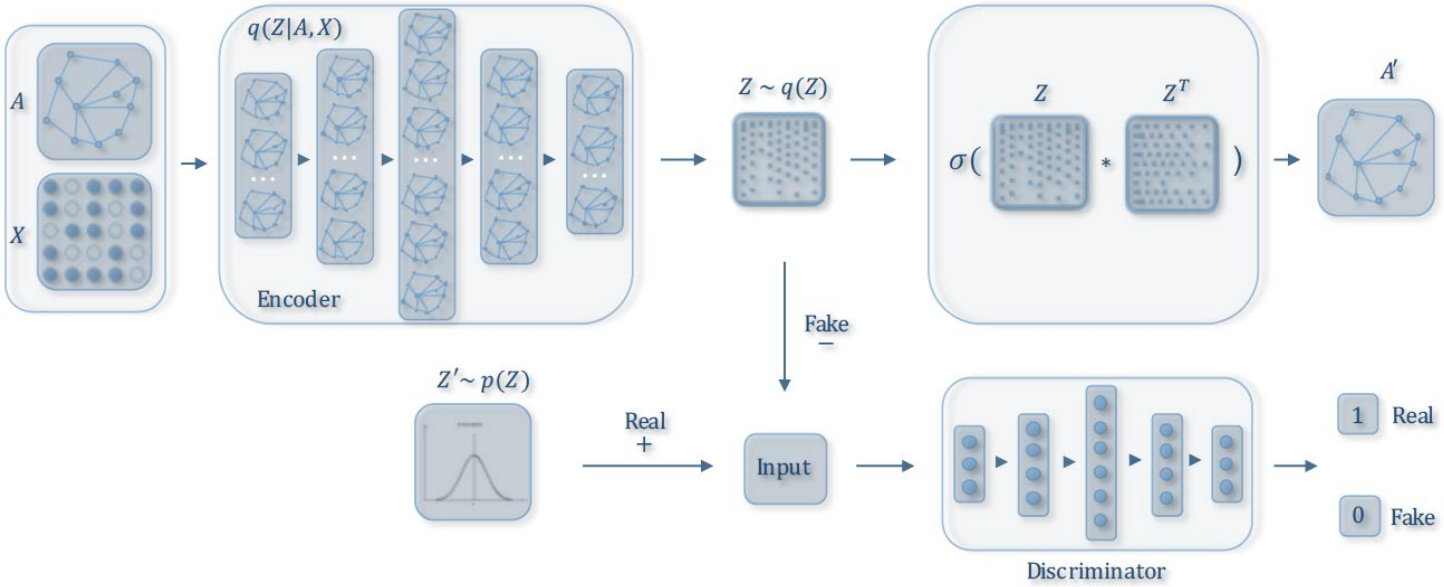
[1] Xu J, Durrett G. Spherical latent spaces for stable variational autoencoders[J]. arXiv preprint arXiv:1808.10805, 2018.

[2] Davidson T R, Falorsi L, De Cao N, et al. Hyperspherical variational auto-encoders[J]. arXiv preprint arXiv:1804.00891, 2018.

对抗正则变分自编码器（ARVGAE）

Adversarial Regulated VGAE

KL散度对判断分布距离存在问题
借鉴GAN的思想，使用网络Discriminator鉴别分布的距离（W-GAN）



[1] Adversarially Regularized Graph Autoencoder for Graph Embedding

实验结果

Experiments

	AUC	AP
GAE	91.14	91.28
VGAE	90.56	91.56
ARVGAE	92.5	92.9
S-VGAE	92.88	93.1

THANKS

THANKS AGAIN

小组分工

Teamwork

姜俊哲：社区发现，网络分析

陈乐偲：结点分类，链接预测

于淼芃：网络生成，部分数据集训练（Citeseer, Pubmed）

节标题

01

Warm-up

02

OHEM

03

Auxiliary Loss

节标题

内容

文本