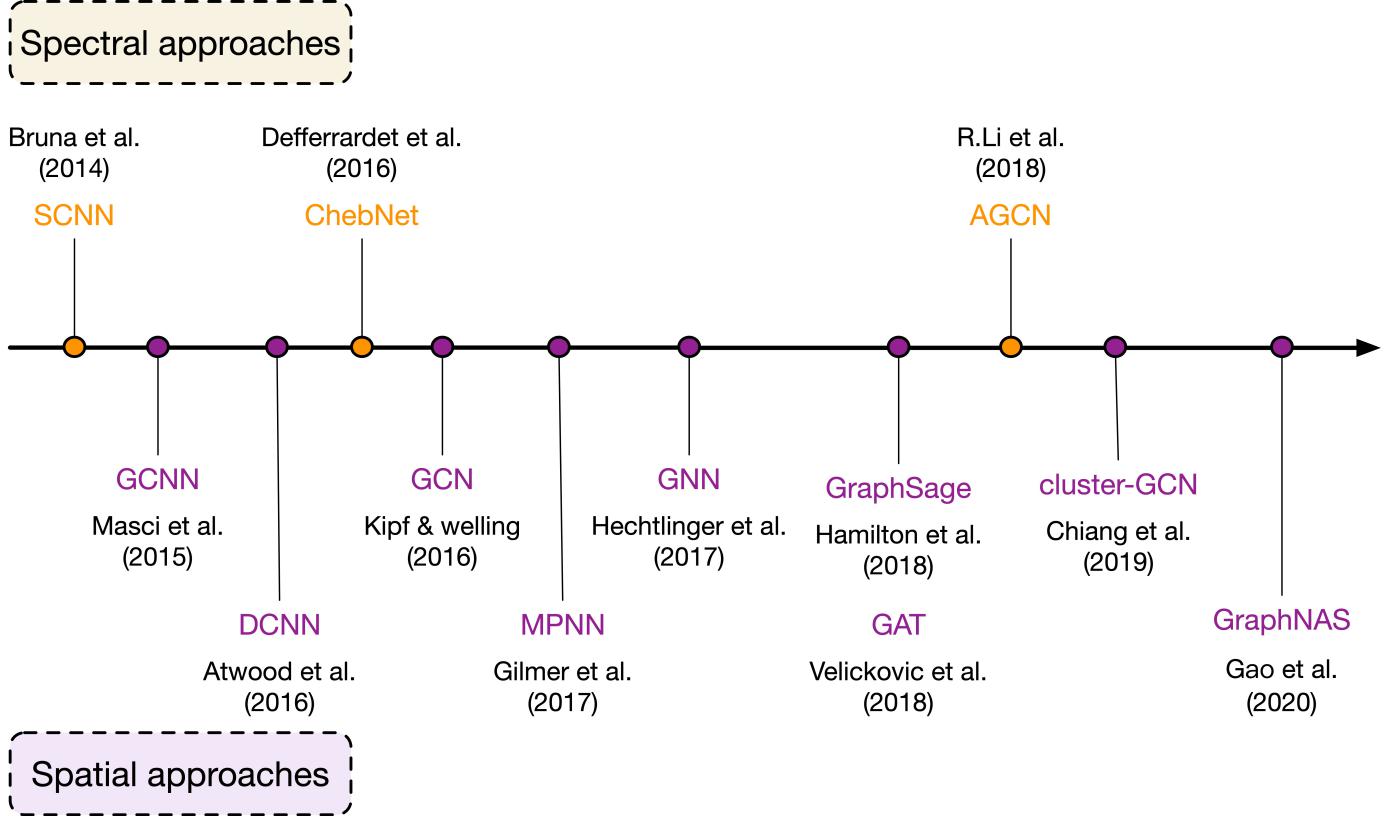
正 文

# 一、项目背景和意义

图（Graph）是一种表示对象之间关联关系的抽象数据结构，包含节点集合和边集合以及它们的属性特征。大规模图数据可以表达丰富的蕴含逻辑关系的人类常识和专家知识。图神经网络可以将不同来源、不同类型的数据融合到一个图里进行分析，从而得到分析孤立数据难以发现的结果。

1.1图神经网络的发展和应用

神经网络在很多任务上已经取得了重要的突破，例如图像分类、视频处理、语音识别等。这些任务处理的通常是欧式空间数据，然而在很多应用中涌现了越来越多以图的形式出现的非欧式空间数据。为了更好地在分析非欧式图数据，图神经网络成为当前的一个研究热点并被广泛应用到工业界[1]。



**图1 图神经网络研究的演进**

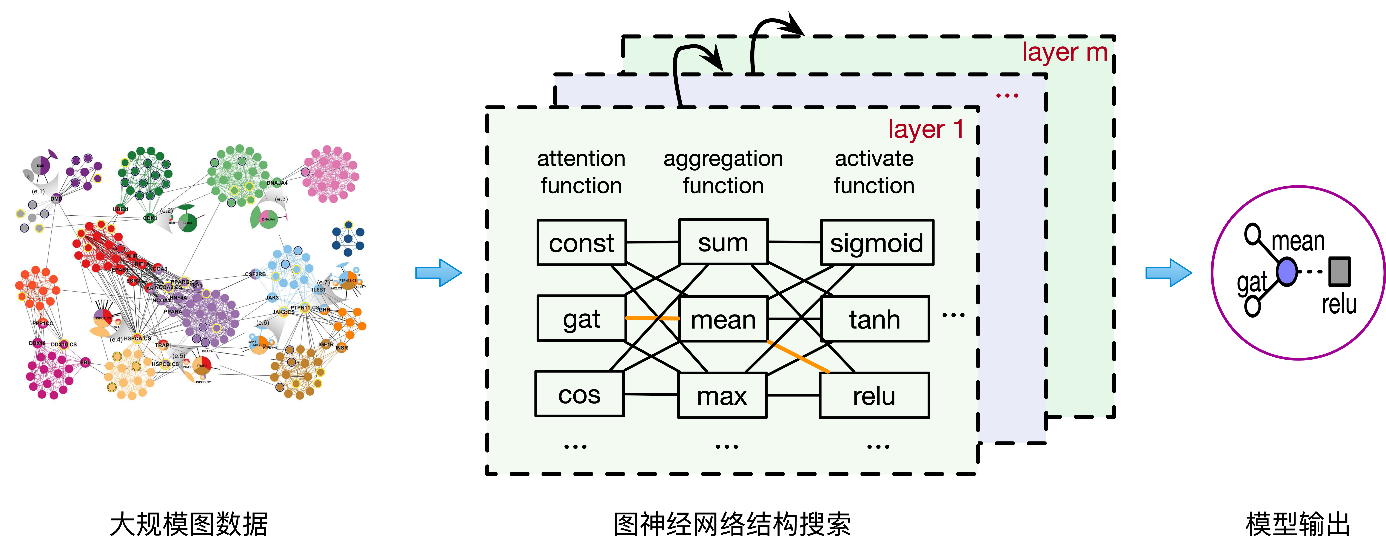
图1展示了图神经网络研究演进的部分代表性工作。较早的谱域图神经网络SCNN[2]利用图的傅里叶变换设计卷积核，ChesNet[3]通过切比雪夫多项式近似SCNN给出的卷积核设计，降低了时间复杂度，2018年提出的AGCN[4]是对ChebNet的改进，进一步加快了谱域图神经网络的建模速度。由于谱域图神经网络的计算代价高、卷积核定义不灵活等缺点，目前图神经网络的研究聚焦于空域图神经网络方法。GCNN[5]、DCNN[6]是早期的空域图神经网络方法，2016年提出的GCN空域图神经网络[7]利用切比雪夫多项式的一阶近似设计卷积核，实现了一种建模速度更快的模型结构，随后提出的MPNN[8]、GNN[9]图神经网络是对GCN图神经网络的改进，随着2018年GAT[10]提出带权重的邻居信息聚合方式，GraphSAGE[11]提出邻居节点采样机制，空域图神经网络方法基于具体的任务和图数据设计的图神经网络结构实现了较快的建模速度与建模准确度。

随着图数据的规模增大，大规模图数据上数据特征分布越来越复杂，基于图数据特征自动搜索出适配数据特征的图神经网络结构，并基于分布式方法在大规模图数据上训练单个图神经网络，成为了在大规模图上加速图神经网络建模过程的研究重点。2019年谷歌提出Cluster-GCN[12]，基于图划分思想对大规模图数据进行分割，为图神经网络难以处理大规模图提出了解决方案，但在不同数据集上，图划分的方式依然会影响建模的准确度，2020年阿里研究团队提出了GraphNAS[13]图神经网络结构搜索方案，设计了一种集中式的图神经网络结构的搜索方式，但还没有实现并行化搜索过程的加速。

图神经网络在不同领域都得到了广泛的应用,例如,电商网络中欺诈用户的检测[14]、自然语言处理中关系抽取[15]、社交网络中链路预测[16]和节点聚类[17]、生物医疗中药物副作用的预测[18]。我们以下几个典型应用进行介绍。金融风险评估。金融行业内部存在大量风控需求，如反洗钱、防身份欺诈、防车险骗保、防金融欺诈、信用卡伪造交易套现等。图神经网络的拓展性和线上预测能力能够根据资金交易关系网络构建动态图模型，发现个体或群体的异常交易行为，对金融业务中的每一笔业务进行风险预测。（2）电商应用。伴随着电商用户的快速增长，平台上用户欺诈行为增多，包括刷单炒信、虚假评价、风险流量等，这些欺诈行为严重影响搜索排序算法的准确率，成为电商生态的重要威胁。阿里巴巴已经利用图神经网络进行欺诈用户的检测，取得了良好的效果。（3）自然语言处理。在自然语言处理中，包括多跳阅读[19]、实体识别[20]、关系抽取[15]以及文本分类[21]都可以运用图神经网络解决。（4）生物医疗。化合物是由原子和化学键构成的，它们天然就是一种图数据的形式，所以图神经网络在生物医疗领域应用特别广泛。包括新药物的发现[22]、化合物筛选[23]、蛋白质相互作用点检测[24]、以及疾病预测[25]。目前国内外已经有很多实验室研究图神经网络在医学方面的应用，这将是图神经网络最有价值的应用方向之一。

在除了上述的应用领域外，包括交通预测[1]、程序推断[26]、优化求解[27]等任务上，图神经网络都开始被人们使用。由于其可以建模在现实生活中常见的图数据，并且通过卷积、注意力或消息传播等机制，能够将网络的拓扑结构和节点属性等信息以神经网络进行捕获和建模，因此图神经网络有广泛的应用前景及重要的研究意义。

1.2大规模图神经网络建模面临的挑战



**图 2 图神经网络建模过程**

一个图神经网络结构可以分为多个组成部分[1]。图2给出了一个示例，列出图神经网络每一层（layer）中包含的注意力机制的类型、聚合方式的类型、激活函数等组成部分。图神经网络建模的目标是在每一个组成部分中选择合适的组件形成一个高效的图神经网络结构。图神经网络建模能根据不同的图数据特征调整图神经网络结构达到建模准确性的要求，很好的适应了图数据分布的复杂多样性。随着图数据规模不断增加，大规模图数据上图神经网络建模面临着复杂度过高的巨大挑战。

目前没有一种通用的图神经网络结构能在不同的图数据上保证建模的准确性，需要基于具体的图数据针对性地设计图神经网络结构并训练图神经网络评估模型准确性。然而，图神经网络建模是一个特别耗时的过程，以六个组成部分构成的两层图神经网络结构为例，不同图神经网络结构的组合规模­量级[13]。即使单个图神经网络训练的时间复杂度也比较高，训练一个GCN[7]网络结构的时间复杂度和空间复杂度分别为和。其中L为图卷积层层数，为邻接矩阵非0元素个数，F为节点特征数，N为节点个数[12]。在图神经网络结构搜索过程中，需要训练大量不同结构的模型，人工调整或穷举法搜索结构的方式工作量巨大且十分耗时，无法在实际应用中的大规模图数据上图神经网络的建模过程[28]。

高效、自动、并行地从图神经网络结构空间中搜索出适配不同图数据特征的模型结构，成为了在大规模图数据上加速图神经网络建模过程的关键挑战。

1.3本项目组的图数据研究工作

本项目组近年来一直着力于研究图数据处理相关理论和应用，在大规模图数据计算、高效图处理算法、图神经网络应用等方面取得了一系列成果。  
 为了更加高效的处理大规模图数据，我们提出了分布式图查询算法[29]、增量的图匹配算法[30、31]以及Top-K的子图查询算法[32]。在最近的研究工作中，我们将重点对图神经网络进行了深入研究和应用。例如，我们提出了一种TypedGAT图神经网络实现了跨语言实体对齐。图数据隐含了复杂多样的特征，多个图数据关联的背后往往蕴含着单个图数据所不具备的知识，我们通过在多种图数据进行特征挖掘，研究跨语言知识图谱的实体对齐[33]。在生物网络中，我们提出了基于图神经网络的数据挖掘方法，例如多模态数据融合预测癌症患者预期寿命[34]、蛋白质相互作用网络对齐方法[35、36]。在社交网络应用中，我们提出了K-匿名的图结构保护方法[37-39]。在金融股票应用中，我们提出了动态图神经网络进行股票预测[40]。  
 传统的图神经网络在面临大规模图数据时，需要对图神经网络结构进行建模以提高其适应度和灵活度。然后，大规模图神经网络建模加速技术研究尚处于起步阶段，无论是理论模型还是应用技术方面都存在许多关键问题需要进一步研究解决。

**参考文献：**

[1] Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[2] Estrach J B, Zaremba W, Szlam A, et al. Spectral networks and deep locally connected networks on graphs. *2nd International Conference on Learning Representations*, ICLR 2014.

[3] Tang S, Li B, Yu H. ChebNet: Efficient and Stable Constructions of Deep Neural Networks with Rectified Power Units using Chebyshev Approximations. *arXiv preprint* arXiv:1911.05467, 2019.

[4] Li R, Wang S, Zhu F, et al. Adaptive graph convolutional neural networks. *AAAI Conference on Artificial Intelligence*. 2018:1801.03226.

[5] Masci J, Boscaini D, Bronstein M, et al. Geodesic convolutional neural networks on riemannian manifolds. *Proceedings of the IEEE international conference on computer vision workshops*. 2015: 37-45.

[6] Atwood J, Towsley D. Diffusion-convolutional neural networks. *Advances in neural information processing systems*. 2016: 1993-2001.

[7] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. *5th International Conference on Learning Representations*, ICLR 2017:1–14.

[8] Gilmer J, Schoenholz S S, Riley P F, et al. Neural message passing for quantum chemistry. *Proceedings of the 34th International Conference on Machine Learning*. 2017: 1263-1272.

[9] Hechtlinger Y, Chakravarti P, Qin J. A Generalization of Convolutional Neural Networks to Graph-Structured Data. *STAT*, 2017, 1050: 26.

[10] Veličković P, Cucurull G, Casanova A, et al. Graph Attention Networks. *International Conference on Learning Representations*. 2018.

[11] Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. *Advances in neural information processing systems*. 2017: 1024-1034.

[12] Chiang W L, Liu X, Si S, et al. Cluster-GCN: An efficient algorithm for training deep and large graph convolutional networks. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019: 257-266.

[13] Gao Y, Yang H, Zhang P, et al. Graphnas: Graph neural architecture search with reinforcement learning. *Proceedings of the 25th International Joint Conference on Artificial Intelligence*(IJCAI).2020:1403-1409.

[14] Wang H, Li Z, Huang J, et al. Collaboration based multi-label propagation for fraud detection. *Proceedings of the 29th International Joint Conference on Artificial Intelligence*. 2020:2477-2483.

[15] Nathani D, Chauhan J, Sharma C, et al. Learning Attention-based Embeddings for Relation Prediction in Knowledge Graphs. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019: 4710-4723.

[16] Zhang M, Chen Y. Link prediction based on graph neural networks. *Advances in Neural Information Processing Systems*. 2018: 5165-5175.

[17] Park H, Neville J. Exploiting Interaction Links for Node Classification with Deep Graph Neural Networks. *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 2019: 3223-3230.

[18] Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 2018, 34: i457–i466

[19] Qiu L, Xiao Y, Qu Y, et al. Dynamically fused graph network for multi-hop reasoning. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019: 6140-6150.

[20] Gui T, Zou Y, Zhang Q, et al. A lexicon-based graph neural network for Chinese ner. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP). 2019: 1039-1049.

[21] Hu L, Yang T, Shi C, et al. Heterogeneous graph attention networks for semi-supervised short text classification. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP). 2019: 4823-4832.

[22] Lin X, Quan Z, Wang Z J, et al. Kgnn: Knowledge graph neural network for drug-drug interaction prediction. *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, 2020: 2739-2745.

[23] De Cao N, Kipf T. MolGAN: An implicit generative model for small molecular graphs. *the ICML 2018 workshop on Theoretical Foundations and Applications of Deep Generative Models*, Stockholm, Sweden,PMLR 80, 2018.

[24] Fout A, Byrd J, Shariat B, et al. Protein interface prediction using graph convolutional networks. *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017:6530-6539.

[25] Chen Huiyuan, and Jing Li. Learning Data-Driven Drug-Target-Disease Interaction via Neural Tensor Network. *Proceedings of the 29th International Joint Conference on Artificial Intelligence*. 2020:3452-3458.

[26] Khalil, Elias, et al. Learning combinatorial optimization algorithms over graphs. *Advances in Neural Information Processing Systems*. 2017:6348-6358.

[27] Li Z, Chen Q, and Vladlen Koltun. Combinatorial optimization with graph convolutional networks and guided tree search. *Advances in Neural Information Processing Systems*. 2018:539-548.

[28] 王靖, 张路, 王鹏宇, 等. 面向图计算的内存系统优化技术综述. *中国科学: 信息科学*, 2019, 49: 295-313.

[29] Gao J, Liu P, Kang X, et al. PRS: Parallel Relaxation Simulation for Massive Graphs. *Computer Journal*, 2016, 59(6), pp. 848-860.

[30] Zhang L, and Gao J. Incremental Graph Pattern Matching Algorithm for Big Graph Data. *Scientific Programming*, 2018, 2018(6749561), pp. 1-8.

[31] 张丽霞, 王伟平, 高建良,王建新. 面向模式图变化的增量图模式匹配. *软件学报*, 2015. 26(11), pp. 2964-2980.

[32] Gao J, Song B, Liu P, et al. Parallel top-k subgraph query in massive graphs: Computing from the perspective of single vertex. *IEEE International Conference on Big Data*, 2016, pp. 636-645,

[33] Xiong F, and Gao J. Entity Alignment for Cross-lingual Knowledge Graph with Graph Convolutional Networks. *The 28th International Joint Conference on Artificial Intelligence,* 2019, pp. 6480-6481.

[34] Gao J, Lyu T, Xiong F, et al. MGNN: A Multimodal Graph Neural Network for Predicting the Survival of Cancer Patients. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020: 1697-1700.

[35] Gao J, Song B, Ke W, et al, BalanceAli: Multiple PPI Network Alignment With Balanced High Coverage and Consistency, *IEEE Transactions on NanoBioscience*, 2017, 16(5), pp. 333-340.

[36] Song B, Gao J, Du H, et al. Aligning Multiple PPI Networks with Representation Learning on Networks. *IEEE International Conference on Bioinformatics and Biomedicine*, 2018, pp.136-141.

[37] Gao J, Song B, Z. Chen, et al. Counter deanonymization query: H-index based K-anonymization privacy protection for social networks. *The 40th International Conference on Research and Development in Information Retrieval*, 2017, pp. 809-812 .

[38] Gao J, Ping Q, and Wang J. Resisting re-identification mining on social graph data. *World Wide Web*, 2018, 21(6), 1759–1771.

[39] Gao J, Wang J, He J, et al. Against Signed Graph Deanonymization Attacks on Social Networks. *International Journal of Parallel Programming*, 2017,47(4),pp.725-739.

[40] Ying X, Xu C, Gao J, et al. Time-aware Graph Relational Attention Network for Stock Recommendation. *ACM International Conference on Information & Knowledge Management* (CIKM),2020:2281–2284.

# 拟解决的关键科学和技术问题

针对图神经网络建模在大规模图中应用的难题，本项目拟解决的如下关键科学和技术问题：

**（1）如何缩小大规模图神经网络建模结构空间加速结构搜索过程**

图神经网络建模过程是自适应搜索图神经网络各个组件的最优组合，其搜索空间规模与图神经网络层数成指数关系。在大规模图中训练一个已确定结构的图神经网络都是非常耗时的过程，因此，在实际大规模图应用中无法遍历评估搜索空间中每一个图神经网络结构。如何缩小大规模图神经网络建模搜索空间是本项目拟解决的一个关键科学问题。在本项目中，我们拟提出基于强化学习的初始图神经网络结构搜索算法来解决图神经网络建模搜索空间巨大的挑战。

**（2）如何实现大规模图神经网络建模的并行计算加速建模过程**

图神经网络在大规模图数据上的应用面临计算效率低的问题，图神经网络建模增加了模型的灵活性和适应度，但是对图神经网络的结构搜索算法的效率提出了更高的要求。因此，如何实现大规模图神经网络建模的并行计算是本项目需要解决的另一个关键科学与技术问题。实现大规模图神经网络建模的并行计算，一方面需要并行化图神经网络结构搜索过程，即通过分布式搜索算法提高搜索效率；另一方面需要加速单个模型的训练过程。在本项目中，我们拟提出基于初始图神经网络结构，使用进化学习来并行化图神经网络结构的搜索过程，并通过适配不同类型图数据的图划分策略并行化单个模型的训练过程。

# 三、研究目标和内容

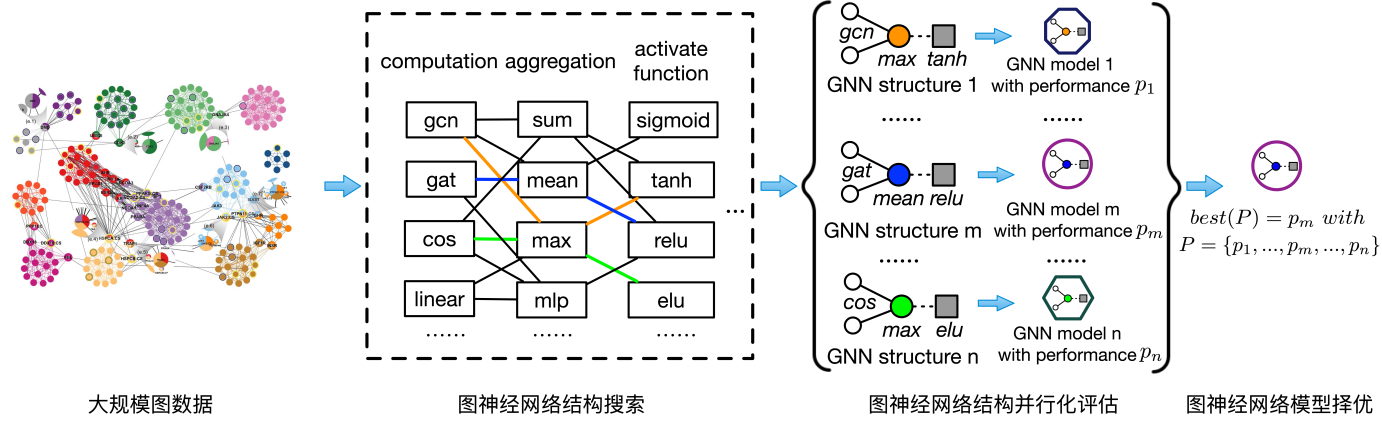
3.1研究目标

**（1）基础理论创新**。图神经网络正成为挖掘图数据背后丰富知识的有效工具，但是面对规模越来越大的图数据，如何加速图神经网络建模过程是大规模图神经网络进一步发展的一个关键瓶颈。本项目将研究大规模图神经网络建模加速技术，力求在基础理论创新上取得突破。首先研究面向图神经网络建模加速的存储优化技术；然后重点研究大规模图神经网络的建模加速，包括基于强化学习与进化学习的并行图神经网络结构搜索算法与基于图划分算法的并行图神经网络训练技术，实现建模过程的分布式计算，有效解决建模过程难以加速的问题。

**（2）应用创新**。本项目提出的基础理论创新成果将在实际的大规模图数据中得以验证和应用。本项目由中南大学高建良教授和阿里巴巴集团人工智能技术总监李朝领衔组成，在长期合作中发现企业大规模图数据处理中需求和瓶颈所在。因此，本项目的另一个目标是将图神经网络建模加速的理论成果应用到企业大规模图数据中，在十亿节点规模的图数据中检验和应用我们的理论成果。我们也计划将本项目研究成果开源集成到GraphScope中，推进我国大规模图计算的发展。

3.2研究内容

在数据爆发式增长的背景下，如何自适应的针对大规模数据获得最有效的图神经网络结构已经开始得到工业界的关注。由于图数据的强耦合性，大规模图神经网络的加速面临着巨大挑战。如图3所示，本项目研究内容将从存储优化、结构搜索加速、搜索与建模并行化等方面展开。



**图3 基于大规模图数据的图神经网络加速**

3.2.1面向图神经网络建模加速的存储优化

在处理大规模图数据时，通常无法将全部图数据放入内存，而图数据的耦合性导致了频繁的访存操作。在图神经网络中，邻居采样和消息聚合操作都需要频繁地访问节点的数据，对图数据的访存速度提出了更高的要求。本项目首先将面向图神经网络建模过程优化图数据的存储机制，减少图神经网络建模过程的访存开销。

**（1）持久化缓存频繁访问节点。**图的结构具有不规则性，图中通常会存在一些度很高的中心节点，中心节点及其邻居节点在图算法的执行过程中通常有更高的概率被访问。由于图神经网络在训练过程中需要在多个图卷积层之间迭代地聚合每个节点的邻居信息，我们提出判别频繁访问节点的算法，在开始建模前，预先缓存部分将被频繁访问的节点，并在图神经网络建模过程中将新出现的频繁访问节点持久化到内存中，降低图神经网络建模的访存开销。

**（2）优化属性信息的缓存机制。**图节点和边的属性通常比节点和边的序号占用更多的存储空间。我们拟研究优化属性信息的缓存机制，一方面让频繁访问的属性信息持久化在内存中，另一方面在多个节点间共享相同属性信息，特别是针对占用空间较大的属性信息。通过对图数据的存储模式进行优化，降低访存/计算比，加速图神经网络建模过程。

**（3）图划分。**大规模图数据的存储需要依赖分布式存储系统。如何针对图数据的稀疏性和强耦合性优化图的存储是本项目的主要研究内容之一。首先确定如何衡量图的局部性。根据图的局部性特点对图进行划分，同时减少跨机器通信的节点数量。研究图划分对负载均衡的影响，根据负载均衡的指标对图划分进行动态调整，从而达到存储、通信和计算的负载均衡。

3.2.2.图神经网络结构搜索加速

一个图神经网络结构可以拆分为多个组成部分，例如注意力机制的类型、邻居节点的聚合方式、激活函数的种类、多头注意力机制的数量、图神经网络隐层的维度和跳跃连接层的计算方式。图神经网络建模的目标是在每一个组成部分中选择合适的组件形成一个高效的图神经网络结构。但是，以六个组成部分为例，不同图神经网络结构的组合规模达到~量级。面对如此规模的结构空间，我们需要研究缩小图神经网络建模结构空间加速结构搜索过程的有效方法。

**（1）基于强化学习的图神经网络初始结构搜索加速。**基于不同特征的图数据集，本项目研究自动从结构空间中搜索出最佳图神经网络结构组合。强化学习是一种以任务为导向加速机器学习收敛的方法，以强化学习构建的搜索算法可以有效收缩结构空间，加速搜索过程，快速找到满足任务需要的搜索结果。

**（2）基于进化学习的图神经网络结构并行搜索加速。**进化学习是一种模拟生物进化过程的学习方法，具备非常好的并行特性，遗传算法是一种典型的进化学习算法，基于遗传算法构建的图神经网络结构搜索算法，能够借助分布式计算框架提高搜索效率。本项目将研究强化学习与遗传算法的融合点，在此基础上提出面向大规模图数据可并行的图神经网络结构搜索方案。

3.2.3基于图划分的图神经网络训练加速

在确定图神经网络模型结构后，需要对其中的每一种结构进行训练并评估其效果。本项目将从图划分的角度研究如何加速一个已确定结构的图神经网络模型的训练过程。

本项目拟提出基于图划分的Partition-GNN方法加速图神经网络模型的训练过程。核心思想是通过图划分算法将全图划分成多个子图，将图神经网络的图卷积层的邻居搜索空间限制在子图中，有效缓解邻居搜索的时间复杂度随图卷积层层数增加而指数级增长的问题，同时提高图卷积层对共享邻居嵌入表示的利用率，减少计算和冗余数据，达到降低模型时间复杂度与空间复杂度的目的。通过将图划分与图神经网络建模过程结合，在搜寻最佳图神经网络结构的过程中，提升图神经网络训练速度和模型处理大规模图的能力。

# 四、研究方案

本项目将面向大规模图数据分析的背景和挑战，采用理论与应用相结合的研究方式，紧密围绕上述提出的研究内容、研究目标和关键科学问题。以“图存储优化🡪初始结构搜索🡪并行搜索加速🡪图应用”为主线，针对大规模图神经网络建模的主要挑战，重点解决大规模图神经网络建模中的关键科学问题并在应用中加以验证。具体研究方案如图4所示。

# 之江项目框架图-怎么做v5(1)

**图4 总体方案架构图**

4.1面向图神经网络建模的存储优化方案

大图的规模给高效的图处理带来了非常大的挑战。例如一个电子商务网络可能包含数百亿个节点和数千亿条边，占用空间超过10TB。我们拟从以下三个方面研究如何提高面向图神经网络建模的访存速度。

**（1）图划分**

在分布式环境下，一个完整的大图被划分成多个子图并被分配到不同的工作节点上独立地存储。图划分的目标是最小化跨机器数据边的数量，从而降低计算过程中跨机器通信的开销。

我们拟提出一种高效的图划分算法，目标是保证划分后的子图内部联系紧密，子图之间联系稀疏，同时采用点分割策略将划分后的图数据分配到不同机器上存储，每条边只存储一次，尽管有的节点会被复制到多台机器上存储，增大了存储开销，但在对图进行基于边的计算时，如图神经网络模型中的邻居聚合操作，能够大幅降低内网通信量。

**（2）属性独立存储**

图中的节点和边通常具有属性，需要在每个机器上存储图的结构和属性信息，结构信息可以用邻接表存储，即对于每个节点v，存储它的邻居集合，然而属性信息不适合存储在邻接表里，原因包括：1）节点/边属性通常比节点/边id占用更多空间，比如节点id只需要8byte，属性可能需要1kb。2）不同节点和边的属性通常有重叠，比如在一个用户关系网络中，男性用户节点的性别都用“man”表示。

基于上述两个原因，需要分别为节点和边的属性建立和两个索引。或的每一个入口都是一个节点v或边e的属性，所以在邻接表中，对于节点u，只存储属性在中的索引，对于每条边(u,v)，只存储属性在中的索引。用平均邻居数量，用表示属性平均长度，用表示节点和边的不同属性的数量，n表示节点数量，将属性与结构信息分开独立存储将使空间复杂度从降到。

**（3）持久化缓存**

本项目拟通过本地缓存一些重要节点及其邻居来减少跨机器通信的开销。如果一个节点v经常被其他节点访问，则应在节点v所在的分区中存储v的邻居，好处是一些节点通过v来访问它们的高阶邻居的速度将大大提升。然而，如果v的邻居数量很多，为v的邻居保存多个备份也会导致很大的存储开销。为了权衡访存速度与空间占用，定义一个指标来评估每个节点的重要性，以此决定一个节点是否值得缓存。用和表示节点v的k阶入度和出度邻居，节点v的k阶重要性如下：

现实中的点通常遵循幂律分布，即只有少量节点的入度和出度很大，基于以下两个重要理论：1）如果入度和出度分布遵循幂律分布，对于任何 ，k-hop的入度和出度邻居的数量同样遵循幂律分布；2）如果入度和出度遵循幂律分布，节点的重要性同样遵循幂律分布。本项目提出对频繁访问节点进行缓存能降低图数据访存开销。

4.2初始结构搜索

图神经网络的初始结构是指经过初始搜索算法确定的图神经网络结构，这些结构在具体图数据上建模具有较好的模型准确性，但还达不到任务的要求。以具有6个组件的两层图神经网络为例，其结构空间规模的量级为，高效地从庞大的结构空间中确定初始图神经网络结构成为了提升并行结构搜索速度的基础。

**（1）基于强化学习的图神经网络初始结构搜索**

强化学习是由策略模型产生一个行动并与环境交互，从交互中获取反馈并更新参数最终得到最优策略模型的一种学习机制，它由以下几个要素组成，1.智能体：指能够通过动作与环境交互的对象，2.环境：指智能体动作作用的对象，3.动作：指所有可能作用于环境中的操作，4.状态：指可被主题感知的关于环境的信息，5.奖励：指由环境反馈给智能体描述上一个动作效果的信息。

将强化学习机制融合到图神经网络初始结构搜索中，需要拆解整个搜索过程并与强化学习的各个要素相对应，1.智能体：借鉴卷积图神经网络结构搜索的设计，使用单层或多层LSTM作为智能体模型，保证在结构的搜索过程中已搜索出的结构组件对即将搜索的结构组件的影响。2.环境：以搜索出的图神经网络结构在验证集中建模并测试模型性能为强化学习的外部环境。3.动作：以不同的图神经网络结构组件为智能体输出的动作。4.状态：以搜索过程中，当前智能体已经确定的图神经网络结构为强化学习状态。5.奖励：以智能体确定的图神经网络结构在验证集上的验证效果为奖励的一部分，通过与历史验证效果的比较，确定每一次奖励的正负性。

为了更好地训练智能体，使用奖励构建策略损失，并基于策略梯度算法训练智能体，让智能体模型更快更稳定地收敛。基于强化学习机制使用大规模图数据训练智能体，训练完成后使用智能体在庞大的图神经网络结构空间中搜索初始的结构。

**（2）基于图神经网络结构搜索的强化学习奖励机制分配研究**

强化学习中策略模型的参数更新过程与环境对动作产生的奖励紧密相关。对图神经网络结构搜索而言，奖励主要取决于策略模型确定的图神经网络结构在验证集上获取的验证效果与历史效果的比较，而在搜索过程中，每一个结构的组件的选择代表一次策略模型产生的动作。本项目将提出一种评估动作贡献度的方法，评估每一个动作对产生此次奖励的影响程度，基于影响程度对每一个动作分配合适奖励值。这将更加有效地更新策略模型参数，加快策略模型的收敛速度与鲁棒性。

4.3并行搜索加速

**（1）基于进化学习的图神经网络结构并行搜索加速**

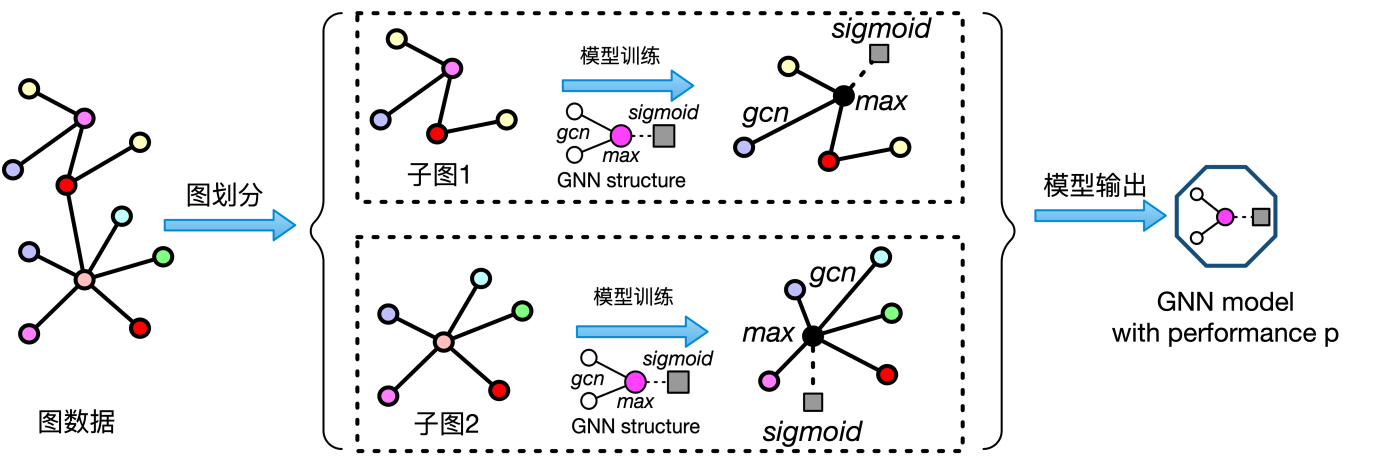
进化学习是一种模拟生物进化过程的学习方法。从随机的初始种群开始，这些算法在先前的精英样本附近产生新的个体。其中遗传算法是一种典型的进化学习算法，主要包含了四个组成部分，1.编码:构建染色体，将问题的解用一种数学方式表达。2.个体：问题中一个解的编码可以看成一个个体，多个个体组成一个种群。3.适应度：评判个体的好坏程度。4.遗传算子：种群进化迭代的方式。遗传算子是构建遗传算法的核心，其中包含了三个组成部分，1.选择：从前代种群中选择多对较优个体，一对较优个体成为一对父母。2.交叉：两个待交叉的不同染色体。3.变异：染色体按照一定的概率进行染色体内某些元素的变异。

进化学习搜索问题解的优势在与可以很好的利用并行机制，使用不同的遗传算子在相同或者不同的随机种子中并行搜索问题的解。本项目将基于强化学习搜索获取的初始图神经网络结构作为遗传算法的初始染色体，将不同遗传算子的特点与图神经网络结构搜索问题相结合，设计适配结构搜索问题的不同遗传算子，使用图神经网络结构在验证集上验证结果评估染色体构造适应度，将不同的遗传算子放入不同的计算节点中并行地对图神经网络结构进行搜索。

由于遗传算法的搜索将从较好的染色体出发，并能基于不同的遗传算子并行地搜索图神经网络结构，搜索速度将显著提升。

**（2）基于Partition-GNN的模型训练加速**

本项目将从降低内存占用和时间复杂度、提高收敛速度这三个角度研究图划分对图神经网络模型训练过程进行加速的可行性和适用于大规模图数据的图划分算法。如图5所示，Partition-GNN的输入为一个完整的图，首先通过图划分算法将图划分为多个子图，然后在每个子图上根据图神经网络结构搜索算法提供的模型结构训练相应的图神经网络模型，最后在图数据的验证集上评估模型。在Partition-GNN算法中，邻居搜索的范围被限制在子图中，不同子图的模型训练过程相互独立，因此可以在分布式环境下并行地在不同子图上训练图神经网络模型。对图进行划分的算法需具备高效性和并行性，算法本身的时间复杂度不能成为加快图神经网络模型训练速度的掣肘，因此本项目提出一种能够并行执行的图划分算法与图神经网络模型的训练过程相结合来加快图神经网络的训练速度。



**图5 Partition-GNN（PGNN）过程**

首先从内存占用的角度研究现有图神经网络方法在模型训练过程中图数据存储方面存在的问题和Partition-GNN的可行性。使用full-batch（全批量）参数更新策略的GNN模型如GCN需要在运算过程中存储全部节点在每一层图卷积层的临时嵌入，导致算法的空间复杂度非常高，不能用于大规模图的训练。图划分预处理操作能够降低分布式集群中单台机器需要处理的数据规模，并减少单个batch中对邻居节点的临时嵌入的存储，因而大幅降低内存占用。

其次从时间复杂度的角度研究现有图神经网络方法在图卷积层计算过程中存在的问题和Partition-GNN的可行性。GraphSAGE提出mini-batch策略更新参数，虽然降低了空间复杂度，但是由于参数需要通过多个batch进行更新，增加了算法的时间复杂度。通过图划分将多个子图分配到不同机器上存储，并将图卷积层的邻居搜索空间限制在子图中，减少聚合邻居信息过程中对噪声节点的吸收和跨机器通信，从而降低图神经网络算法的时间复杂度。

最后从模型收敛速度的角度研究现有图神经网络方法在模型训练过程中计算损失与更新参数的策略存在的问题和Partition-GNN的可行性。GCN采用full-batch策略计算损失，模型的收敛速度快但空间开销大。初步的改进策略是通过图划分算法将全图划分为内部联系紧密而外部联系稀疏的多个子图，集群中的每台机器分别负责一个子图的计算工作，但是由于子图之间的边被切开，训练过程相对独立，可能导致划分后的子图标签分布与全图不一致。为了解决上述问题，本项目提出将多个子图按照一定的数量随机划分到同一台机器中，并保留子图之间的边，以此解决子图之间标签分布不一致的问题，并能在更短的时间内收敛到与采用full-batch参数更新策略的GNN模型一样的结果。

综上所述，我们拟提出基于图划分算法的Partition-GNN图神经网络训练加速方法。由于图划分缩小了邻居搜索的范围，降低了噪声邻居的影响，减少了模型训练过程中的重复计算和内存中的冗余数据，使得模型能够支持在更大规模的图和更深的图卷积模型上更快地训练，并达到更好的效果。

4.4图神经网络建模加速的应用

图神经网络的应用领域非常广泛，例如通过交通网络监控和应对道路事故、通过生物图研究药物模型和预测疾病爆发、通过社交网络分析舆情和推荐产品等。本项目拟重点在电商网络、社交网络和生物网络中进行应用研究。

**（1）电商网络应用。**本项目由中南大学高建良教授和阿里巴巴集团人工智能技术总监李朝领衔组成，在长期合作中发现大规模电商图数据处理中需求和瓶颈所在。因此，本项目将图神经网络建模加速的理论成果应用到电商大规模图数据中，在十亿节点规模的图数据中检验和应用我们的理论成果。我们也计划将本项目研究成果集成到阿里巴巴开源项目GraphScope中，推进我国大规模图计算的发展。

**（2）生物网络应用。**我们的方案是将图神经网络建模加速技术应用在生物网络中，利用图分析方法解决生物网络中的实际问题。例如药物与药物之间相互作用网络中药物副作用的预测，疾病与生物实体相互作用网络中疾病相似性的研究，以及蛋白质相互作用网络中蛋白质功能分类预测等各种生物网络中的任务，我们将各种生物网络看做以生物实体为节点，生物实体间的相互作用为边的图，利用图计算方法将生物网络中需要解决的问题视为图分析中链路预测、相似性研究以及节点分类等任务。

**（3）金融知识图谱应用。**我们的方案是通过建模图神经网络，从金融事理图谱中挖掘事件的特征。事理图谱节点代表事件，有向边代表事件之间的顺承、因果、上下位等逻辑关系。挖掘知识图谱的实体、关系特征不仅仅要从实体本身入手，更要学习其连接的关系、周围的实体信息的结构信息。通过研究事件对金融市场时间序列数据的影响规律，主要分析不同属性的事件的相互影响，以及随时间对金融市场变化的不同程度的影响，构建金融事理图谱-金融市场时间序列数据关联模型；研究结合金融市场时间序列数据和金融事理图谱的时间感知图神经网络模型，得到具有时间相关性的融合特征向量进行后续投资风险预测任务。针对投资风险预测的金融事理图谱的图神经网络方法进行研究，为投资者、各金融机构及政府部门的市场预测、风险管理和管理决策提供理论支持。

# 预期成果

**（1）**在大规模图神经网络建模加速方向取得理论突破，在CCF推荐列表或IEEE/ACM Transactions等重要学术会议以及国内外重要期刊发表6­10篇论文。

**（2）**系统总结和提炼本项目的创新成果，在亿量级以上规模的大规模图中应用，申请国家专利1­2项。

# 六、申请人简介及研究团队基础

本项目由中南大学计算机学院高建良教授和阿里巴巴人工智能技术总监李朝（兼任之江实验室青年人才委员会委员）团队联合申报。项目团队长期从事图计算相关研究和应用开发，已经合作发表一系列图计算方向的高水平论文。同时，项目团队参与开发了阿里巴巴GraphScope项目（<https://github.com/alibaba/GraphScope>），本项目将力争进一步推进图计算理论和应用研究。

**6.1申请人简介**

高建良，博士，中南大学计算机学院教授、博士生导师。先后毕业于国防科技大学和中国科学院大学。近年来主要从事大规模图计算技术及其应用研究，2018年出版**《大规模图数据计算技术》**专著一部。

作为项目负责人，主持了国家自然科学基金项目、中国博士后特别资助项目、计算机体系结构国家重点实验室开放课题等项目。在IEEE Transactions期刊和CCF A/B类会议发表学术论文多篇，并与国内外学者建立良好的合作关系，目前担任阿里巴巴集团项目合作专家。在大数据处理领域取得一系列成绩，曾任IEEE Conference on Big Data 2016大会主席（General Chair）。作为学术带头人，带领中南大学团队建设了我国第一批数据科学与大数据技术专业，编写出版《Spark大数据编程基础》教材。

**代表性成果列表**：

[1] **高建良**，张丽霞; 大规模图数据计算技术,长沙：中南大学出版社, 2018.（ISBN：9787548733607） (学术专著)

[2] **Jianliang Gao**, Tengfei Lyu, Fan Xiong, Jianxin Wang, Weimao Ke, Zhao Li: MGNN: A Multimodal Graph Neural Network for Predicting the Survival of Cancer Patients. *International ACM Conference on Research and Development in Information Retrieval* (SIGIR),2020.(CCF A类会议)

[3] Xiaoting Ying, Cong Xu, **Jianliang Gao**, Jianxin Wang, Zhao Li: Time-aware Graph Relational Attention Network for Stock Recommendation. *ACM International Conference on Information & Knowledge Management* (CIKM),2020.(CCF B类会议)

[4] Zhao Li, Chenyi Lei, Pengcheng Zou, Donghui Ding, Shichang Hu, Zehong Hu, Shouling Ji, **Jianliang Gao**: Attention with Long-Term Interval-Based Gated Recurrent Units for Modeling Sequential User Behaviors. *International Conference on Database Systems for Advanced Applications* (DASFAA),2020.(CCF B类会议)

[5] **Jianliang Gao**, Ling Tian, Tengfei Lv, Jianxin Wang, Bo Song, Xiaohua Hu, Protein2Vec. Aligning Multiple PPI Networks with Representation Learning, I*EEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019, DOI: 10.1109/TCBB.2019.2937771. (SCI)

[6] **Jianliang Gao**, Jianxin Wang, Jianbiao He, Fengxia Yan: Against Signed Graph Deanonymization Attacks on Social Networks. *International Journal of Parallel Programming*. 47(4): 725-739,2019. (SCI)

[7] Jun Gao, **Jianliang Gao\***, A Similarity Measurement Method Based on Graph Kernel for Disconnected Graphs. IJCAI 2019. (CCF A类会议)

[8] Fan Xiong, **Jianliang Gao\***, Entity Alignment for Cross-lingual Knowledge Graph with Graph Convolutional Networks. IJCAI 2019. (CCF A类会议)

[9] Hongliang Du, Zhiyi Jiang, **Jianliang Gao\***, Who is Who: Name Disambiguation in Large-Scale Scientific Literature. ICDM Workshops 2019: 1037-1044.

[10] **Jianliang Gao**, Jianxin Wang, Ping Zhong, Haodong Wang, On Threshold-Free Error Detection for Industrial Wireless Sensor Networks, *IEEE Transactions on Industrial Informatics*, 14(5): 2199-2209, 2018. (中科院一区SCI，高被引论文)

[11] **Jianliang Gao**, Bo Song, Weimao Ke, Xiaohua Hu, BalanceAli: Multiple PPI Network Alignment With Balanced High Coverage and Consistency,*IEEE Transactions on NanoBioscience*, 16(5): 333-340,2017. (SCI)

[12] **Jianliang Gao**, Bo Song, Zheng Chen, Weimao Ke, Wanying Ding, Xiaohua Hu, Counter Deanonymization Query: H-index Based k-Anonymization Privacy Protection for Social Networks, *International ACM Conference on Research and Development in Information Retrieval* (SIGIR), 2017. (CCF A类会议)

[13] **Jianliang Gao**; Qing Ping; Jianxin Wang, Resisting re-identification mining on social graph data, *World Wide Web*, pp. 1759-1771, 2018. (SCI)

[14] **Jianliang Gao**, Ping Liu, Xuedan. Kang, Lixia Zhang, and Jianxin Wang, PRS: Parallel Relaxation Simulation for Massive Graphs, *Computer Journal*, 2016, 59(6), pp. 848-860.(SCI)

[15] **Jianliang Gao**, Bo Song, Ping Liu, Weimao Ke, Jianxin Wang, and Xiaohua. Hu, Parallel Top-k Subgraph Query in Massive Graphs: Computing from the Perspective of Single Vertex, *IEEE International Conference on Big Data*, pp. 636-645, 2016.

**6.2 研究团队基础**

研究团队由中南大学和阿里巴巴集团长期从事图计算的合作者组成。中南大学参与团队由高建良教授牵头，阿里巴巴集团参与团队由算法专家李朝牵头。

中南大学是中南大学是教育部直属全国重点大学，国家“211工程”首批重点建设高校，国家“985工程”部省重点共建高水平大学和国家“2011计划”首批牵头高校，2017年9月入选世界一流大学A类建设高校。中南大学计算机学院拥有计算机科学与技术A类学科，在计算机理论与算法、大数据前沿研究方面具有优势，2006年获得“计算机优化算法及其应用技术”教育部长江学者创新团队，2017年通过教育部高等学校学科创新引智计划项目“医学大数据分析理论与应用”。

**李朝**，现任阿里巴巴人工智能技术总监，浙江省千人，之江实验室青年人才委员会委员，专注于电子商务排名和推荐系统。曾在TCL Research America担任推荐系统方面的首席数据科学家。获得了美国国家科学基金会EPSCoR的奖学金，并在吴信东教授的指导下完成了佛蒙特大学计算机科学系的博士学位。在此之前，在CSC(中国留学基金委)的资助下获得了华南理工大学的硕士学位。目前主要研究方向为知识图、图计算、网络表示学习、多智能体强化学习、大数据驱动安全。是CCF数据库技术委员会委员，2019年CIKM大数据挑战赛获得冠军。

**代表性论文：**

[1] **Zhao Li**, Xin Sheng, Yuhang Jiao, Xuming Pan, Pengchen Zou, Xianling Meng, Chenwei Yao, Jiajun Bu：Hierarchical Bipartite Graph Neural Networks: Towards Large-Scale E-commerce Applications. *IEEE 36th International Conference on Data Engineering* (ICDE), 2020: 1677-1688. (CCF A类会议)

[2] **Zhao Li**, Long Zhang, Chenyi Lei, Xia Chen, Jianliang Gao, Jun Gao:. Attention with long-term interval-based deep sequential learning for recommendation. *Complexity*, 2020.

[3] **Zhao Li**, Haobo Wang, Donghui Ding, Shichang Hu, Zhen Zhang, Weiwei Liu, Jianliang Gao, Zhiqiang Zhang, Ji Zhang: Deep Interest-Shifting Network with Meta-Embeddings for Fresh Item Recommendation. *Complexity*, 2020.

[4] **Zhao Li**, Yuying Xing, Xiaoming Huang, Haobo Wang, Jianliang Gao, Guoxian Yu： Large-scale online multi-view graph neural network and applications. *Future Generation Computer Systems*, 2020.

[5] Guoxian Yu, Xia Chen, Carlotta Domeniconi, Jun Wang, **Zhao Li**, Zili Zhang, Xiangliang Zhang： CMAL:Cost-effective Multi-label Active Learning by Querying Subexamples. *IEEE Transactions on Knowledge and Data Engineering*, 2020. (CCF A类期刊)

[6] Haobo Wang, **Zhao Li**, Jiaming Huang, Pengrui Hui, Weiwei Liu , Tianlei Hu, Gang Chen： Collaboration Based Multi-Label Propagation for Fraud Detection. IJCAI. 2020. (CCF A类会议)

[7] Zhen Zhang, Jiajun Bu, Martin Ester, Jianfeng Zhang, Chengwei Yao, **Zhao Li** and Can Wang： Learning Temporal Interaction Graph Embedding via Coupled Memory Networks. *Proceedings of The Web Conference 2020*. 2020: 3049-3055 (CCF A类会议).

[8] Junshuai Song, **Zhao Li**, Zehong Hu, Jun Gao：PoisonRec: An Adaptive Data Poisoning Framework for Attacking Black-box Recommender Systems. *IEEE 36th International Conference on Data Engineering* (ICDE). IEEE, 2020: 157-168. (CCF A类会议)

[9] Shouling Ji, Qinchen Gu, Haiqin Weng, Qianjun Liu, Pan Zhou, Jing Chen, **Zhao Li**, Raheem Beyah, Ting Wang: De-Health: All Your Online Health Information Are Belong to Us. *2020 IEEE 36th International Conference on Data Engineering* (ICDE). IEEE, 2020: 1609-1620. (CCF A类会议)

[10] Chenyi Lei, Lei Wu, Dong Liu, **Zhao Li** , Guoxin Wang, Haihong Tang, Houqiang Li ： Multi-Question Learning for Visual Question Answering. *AAAI*. 2020: 11328-11335. (CCF A类会议)

[11] Xuanwu Liu, **Zhao Li**, Yuanhui Mao, Lixiang Lai, Ben Gao, Yao Deng and Guoxian Yu： Dynamical User Intention Prediction via Multi-modal Learning. *International Conference on Database Systems for Advanced Applications*. Springer, Cham, 2020: 519-535.

[12] Yuying Xing, **Zhao Li**, Pengrui Hui, Jiaming Huang, Xia Chen, Long Zhang and Guoxian Yu：Link Inference via Heterogeneous Multi-view Graph Neural Networks. *International Conference on Database Systems for Advanced Applications*. Springer, Cham, 2020: 698-706.

[13] Jianliang Gao, Tengfei Lyu, Fan Xiong, Jianxin Wang, Weimao Ke, **Zhao Li**: MGNN: A Multimodal Graph Neural Network for Predicting the Survival of Cancer Patients. *International ACM Conference on Research and Development in Information Retrieval* (SIGIR),2020.(CCF A类会议)

[14] Xiaoting Ying, Cong Xu, Jianliang Gao, Jianxin Wang, **Zhao Li**: Time-aware Graph Relational Attention Network for Stock Recommendation. *ACM International Conference on Information & Knowledge Management* (CIKM),2020.(CCF B类会议)

[15] Shixun Huang, Yuchen Li, Zhifeng Bao, **Zhao Li**: Towards Efficient Motif-based Graph Partitioning: An Adaptive Sampling Approach.ICDE 2021. (CCF A类会议, Accepted)

**专利：**

1. 2019年，鲁鸣鸣、张坤芳、张阳、蔡佳文、田卓林、毕文杰、刘海英, 一种基于层级 时序图卷积网络的路网交通流量预测，发明专利申请号 201910164972.7
2. 2019年，鲁鸣鸣、谭丁武、毕文杰、刘海英、郑一基、易驰，一种计算机自主学习源代码的方法，发明专利申请号 201910178229.7
3. 2019年，鲁鸣鸣、刘海英、吴君彦、毕文杰，一种融合胶囊机制的图卷积网络论文分类方法，发明专利申请号 201910168853.9
4. 2019年，鲁鸣鸣、刘海英、伍谷丰、王建新、潘毅、毕文杰，一种基于力导图的图神经网络可视分析方法，发明专利申请号 201910167501.1
5. 2019年，鲁鸣鸣、谢尚晟，一种基于梯度的图神经网络可解释性方法，发明专利申请号 201911069788.0