ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2014/08(00)-0000-00
doi: 10.3778/j.issn.1673-9418.

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-89056056

基于低秩结构和局部约束矩阵估计的链接预测方法

刘冶[1], 印鉴[1+], 邓泽亚[1], 王智圣[1], 潘炎[2]

1. 中山大学 信息科学与技术学院, 广东省 广州市 510006
2. 中山大学 软件学院, 广东省 广州市 510006

Link Prediction with Low-rank Structure and Local Constraint Matrix Pursuit

LIU Ye[1], YIN Jian[1+], DENG Zeya[1], WANG Zhisheng[1], PAN Yan[2]

1. School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510006, China
2. School of Software, Sun Yat-sen University, Guangzhou 510006, China
+ Corresponding author: E-mail: issjyin@mail.sysu.edu.cn

LIU Ye, YIN Jian, DENG Zeya, et al. Link Prediction with Low-rank Structure and Local Constraint Matrix Pursuit. Journal of Frontiers of Computer Science and Technology, 2014, *(*): *-*.**

Abstract: The link prediction problem is an significant subfield of link mining.With the existed data from the network structure, the unknown relationship of nodes in the network can be predicted by link prediction models. In the big data era, the study of link prediction for social network on the web and other complex networks has attracted popular research interest. Some methods with link prediction algorithms have been widely used in social network relationship mining, individual recommendation and biological pharmacy. In the problem of complex network link prediction, the similarity matrix is selected for representing the probability of existed links between any nodes. Therefore, the calculation method for estimating the similarity matrix becomes the most crucial step. Recent years, most research focus on the methods based on data analysis which construct the similarity relationship matrix by data-dependent machine learning and optimization algorithm. In this paper, an novel data-dependent link prediction

*The National Natural Science Foundation of China under Grant Nos.61033010, 61272065 (国家自然科学基金); the Natural Science Foundation of Guangdong Province under Grant Nos.S2011020001182, S2012010009311 (广东省自然科学基金); the Research Foundation of Science and Technology Plan Project in Guangdong Province under Grant Nos.2011B040200007, 2012A010701013 (广东省科技计划项目); the Research Foundation of Science and Technology Plan Project in Guangzhou City under Grant Nos.11A31090341, 11A53010726(广州市科技计划项目).
Received 2014-**, Accepted 2014-**.

method is proposed with the global low-rank structure assumption and local constraint of node feature in the network. The new proposed method is designed for scalable divide-and-conquer calculation for complex network and suitable for distributed computation. Extensive experiments on several real-world datasets show that the proposed link prediction measure obtains competitive performance compared with the baselines, the result also indicates the new algorithm is effective, robust and scalable for complex network.

Key words: link prediction; low-rank estimation; local constraint; social network; data mining

摘　要: 网络链接预测问题是链接挖掘的一个重要部分，指的是通过已知的网络结构数据预测网络中尚未连接的任意节点间产生链接的可能性。在大数据时代，互联网社会网络和其他复杂网络中的链接预测问题研究成为热门领域。链接预测相关的方法已被广泛地应用于社会网络关系挖掘、个性化推荐和生物制药等领域。在复杂网络的链接预测问题中，通常利用相似性矩阵来表示网络中任意两个节点之间存在链接的可能性，因此相似性矩阵的计算是链接预测中的至关重要的一步。近年来的研究中，大多数方法是基于已知网络中数据的分析，通过网络潜在结构设计机器学习算法构造相似性矩阵。在全局低秩的网络结构假设下，结合网络中节点特征的局部约束，提出了一种基于数据的链接预测优化算法，并针对复杂网络数据链接预测问题设计了可扩展的分治方法，便于分布式环境中对大规模数据求解。通过在多个真实数据集上的实验和结果分析，提出的基于低秩结构和局部约束矩阵估计的链接预测分治方法能够取得较好的效果，并对复杂的网络结构数据具有较强的可扩展性。

关键词: 链接预测；低秩估计；局部约束；社会网络；数据挖掘

文献标志码：A　　**中图分类号**：****

1 引言

网络关系挖掘中的链接预测（link prediction）问题是近年来数据挖掘和机器学习领域一个新兴的热点领域。链接预测的主要研究内容是通过已知网络中建立连接的数据，构造链接预测模型[1]。在实际的应用中，链接预测被用于社会网络分析、个性化推荐和生物基因工程等领域。链接预测问题主要分为两种类型，一种是对于给定包含已知关系的网络，预测网络中可能存在的未知关系，另一种是以当前时刻网络中存在的链接关系，预测下一时刻网络中会存在的链接。事实上，这两种链接预测问题都可以转化为相同的链接预测模型进行求解，对于给定已知的网络数据，通过网络的全局结构和局部特征建立模型，预测链接存在的可能性。

对于链接预测问题的研究，通常会定义描述网络中节点间相似性的矩阵，通过相似性矩阵来描述链接存在的概率，而相似性矩阵可以通过对已知的网络数据进行分析和挖掘获得。因此，链接预测的主要任务是利用网络中可观察到节点间的关系设计预测未知部分的链接是否存在的模型[2]。链接预测模型既可以用于静态的网络数据，还可以扩展到动态变化的网络中，利用当前时刻的网络数据和结构预测一段时间后的链接关系。

在社会网络和其他复杂网络的分析和挖掘中，通常会将链接预测问题转化为具体的相似性矩阵构造和估计，并利用机器学习的优化算法求解[3]。通过问题转化，链接预测中的一个基本问题是链接关系存在可能性的计算。对于给定的网络，可以根据网络中的链接关系定义为有向图或无向图，其中包含表示网络中节点的顶点集和表示网络链接存在关系可能性的边集。通过对网络数据的分析，可以构造机器学习模型进行训练和预测，模型建立过程需要通过数据挖掘技术对网络中的数据信息提取，对数据的全局和局部特性进行分析，建立网络链接关系的特征描述，提高学习和预测的效果。

相似性矩阵的构造问题可以通过分析网络拓扑结构设计节点相似度计算公式求解，这类方法在机

器学习领域属于无监督的算法，应用于快速地计算网络节点间链接关系，并通过简化计算方式，已经扩展到大规模的系统[4]。网络拓扑结构的节点相似度计算方法[2][5]存在局限性，随着社会网络和其他复杂网络的数据量的提升，传统的相似性方法在计算可扩展性和预测结果可靠性上面临挑战[14]。近年来，基于数据的机器学习方法的监督[12]和半监督[6]算法开始被用于链接预测中的相似性矩阵估计问题。监督和半监督学习方法需要通过网络数据的潜在结构和特征构造假设，同时处理大规模数据网络中链接预测问题中噪声数据和缺失数值对链接关系预测结果的影响。

基于监督和半监督学习的链接预测方法需要构造机器学习优化模型求解，模型中需要对相似性矩阵的全局结构和局部特征进行约束，便于优化算法估计求解。对于相似性矩阵构造的全局结构约束，目前比较有效的是低秩（low-rank）和稀疏两种[3][30]。由于网络关系中有“物以类聚，人以群分”的特点，这种特点实际上可以转化为矩阵的低秩结构，能够有效地预测链接关系存在的可能性，同时在模型中通过加入稀疏性约束，这类约束能够有效地降低噪声数据和缺失数值的影响，使得相似性矩阵更加逼近大规模数据预测中的稀疏特点。

结合低秩和稀疏的特性，就可以将问题转化为利用优化理论求解复杂的机器学习模型，基于低秩和稀疏约束模型的求解方法已经被广泛应用[10][13]。为了进一步提高链接预测的效果，在构造相似性矩阵的过程中可以利用网络中节点的特征信息，这种通过特征作为局部约束信息构造优化算法的思想在图像识别[23]和推荐系统[24]中已经得到有效应用。通过对社会网络和其他复杂网络中节点特征的数据描述信息提取，利用节点本身的特征作为相似性矩阵模型中的局部特征约束，这样做可以在全局结果约束的基础上取得更好的效果。

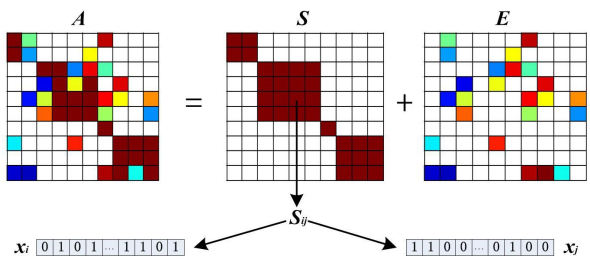


Fig.1 Overview of the proposed link prediction model with global low-rank structure and local characteristic

图 1 基于全局低秩和局部特征的链接预测模型示意图

本文工作的主要贡献包括：

1）通过低秩网络结构假设，在全局低秩结构的基础上加入稀疏噪声数据约束，同时利用网络节点的特征构造局部约束，提出了一种链接预测模型；

2）对于提出的链接预测模型，基于增广拉格朗日乘子法（Augmented Lagrangian Multiplier，ALM）框架[19]，提出了求解全局结果低秩和局部特征约束链接预测模型的优化算法，求解相似性矩阵近似估计问题；

3）针对规模较大的数据，设计了针对相似性矩阵求解过程中的分块求解方法，便于链接预测模型在分布式系统计算中提高计算效率。

链接预测模型中包含低秩与稀疏的双重约束，如图 1 所示，目的在于学习网络数据的内在结构的同时降低噪声对训练结果的影响，而局部特征约束是通过利用网络节点的额外信息增强预测效果。同时，相似性矩阵分块求解方法便于在较大规模数据上通过并行计算加快运行速度。通过在多个真实数据集的实验和结果分析，将本文提出的方法用于链接预测问题的相似性计算上，在实验表现上与基准方法（baseline）相比取得较好的效果，同时使用分块矩阵估计的方法使得预测模型在较大规模数据上具备计算可扩展性。

2 相关工作

链接关系普遍存在于数据集中的样本之间,链接挖掘是针对链接关系相关研究的一个新兴领域[1]。链接预测属于链接挖掘的范畴，任务是根据已有的数据和相关信息预测链接存在的可能性。链接预测模型目前可以被运用于社会网络关系挖掘、个性化推荐和生物制药等领域。在基于互联网的社会网络上，最基本的要素是社会关系，使用预测模型对社会链接关系进行分析是链接预测研究的重要方向。

目前对链接预测的相关研究较为丰富，比较系统地提出了链接预测的概念和任务，同时对链接预测的模型和方法做了相关研究。Liben-Nowell 等[2]首先提出了社会关系网络上的链接预测问题，阐述了链接预测问题的概念和相关研究进展。Getoor 等[1]从链接挖掘宏观角度，对链接预测及其相关领域的重要问题和方法进行了综述。Lü等[5]对复杂网络下的链接关系挖掘和预测的方法进行了总结，将链

接预测模型进行系统地分类和总结。比较和评价链接预测模型的效果需要使用适当的评估方法，Bradley[16]较为全面地介绍了链接预测常用的评估方法 AUC 度量。

在最初的相关研究工作中，基于相似性度量的网络拓扑方法用于设计链接预测模型，包括 Katz[31]、Common Neighbors[2]等方法被称为无监督学习方法，这类传统方法主要基于相似性度量，将数据转化为相似性矩阵的形式作为算法模型的输入，通过简单的运算求解。在无监督的方法中，Katz 算法效果较好且计算过程简单，在此基础上 Foster 等[31]提出了一种快速的无监督学习 Katz 算法。这类无监督学习的链接预测模型优点在于运算简单，便于广泛运用到各个领域，但主要问题在于在复杂网络数据中的预测准确性有限，并且对噪声数据具有敏感性。随着链接预测在大数据中应用需求的出现，在一些重要的工业界系统中，为了满足大规模链接预测的计算性能要求，出现了对无监督学习的链接预测模型进行简化改进的方法，Song 等[4]对关系挖掘中大规模数据链接预测的运算性能瓶颈进行了归纳和研究，提出了在效率和效果之间做出平衡大规模计算方法。

为了尝试解决基于拓扑的链接预测模型存在的问题，链接预测的相关研究从最初的无监督学习模型逐步向基于数据的监督学习和半监督学习模型发展。一些传统的监督学习[12][35]和半监督学习[6]的方法被用于链接关系分析和预测，Al 等[12]系统总结了有监督学习进行链接预测的方法。在监督学习模型的相关方法中，传统的无监督方法常被用于关系特征的构造，通过特征向量的形式描述实体关系对的特征，算法模型需要构造关系对的特征向量，然后通过描述训练集上的关系对特征向量和关系相似性标签，运用监督学习的技术在测试集上进行链接预测。监督学习和半监督学习模型可以提高链接预测的效果，但是关键在于构造数据集中关系对的特征，通常特征的构造需要依赖于基于拓扑的无监督学习方法的输出。

在链接预测模型广泛运用的社会网络和生物信息领域，随着用于预测的数据信息日益丰富，链接预测的研究从单源浅层网络向多源深层网络[32][33]发展，同时结合多个文本和声音的数据，使得问题的形式更加丰富，也为相关研究带来了挑战。社会网络挖掘目前在链接预测方面已经有较多的工作，随着近年来文本和图像的数据挖掘处理技术的发展，开始出现通过对文本[29]和图像[36]提取特征增强社会关系挖掘效果的研究。移动互联网的发展使得移动社交相关的网络数据逐渐丰富多样，Lichtenwalter 等[35]给出了社交网络链接预测的一些新的研究角度。链接预测也逐步和其他相关领域融合和发展，Huang 等[34]将链接预测和协同过滤联系并做了相关的研究。

传统的链接预测模型中主要依赖于单一的相似性矩阵的构造，而相似性矩阵描述的仅仅是链接是否存在，影响了对链接预测的准确率提升。由于单一相似性信息的局限性，多源网络的研究逐步兴起，Davis 等[17]对复杂的多源网络中的链接预测进行了研究，Lu 等[26]在多个源数据的网络中进行有监督链接预测，取得了一定的成果。社会网络关系的变化使得网络的动态也成为研究方向，Raymond 等[6]提出了一种用于静态和动态图的快速的半监督链接预测算法。Ge 等[27]研究了多数据源社会网络关系挖掘中链接预测的冷启动问题。

基于数据的监督学习和半监督学习链接预测模型需要借助机器学习领域的优化算法求解。通常在链接预测模型中，会对数据潜在的全局结构进行研究并提出特定的假设。Hsich 等[9]提出了基于随机梯度下降（Stochastic Gradient Descent，SGD）优化方法[7]的低秩矩阵分解方法，该方法在低秩矩阵估计方面效果较好，同时也存在大规模数据下参数设定范围较大的问题。Menon 等[28]基于相似性矩阵全局低秩的假设，运用矩阵分解的方法求解链接预测问题。Richard 等[3]对社会网络的数据提出了同时满足低秩和稀疏的假设，针对链接预测的相似性矩阵设计了估计低秩和稀疏矩阵的算法，并通过实验发现同时使用稀疏和低秩约束进行矩阵估计所得效果优于仅使用低秩约束。Liu 等[30]在链接预测问题的相似性矩阵全局低秩和稀疏的双重约束的基础上，增加了矩阵结构的等式约束，并引入了支持向量机（Support Vector Machine，SVM）中常用的 Hinge 损失函数[11]，设计了相似性矩阵近似求解方法。为了进一步提高低秩矩阵估计的效果，在全局结果约束的基础上可以结合矩阵相关的特征增加局部特征约束，Lee 等[24]将这种思想运用到推荐算法中，Zheng 等[23]则在图像研究中也使用了类似的局部约束。为解决大规模数据下低秩矩阵约束求解问题，Pan 等[25]基于对矩阵分块求解的相关研究工

作[20][21][22]，提出了一种使用分治方法求解低秩矩阵约束问题的算法。

受到相关工作的启发，本文的工作引入了局部特征约束，在全局低秩约束的基础上，进一步通过网络中节点之间的局部特征关系差异增加约束项。同时，为了克服链接预测的噪声问题，在全局低秩约束和局部特征约束上，对输入数据构造的相似性矩阵加入等式约束，将输入数据分解为全局低秩矩阵和稀疏噪声矩阵。在模型训练求解的优化方法上，通过基于增广拉格朗日乘子法（Augmented Lagrangian Multiplier，ALM）[19]框架，设计对应的链接预测相似性矩阵估计方法，该方法能在优化求解过程中交替更新各个相关变量，且每一步均有闭式解（closed-form solution）。此外，为满足大规模数据问题，在优化求解框架的基础上，提出了矩阵分块估计的方法，使得本文提出的链接预测模型具有可扩展性。

3 链接预测方法

3.1 问题模型

在链接预测问题的建模和求解过程中，需要将社会网络或其他复杂网络数据转换为形式化的表示。对于链接预测问题中的网络数据，可以用图 $G=(V, E)$进行描述，其中图 G 根据网络中的关系可以是有向图或者是无向图，顶点集 V 表示图中的顶点即网络中的节点，边集 E 表示图中的边即网络中节点间的关系。链接预测与关系挖掘的很多问题都能够转化为图 G 中边集 E 的关系，同时可以从概率的角度定义图 G 对顶点集 V 中任意两个顶点之间的边存在的概率，对应网络中节点之间存在某种关系的可能性，这样就可以运用机器学习的优化算法通过训练建立模型并进行链接预测。

链接预测模型的输入通常是相似性矩阵，定义网络节点间的相似关系可以由邻接矩阵 $\boldsymbol{A}$ 描述，相似性矩阵 $\boldsymbol{A}$ 中的一个元素表示为 $\boldsymbol{A}_{ij}$。同时，规定当网络中节点 i 与节点 j 存在链接关系时 $\boldsymbol{A}_{ij}$ 的值为 1，反之节点 i 与节点 j 不存在链接或者链接关系未知的情况下，则设 $\boldsymbol{A}_{ij}$ 值为 0。通常，链接预测模型的目标是预测网络中任意两个节点之间是否存在链接，而存在链接的可能性一般用概率表示。

$$A_{ij}=\begin{cases}0 & \text{节点 } i \text{ 和节点 } j \text{ 之间存在链接}\\ 1 & \text{节点 } i \text{ 和 } j \text{ 之间不存在链接或关系未知}\end{cases}$$

链接预测需要解决的问题是，基于已知的网络中存在的链接数据预测网络中可能存在的未知链接，或者以当前网络中链接预测网络中链接的变化趋势。基于数据的链接预测，需要依赖相似性矩阵作为数据，通过模型优化求解的过程，输出网络中任意两个节点之间存在链接的概率。因此，设计链接预测模型的关键问题是对输入的相似性矩阵数据，通过合理的假设对全局和局部结构进行描述，并使用适当的矩阵估计方法给予输入数据获得符合要求矩阵输出。

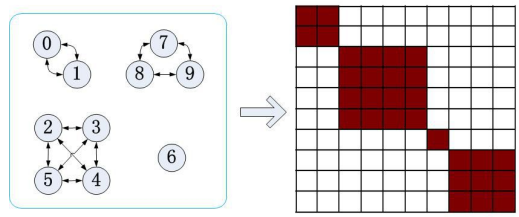


Fig.2 Example of the relationship between the aggregation of nodes in the network and the property of low-rank matrix

图 2 网络节点聚集和矩阵低秩性质的关系示意图

链接预测输出的相似性矩阵通过概率描述网络中节点间存在链接的可能性大小。通过对客观世界真实数据的分析，链接预测的网络关系可以聚集成簇[3][30]，例如在真实的社会网络关系中，人与人之间会形成不同的圈子。如图 2 所示，类似的聚集结构特点可以在相似性矩阵中形式化地表示成低秩（low-rank）约束，在图像识别[23]、推荐系统[24]、排序学习（learning to rank）[10]以及多视图聚类（multi-view clustering）[13]等领域已经被广泛运用，同时在链接预测的相关研究[3][28][30]中低秩也被选择作为全局结构约束。对于输入的相似性矩阵 $\boldsymbol{A}$，设链接预测模型优化求解的目标矩阵为 $\boldsymbol{S}$，则矩阵的秩可以形式化记为 $rank(\boldsymbol{S})$。

除了全局结构存在低秩的特点，在真实数据的环境下，数据的噪声普遍存在，在模型中需要使用适当的去噪方法[13][25]。链接预测模型输入的相似性矩阵客观上会受到稀疏噪声数据的影响，令矩阵 $\boldsymbol{E}$ 表示描述噪声的矩阵，根据相似性矩阵的定义，可以限制噪声矩阵 $\boldsymbol{E}$ 中的非零元素个数保持噪声矩阵的稀疏特点，对应地数学上稀疏性约束使用 l_0-范数来描述，l_0-范数的数学符号记为$\|\boldsymbol{E}\|_0$。

网络结构中的局部节点特征对链接预测的效果也存在影响，例如社会关系网络中，两个人之间是否相似不仅和所在的圈子有关系，同时和这两个人

本身的年龄、身份和工作等自身具有的特征密切相关。局部的特征可以用来构造局部特征矩阵描述网络中节点间的特点，在图像识别的工作中[23]通过局部特征的相似性可以提升算法效果，因此在链接预测中可以利用网络中节点本身的特征使得模型更加准确。

在定义链接预测模型的目标函数之前，需要确定对网络中节点之间的局部特征的计算方法。设 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 为描述网络中任意两个节点的特征向量，定义 f_{ij} 为特征向量 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的欧式距离的平方，即 f_{ij} 是特征向量 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 的欧几里德范数（Euclidean norm）的平方，则 f_{ij} 的定义形式如下：

$$f_{ij} = ||x_i - x_j||_2^2 \tag{1}$$

通过定义 f_{ij} 可以在链接预测中度量任意两个特征向量的局部相似性，进一步地，需要通过约束建立预测的目标矩阵 $\boldsymbol{S}$ 与局部相似性的关系。定义 $g(\boldsymbol{S})$ 表示描述预测的目标矩阵 $\boldsymbol{S}$ 和局部特征约束关系之间联系的函数，函数值表示局部特征和目标矩阵 $\boldsymbol{S}$ 之间存在的不一致性，因而在优化求解的过程中需要最小化函数的值。对于局部约束的评估函数 $g(\boldsymbol{S})$，通过距离函数 f_{ij} 计算得到的值越大表示对应网络中节点 i 和节点 j 的特征向量相似度越低，这时对应节点 $\boldsymbol{S}_{ij}$ 计算得到的值应该更趋近于 0，函数定义如下：

$$g(S) = \sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij}|S_{ij}| \tag{2}$$

综上所述，对于链接预测模型中的目标矩阵 $\boldsymbol{S}$，定义了全局低秩约束 $rank(\boldsymbol{S})$，同时通过评估函数 $g(\boldsymbol{S})$ 描述局部特征约束。此外，为了克服模型中噪声的干扰，链接预测模型中假设通过输入的相似性矩阵 $\boldsymbol{A}$ 得到目标矩阵 $\boldsymbol{S}$，同时分解出稀疏的噪声矩阵 $\boldsymbol{E}$。本文基于全局低秩约束、局部特征约束以及稀疏噪声约束的链接预测模型的优化问题可以形式化定义如下：

$$\begin{aligned} &\min_{S,E} \ rank(S) + \lambda_1||E||_0 + \lambda_2 g(S) \\ &s.t \ \ A = S + E. \end{aligned} \tag{3}$$

在优化问题的形式化定义中，针对目标矩阵使用了全局低秩约束 $rank(\boldsymbol{S})$ 和局部特征约束 $g(\boldsymbol{S})$，通过稀疏约束 $||\boldsymbol{E}||_0$ 控制噪声数据，并通过等式约束输入矩阵 $\boldsymbol{A}$、目标矩阵 $\boldsymbol{S}$ 和噪声矩阵 $\boldsymbol{E}$ 之间的关系，优化问题中的 λ_1 和 λ_2 是调节各个约束权重的参数。

通过对链接预测问题的分析和建模，已经得到了完整的优化问题形式化定义。但是，优化问题中求解低秩约束 $rank(\boldsymbol{S})$ 和稀疏约束 $||\boldsymbol{E}||_0$ 是 NP 难问题，使得优化问题在有效时间内求解存在困难。因此，通常的解决办法是放松约束条件转化为凸优化问题。对于低秩约束 $rank(\boldsymbol{S})$，使用核范数 $||\boldsymbol{S}||_*$ 代替，核范数 $||\boldsymbol{S}||_*$ 表示矩阵 $\boldsymbol{S}$ 的奇异值（singular value）之和，同时使用 l_1-范数 $||\boldsymbol{E}||_1$ 代替稀疏噪声约束 $||\boldsymbol{E}||_0$ 约束，矩阵的 l_1-范数表示矩阵中元素的绝对值之和。此外，在问题的约束中目标矩阵 $\boldsymbol{S}$ 重复出现不便于元素迭代依次求解，因而引入矩阵 $\boldsymbol{D}$ 作为中间变量。通过问题的转化，可以得到如下的优化问题形式化定义：

$$\begin{aligned} &\min_{S,E,D} \ ||S||_* + \lambda_1||E||_1 + \lambda_2 g(D) \\ &s.t \ \ A = S + E \\ &\quad\ \ S = D. \end{aligned} \tag{4}$$

通过上述的定义和转化可以得到链接预测优化问题完整的形式，现在的问题是需要设计优化求解算法通过输入矩阵 $\boldsymbol{A}$ 得到目标矩阵 $\boldsymbol{S}$。本文中通过运用增广拉格朗日乘子法框架[19]，设计了针对转化后的最优化问题中的矩阵元素迭代依次求解的优化算法，可以求解本文提出的基于全局低秩约束和局部特征约束的问题，算法将问题分解为对应全局低秩约束 $||\boldsymbol{S}||_*$、稀疏噪声约束 $||\boldsymbol{E}||_1$ 和局部特征约束 $g(\boldsymbol{D})$ 的子问题并依次求解，而且每个迭代中针对每个约束的求解过程均有闭式解（closed-form solution）。

3.2 优化算法

接下来，我们将重点关注形式化定义的预测模型优化的具体过程，给出以原始相似性矩阵 $\boldsymbol{A}$ 为输入，目标矩阵 $\boldsymbol{S}$ 为输出的算法流程，根据对矩阵 $\boldsymbol{A}$ 和矩阵 $\boldsymbol{S}$ 的定义可知，当链接预测的网络中有 m 个节点时，输入矩阵 $\boldsymbol{A}$ 和目标矩阵 $\boldsymbol{S}$ 为 m 行 m 列的方阵，具体的算法流程如下所示：

算法 1 基于增广拉格朗日乘子法（Augmented Lagrangian Multiplier，ALM）框架求解全局低秩（global low-rank）约束和局部特征（locality characteristics）约束的链接预测模型算法（ALM-GLRLC）

输入：相似性矩阵 $\boldsymbol{A}$，预测模型权重参数 λ_1 和 λ_2

输出：模型优化求解得到的目标矩阵 $\boldsymbol{S}$

算法开始：

第 1 步：矩阵变量的初始化。将算法中涉及的矩阵变量 $\boldsymbol{S}$、$\boldsymbol{E}$、$\boldsymbol{D}$ 初始化为零矩阵；

第 2 步：参数变量的初始化。将迭代次数变量 *iter* 初始化设置为 0，算法两轮迭代之间的误差阈值 ε设为 10^{-8}，增广拉格朗日乘子法框架中涉及的参数 μ和ρ分别设为 10^{-8} 和 1.2 的数值；

第 3 步：对矩阵变量 $\boldsymbol{S}$、$\boldsymbol{E}$、$\boldsymbol{D}$ 迭代依次求解的过程。具体的步骤如下：

① 迭代次数计数变量 *iter* 的数值增加 1；

② 根据公式（11）更新矩阵 $\boldsymbol{S}$；

③ 根据公式（17）更新矩阵 $\boldsymbol{E}$；

④ 根据公式（22）更新矩阵 $\boldsymbol{D}$；

⑤ 根据公式（23）和公式（24）更新 $\boldsymbol{Y}_1$ 和 $\boldsymbol{Y}_2$；

⑥ 将μ更新为ρ乘以μ的值，如果此时μ大于阈值 10^{10} 则将μ设置为 10^{10}；

第 4 步：判断目标矩阵是否满足收敛条件。如果矩阵关系同时满足$\|\boldsymbol{A}\text{-}\boldsymbol{S}\text{-}\boldsymbol{E}\|_\infty$小于误差阈值$\varepsilon$和$\|\boldsymbol{S}\text{-}\boldsymbol{D}\|_\infty$小于误差阈值$\varepsilon$两个收敛条件，则满足收敛条件执行第 6 步，否则执行算法第 5 步验证迭代次数；

第 5 步：判断迭代次数是否小于最大迭代次数。如果迭代次数超过设定最大迭代次数，则执行第 6 步，否则跳转回第 3 步继续下一轮迭代；

第 6 步：输出链接预测模型通过优化求解得到的满足收敛条件的目标矩阵 $\boldsymbol{S}$。

算法结束。

上述算法基于增广拉格朗日乘子法框架，通过不断迭代优化最小化损失函数，每次迭代中交替地更新各个矩阵变量，直到结果满足最优化收敛条件。由此可见，在整个算法中最重要的部分就是每次迭代中交替更新矩阵变量的步骤，因而下面将着重详细推导出各个矩阵变量的更新规则。

首先，我们将公式（4）定义的链接预测模型优化问题形式化表示包含了全局低秩约束$\|\boldsymbol{S}\|_*$、稀疏噪声约束$\|\boldsymbol{E}\|_1$和局部特征约束 $g(\boldsymbol{D})$，同时还具有两个对于相关矩阵的等式约束，这种形式在优化问题上不便于求解，需要将两个等式约束进行转化，因而下面将公式（4）的优化问题改写成对应的拉格朗日函数（Lagrangian function）的形式：

$$\begin{aligned} &L(S,E,D,Y_1,Y_2;\mu)\\ &=||S||_*+\lambda_1||E||_1+\lambda_2 g(D)\\ &+<Y_1,A-S-E>+\frac{\mu}{2}||A-S-E||_F^2\\ &+<Y_2,S-D>+\frac{\mu}{2}||S-D||_F^2 \end{aligned} \tag{5}$$

其中，符号$\|\boldsymbol{X}\|_F$表示矩阵 $\boldsymbol{X}$ 弗罗贝尼乌斯范数（Frobenius norm）即矩阵 $\boldsymbol{X}$ 每个位置绝对值平方和的开方的值，符号<$\boldsymbol{X}_1$,$\boldsymbol{X}_2$>表示矩阵的内积，$\boldsymbol{Y}_1$和 $\boldsymbol{Y}_2$为拉格朗日乘子（Lagrangian multiplier），μ为大于零的数值表示适应性惩罚（adaptive penalty）参数。

然后，由于增广拉格朗日乘子法框架可以较好地平衡优化求解过程的效率和准确性，算法使用该框架在每轮迭代中依次地更新矩阵变量 $\boldsymbol{S}$, $\boldsymbol{E}$, $\boldsymbol{D}$, $\boldsymbol{Y}_1$, $\boldsymbol{Y}_2$，当更新其中的任意一个矩阵变量时，将其他矩阵的值固定看作常量，因而通过多轮迭代，每轮迭代依次求解矩阵变量，直到算法收敛可以得到目标解。算法框架的具体更新策略推导如下：

更新矩阵 $\boldsymbol{S}$：

根据公式（5），在优化问题的拉格朗日函数中，将矩阵变量 $\boldsymbol{S}$ 之外的其他矩阵变量的值固定时，可以得到关于矩阵变量 $\boldsymbol{S}$ 的最优化子问题：

$$\begin{aligned} &\min_S\ ||S||_*\\ &+<Y_1,A-S-E>+\frac{\mu}{2}||A-S-E||_F^2\\ &+<Y_2,S-D>+\frac{\mu}{2}||S-D||_F^2 \end{aligned} \tag{6}$$

为了便于求解推导，设矩阵变量 $\boldsymbol{T}_{S1}$和 $\boldsymbol{T}_{S2}$如下：

$$\begin{aligned} T_{S1}&=A-E+\frac{Y_1}{\mu}\\ T_{S2}&=D-\frac{Y_2}{\mu} \end{aligned} \tag{7}$$

在矩阵变量 $\boldsymbol{S}$ 的最优化子问题公式（6）中，除去$\|\boldsymbol{S}\|_*$的低秩约束形式，剩下的部分可以化简如下：

$$\begin{aligned} &<Y_1,A-S-E>+\frac{\mu}{2}||A-S-E||_F^2\\ &+<Y_2,S-D>+\frac{\mu}{2}||S-D||_F^2\\ &=\frac{\mu}{2}||A-S-E+\frac{Y_1}{\mu}||_F^2+\frac{\mu}{2}||S-D+\frac{Y_2}{\mu}||_F^2\\ &=\frac{\mu}{2}||S-(A-E+\frac{Y_1}{\mu})||_F^2+\frac{\mu}{2}||S-(D-\frac{Y_2}{\mu})||_F^2\\ &=\mu||S-\frac{1}{2}(T_{S1}+T_{S2})||_F^2+constant \end{aligned} \tag{8}$$

从公式（8）的推导结果，可以得到由带参数μ的矩阵弗罗贝尼乌斯范数（Frobenius norm）和常数项组成的形式，为了进一步简化问题的表示，令矩阵变量 $\boldsymbol{Z}_S$为如下形式：

$$Z_S=\frac{1}{2}(T_{S1}+T_{S2}) \tag{9}$$

将公式（9）的矩阵变量 $\boldsymbol{Z}_S$代入公式（8）的推导结果，并加入$\|\boldsymbol{S}\|_*$的低秩约束，可以得到经过推导简化后关于矩阵变量 $\boldsymbol{S}$ 的最优化问题形式：

$$\min_S\ ||S||_*+\mu||S-Z_S||_F^2+constant \tag{10}$$

经过简化后的公式（10）描述的优化问题有闭

式解（closed-form solution），可以通过奇异值阈值方法（Singular Value Threshold method）求解[18]，令 $U\Sigma V$ 为 Z_S 的奇异值分解的结果，则可以得到矩阵变量 S 的更新公式如下：

$$S = U\mathbb{S}_{1/2\mu}(\Sigma)V^T \tag{11}$$

在矩阵变量 S 的更新公式（11）中，使用了收缩算子（shrinkage operator）[19]，形式如下：

$$\mathbb{S}_\eta(X) = \max(X+\eta, 0) + \min(X-\eta, 0) \tag{12}$$

根据上述推导结果，通过公式（11）和公式（12）就可以计算得到需要更新的矩阵变量 S 的值。

更新矩阵 E：

关于矩阵 E 的优化子问题形式与矩阵 S 的类似，当更新矩阵变量 E 时，固定其他变量，由公式（5）拉格朗日函数可以得到关于矩阵 E 的子问题：

$$\begin{aligned}&\min_E \ \lambda_1||E||_1 \\ &+ < Y_1, A-S-E > + \frac{\mu}{2}||A-S-E||_F^2\end{aligned} \tag{13}$$

同样地，为了便于推导关于矩阵 E 的子问题的闭式解，设矩阵中间变量 Z_E 为：

$$Z_E = A - S + \frac{Y_1}{\mu} \tag{14}$$

在关于矩阵变量 E 的最优化子问题公式（13）中，考虑除了 $\|E\|_l$ 的稀疏约束剩下的部分，并进行合并化简，并将公式（14）代入，具体推导过程如下：

$$\begin{aligned}&< Y_1, A-S-E > + \frac{\mu}{2}||A-S-E||_F^2 \\ &= \frac{\mu}{2}||A-S-E+\frac{Y_1}{\mu}||_F^2 \\ &= \frac{\mu}{2}||E-(A-S+\frac{Y_1}{\mu})||_F^2 \\ &= \frac{\mu}{2}||E-Z_E||_F^2\end{aligned} \tag{15}$$

在公式（15）中加入 $\|E\|_l$ 的稀疏约束，便得到经过推导简化后关于矩阵变量 E 子问题的形式：

$$\min_E \ \lambda_1||E||_1 + \frac{\mu}{2}||E-Z_E||_F^2 \tag{16}$$

针对公式（16）中的矩阵变量 E 子问题，同样具有闭式解，求解形式如下：

$$E = \mathbb{S}_{\lambda_1/\mu}(Z_E) \tag{17}$$

在关于矩阵变量 E 子问题的闭式解中，也使用了公式（12）进行计算。

更新矩阵 D：

矩阵 D 是为了简化求解过程加入的辅助矩阵，同样地固定其他矩阵变量，则得到以下与矩阵变量 D 有关的优化子问题：

$$\begin{aligned}&\min_D \ \lambda_2 g(D) \\ &+ < Y_2, S-D > + \frac{\mu}{2}||S-D||_F^2\end{aligned} \tag{18}$$

定义矩阵中间变量 Z_D，便于表示推导出的关于矩阵变量 D 的优化子问题，矩阵 Z_D 的等价关系如下：

$$Z_D = S + \frac{Y_2}{\mu} \tag{19}$$

在公式（18）中描述的关于矩阵 D 的子问题中，不考虑局部约束的评估函数 $g(D)$，余下部分化简：

$$\begin{aligned}&< Y_2, S-D > + \frac{\mu}{2}||S-D||_F^2 \\ &= \frac{\mu}{2}||S-D+\frac{Y_2}{\mu}||_F^2 \\ &= \frac{\mu}{2}||D-(S+\frac{Y_2}{\mu})||_F^2 \\ &= \frac{\mu}{2}||D-Z_D||_F^2\end{aligned} \tag{20}$$

通过公式（20）的推导得到结果，加入局部约束的评估函数 $g(D)$，得到化简的关于矩阵 D 的子问题：

$$\min_D \ \lambda_2 g(D) + \frac{\mu}{2}||D-Z_D||_F^2 \tag{21}$$

上述推导得到的公式（21）描述了关于矩阵变量 D 的优化子问题，求解该问题需要进行转化，通过观察可以发现，公式（2）定义的局部约束评估函数 $g(D)$与公式（16）中关于矩阵变量 E 的 $\|E\|_l$ 的稀疏约束十分类似，局部约束评估函数 $g(D)$可以看做是矩阵每个位置绝对值带系数的 l_1-范数稀疏约束，对应于公式（17）的求解过程，关于矩阵 D 的子问题同样可以求解，具体做法是对于矩阵 D 中的任意元素 D_{ij}，在求解过程中将评估函数 $g(D)$中绝对值之前由公式（1）定义的距离度量结果看做系数，代入公式（12）中，具体得到的更新公式为：

$$D_{ij} = \mathbb{S}_{\lambda_2 f_{ij}/\mu}(Z_{Dij}) \tag{22}$$

由上可知，关于矩阵变量 D 的优化子问题也同样具有闭式解。

更新矩阵 Y_1 和 Y_2：

由拉格朗日乘子法框架可以得到拉格朗日乘子 Y_1 和 Y_2 的更新公式如下：

$$Y_1 = Y_1 + \mu(A-S-E) \tag{23}$$

$$Y_2 = Y_2 + \mu(S-D) \tag{24}$$

更新矩阵 Y_1 和 Y_2 是增广拉格朗日乘子法框架的常规步骤，其中涉及的参数变量μ需要在链接预测模型算法开始时初始化，并每轮迭代中不断更新。

在处理大规模复杂网络的矩阵数据时，为了提

高模型运算效率，通常需要设计针对优化问题的矩阵分块方法，使得模型可以对完整的输入矩阵划分为多个子矩阵，然后对子问题进行计算，最后合并子问题的结果得到目标矩阵，对于本文提出的链接预测优化问题，可以运用文献[20][21][22]对类似问题的矩阵分治法求解，使得处理在大规模网络数据链接预测问题时可以进行并行处理从而缩短计算时间。

4 实验和分析

在实验部分，我们实现了本文提出的基于低秩结构和局部约束矩阵分块估计的链接预测方法（ALM-GLRLC），并使用链接预测模型常用度量标准作为评价指标，将本文的算法与链接预测问题中常用有效的算法在多个真实数据集上进行测试，并分析和对比运行结果。

4.1 数据集和对比方法

本文的实验中选择了用于链接预测的三个真实世界的网络数据集，分别是 CiteSeer 网络、Cora 网络和 WebKB 网络[15]，用于对比本文提出的算法与基准算法在评估度量方法上的测试效果。

由于链路预测模型被广泛应用于网络关系预测中，本文首先选择了 CiteSeer 网络数据集，该数据集描述了计算机科学界不同领域论文引用的网络数据，其中包含了论文的引用关系以及每篇论文经过分析得到的词属性列表和所属类别，在数据预处理中，论文的引用关系被用来构建链接预测相似性矩阵输入，而每篇论文的词属性列表和所属类别则用于生成网络中节点的特征向量，通过预处理的步骤便可以得到符合本文算法模型的输入要求数据。

类似地，在 Cora 网络数据中主要包含了机器学习特定领域的相关论文引用数据，对该网络数据的预处理过程和 CiteSeer 数据集相同，输入相似性矩阵表示论文的引用关系，同时通过论文的词属性列表和类别标签构造网络节点的特征向量。

为了表明本文模型的可扩展性，实验中还选取了 WebKB 网络数据集，该数据集主要描述了多个大学网页链接的网络数据，而每个网页同样包含了对网页中内容进行提取之后的关键词属性列表和类标签，在数据预处理时随机选取了其中一个大学的网络数据，生成符合要求的链接预测模型输入，验证算法在不同的链接关系网络上的有效性。

在构造输入相似性矩阵 $\boldsymbol{S}$ 时，如果对于任意节点 i 和节点 j 存在链接关系，则设置 $\boldsymbol{S}_{ij}$ 的值为 1，反之，将 $\boldsymbol{S}_{ij}$ 的值设为 0 表示节点 i 和节点 j 不存在链接关系或链接关系不明确。

各个数据集的基本信息如下表 1 所示：

Table1 Statistics of the datasets

表 1 数据集的节点与边统计数据

数据集	节点数	边数
CiteSeer	3312	4591
Cora	2708	5429
WebKB	187	310

链接预测问题可以通过基于相似性网络拓扑结构的方法计算结果，这类经典的方法包括 Common Neighbors[2]、Katz[31] 和 Rooted Page Rank[2][4]，本文的实验中选取这三种方法作为比较的基准算法。经典方法中的 Common Neighbors（CN）算法通过计算网络节点对之间的共有邻接节点数目，以此数值作为节点相似性的评估，该方法有点在于计算简单，但在大规模稀疏网络上影响算法效果。在无监督学习方法中，Katz 和 Rooted Page Rank（RPR）的效果较好，特别 Katz 受到广泛运用，这两种方法实现简单，在大规模数据中需要考虑矩阵求逆的效率问题。

近年来关于链接预测的研究中也提出了基于数据的监督学习方法，文献[3]中使用了和本文提出模型类似的全局低秩假设作为约束，并使用增量邻近下降法（Incremental Proximal Descent，IPD）[8]解决低秩与稀疏约束（low-rank and sparse constraint）的矩阵估计问题，该方法在基于数据的监督学习方法中取得较好的效果，因而实验中选择并仔细实现了该方法作为基准算法，并记作 IPD-LRSP 方法。本文提出的方法与基准算法相比最显著的特征是同时利用了全局结构和局部特征，使得网络数据得到更加充分利用。

4.2 评估函数

本文的实验中使用了文献[16]中给出的经典评估度量方法 AUC（Area Under the receiver operating characteristic Curve）来作为衡量链路预测模型效果的指标，该方法在链接预测的研究实验中被普遍使用。AUC 度量用于评估链接预测模型的输出结果与随机算法的结果相比提高的程度[5]，描述测试集中边的评估值优于随机选择一条不存在的边的概率。

对于 CiteSeer 和 Cora 网络数据集，实验中首先

随机选择描述网络的相似性矩阵中10%的节点及对应的边，并在测试中分别对随机选取的数据加入5%到30%不同比例的高斯分布（Gaussian distribution）噪声，将添加噪声后的数据作为链接预测模型的输入，对于每组噪声参数，模型重复10次实验并将计算得到的AUC取平均得到实验评估结果。同样地，在WebKB网络数据集中由于节点数量较少，每次会在相似性矩阵中选取50%的节点及对应的边作为数据添加噪声。

4.3 实验结果分析

本文的实验为了验证提出的链接预测模型算法的效果，在实验中将本文提出的方法ALM-GLRLC与Common Neighbors（CN）、Katz、Rooted Page Rank（RPR）和IPD-LRSP做对比，使用了由粗到精的方法确定模型参数取值范围并选取最佳的运行结果，在不同噪声下的复杂网络数据集上测试不同模型获得的目标矩阵在评估方法AUC度量下的效果。

Table2 The AUC results of the CiteSeer dataset

表2 CiteSeer数据集在不同的噪声下的AUC值

算法	5%	10%	15%	20%	25%	30%
CN	0.524	0.556	0.529	0.511	0.572	0.535
Katz	0.947	0.871	0.837	0.781	0.758	0.704
RPR	0.957	0.869	0.815	0.795	0.733	0.698
IPD-LRSP	0.942	0.889	0.842	0.795	0.757	0.691
ALM-GLRLC	**0.960**	**0.918**	**0.891**	**0.857**	**0.829**	**0.761**

从表2的实验结果分析可知，在CiteSeer网络数据集下，基准算法中Common Neighbors（CN）由于模型简单并不能获得较好的效果，而Katz、Rooted Page Rank（RPR）和IPD-LRSP的测试效果总体上比较接近，本文提出的链接预测算法ALM-GLRLC在不同噪声下都取得比所有对比基准算法中最优值更好的效果。从总体上看，从5%逐步提高到30%比例的高斯分布噪声会使得大多数链接预测算法的效果受到影响并呈现在AUC度量的评估值上，本文提出的ALM-GLRLC算法在不同比例的高斯噪声上均取得更好的效果。在5%较低的高斯噪声影响下，由于ALM-GLRLC的结果比基准算法中的最优值略微提升0.3%，这说明在较低的噪声下，网络数据的全局结构特征受到较小的影响，通过网络数据的全局结构可以较好地获得目标矩阵输出，而在较高的30%高斯噪声下，ALM-GLRLC比基准算法中的最优值大幅度提升8.1%，说明在高噪声的复杂网络下，网络的全局结构受到较大的影响，此时ALM-GLRLC算法同时考虑全局结构约束和局部特征约束的优点得到充分发挥，通过局部节点特征可以明显提升预测效果。

Table3 The AUC results of the Cora dataset

表3 Cora数据集在不同的噪声下的AUC值

算法	5%	10%	15%	20%	25%	30%
CN	0.566	0.572	0.566	0.559	0.527	0.513
Katz	0.958	0.913	0.846	0.827	0.739	0.680
RPR	0.941	0.888	0.812	0.814	0.726	0.681
IPD-LRSP	0.962	0.908	0.849	0.812	0.721	0.689
ALM-GLRLC	**0.970**	**0.916**	**0.860**	**0.857**	**0.769**	**0.751**

从表3可以看出，在Cora数据集的测试结果和CiteSeer数据集类似，本文提出的ALM-GLRLC算法在较低的5%和较高的30%高斯噪声的情况下，分别比基准算法中的最优值提升了0.8%和9%。

Table4 The AUC results of the WebKB dataset

表4 WebKB数据集在不同的噪声下的AUC值

算法	5%	10%	15%	20%	25%	30%
CN	0.577	0.603	0.589	0.578	0.561	0.535
Katz	0.969	0.921	0.855	0.801	0.752	0.669
RPR	0.952	0.897	0.827	0.776	0.721	0.651
IPD-LRSP	0.972	0.911	0.832	0.780	0.755	0.658
ALM-GLRLC	**0.977**	**0.938**	**0.867**	**0.824**	**0.784**	**0.737**

从表4的测试结果可以得出，本文提出的链接预测ALM-GLRLC算法在不同类型的数据集上都有较好的效果，在WebKB数据集上，本文的方法和基准算法的最优结果作对比，在5%和30%的高斯噪声下，本文的链接预测模型分别提升了0.5%和10.2%。

通过对上述实验结果的分析可知，本文提出的基于全局结构和局部特征约束的ALM-GLRLC算法在多个数据集和不同噪声下能获得比基准算法更优解。

5 总结

本文针对链接预测问题的相关方法较少利用网络中节点本身的特征这一问题，设计了一种基于全局结构约束和局部特征约束的链接预测优化问题模型，并运用增广拉格朗日乘子法框架提出了迭代求解优化问题的链接预测算法，该算法在大规模的网络数据链接预测问题中可通过矩阵分块求解的分治法保证计算扩展性。在多个真实网络数据集上的实验结果说明，本文提出的基于低秩结构和局部特征约束矩阵估计的链接预测算法能够取得比基准算法更好的效果，具备鲁棒性与有效性，并且对高噪声的复杂网络结构数据具有较强的可扩展

性。随着数据规模的增加，链接预测问题可以向多数据源方向扩展，因此基于本文的思想针对多源融合问题设计链接预测模型是下一步的研究目标。

References:

[1] Getoor L, Diehl C P. Link mining: a survey[J]. ACM SIGKDD Explorations Newsletter, 2005, 7(2): 3-12.

[2] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks[J]. Journal of the American society for information science and technology, 2007, 58(7): 1019-1031.

[3] Richard E, Savalle P, Vayatis N. Estimation of Simultaneously Sparse and Low Rank Matrices[C]//Proceedings of the 29th International Conference on Machine Learning. 2012: 1351-1358.

[4] Song H H, Cho T W, Dave V, et al. Scalable proximity estimation and link prediction in online social networks[C]//Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. ACM, 2009: 322-335.

[5] Lü L, Zhou T. Link prediction in complex networks: A survey[J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(6): 1150-1170.

[6] Raymond R, Kashima H. Fast and scalable algorithms for semi-supervised link prediction on static and dynamic graphs[M]. Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2010: 131-147.

[7] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8): 30-37.

[8] Bertsekas D P. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey[J]. Optimization for Machine Learning, 2011, 2010: 1-38.

[9] Hsieh C J, Chiang K Y, Dhillon I S. Low rank modeling of signed networks[C]//Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012: 507-515.

[10] Pan Y, Lai H, Liu C, et al. Rank Aggregation via Low-Rank and Structured-Sparse Decomposition[C]//AAAI. 2013.

[11] Gentile C, Warmuth M K. Linear hinge loss and average margin[C]//NIPS. 1998, 11: 225-231.

[12] Al Hasan M, Chaoji V, Salem S, et al. Link prediction using supervised learning[C]//SDM 2006: Workshop on Link Analysis, Counter-terrorism and Security. 2006.

[13] Xia R, Pan Y, Du L, et al. Robust Multi-View Clustering via Low-rank and Sparse Decomposition[C]//AAAI. 2014.

[14] Sarkar P, Moore A W, Prakash A. Fast incremental proximity search in large graphs[C]//Proceedings of the 25th international conference on Machine learning. ACM, 2008: 896-903.

[15] De A, Ganguly N, Chakrabarti S. Discriminative link prediction using local links, node features and community structure[C]//Data Mining (ICDM), 2013 IEEE 13th International Conference on. IEEE, 2013: 1009-1018.

[16] Bradley A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms[J]. Pattern recognition, 1997, 30(7): 1145-1159.

[17] Davis D, Lichtenwalter R, Chawla N V. Multi-relational link prediction in heterogeneous information networks[C]//Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on. IEEE, 2011: 281-288.

[18] Cai J F, Candès E J, Shen Z. A singular value thresholding algorithm for matrix completion[J]. SIAM Journal on Optimization, 2010, 20(4): 1956-1982.

[19] Lin Z, Chen M, Ma Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices[J]. arXiv preprint arXiv:1009.5055, 2010.

[20] Mackey L W, Talwalkar A, Jordan M I. Divide-and-Conquer Matrix Factorization[C]//NIPS. 2011: 1134-1142.

[21] Williams C, Seeger M. Using the Nyström method to speed up kernel machines[C]//Advances in Neural Information Processing Systems 13. 2001.

[22] Goreinov S A, Tyrtyshnikov E E, Zamarashkin N L. A theory of pseudoskeleton approximations[J]. Linear Algebra and Its Applications, 1997, 261(1): 1-21.

[23] Zheng Y, Zhang X, Yang S, et al. Low-rank representation with local constraint for graph construction[J]. Neurocomputing, 2013, 122: 398-405.

[24] Lee J, Kim S, Lebanon G, et al. Local low-rank matrix approximation[C]//Proceedings of The 30th International Conference on Machine Learning. 2013: 82-90.

[25] Pan Y, Lai H, Liu C, et al. A Divide-and-Conquer Method for Scalable Low-Rank Latent Matrix Pursuit[C]//Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013: 524-531.

[26] Lu Z, Savas B, Tang W, et al. Supervised link prediction using multiple sources[C]//Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010: 923-928.

[27] Ge L, Zhang A. Pseudo Cold Start Link Prediction with Multiple Sources in Social Networks[C]//SDM. 2012: 768-779.

[28] Menon A K, Elkan C. Link prediction via matrix factorization[M]. Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2011: 437-452.

[29] Tang J, Wang X, Liu H. Integrating social media data for

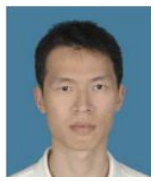
community detection[M]. Modeling and Mining Ubiquitous Social Media. Springer Berlin Heidelberg, 2012: 1-20.
[30] Liu Y, Wang Z, Yin J, et al. A Scalable Proximity Measure for Link Prediction via Low-rank Matrix Estimation[C]//3rd International Conference on Computer Science and Service System. Atlantis Press, 2014.
[31] Foster K C, Muth S Q, Potterat J J, et al. A faster Katz status score algorithm[J]. Computational & Mathematical Organization Theory, 2001, 7(4): 275-285.
[32] Li X, Du N, Li H, et al. A Deep Learning Approach to Link Prediction in Dynamic Networks[C]//SIAM International Conference on Data Mining, 2014.
[33] Ge L, Gao J, Li X, et al. Multi-source deep learning for information trustworthiness estimation[C]//Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013: 766-774.
[34] Huang Z, Li X, Chen H. Link prediction approach to collaborative filtering[C]//Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries. ACM, 2005: 141-142.
[35] Lichtenwalter R N, Lussier J T, Chawla N V. New perspectives and methods in link prediction[C]//Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010: 243-252.
[36] McAuley J, Leskovec J. Image labeling on a network: using social-network metadata for image classification[M]. Computer Vision – ECCV 2012. Springer Berlin Heidelberg, 2012: 828-841.

LIU Ye was born in 1989. He is a master candidate at Sun Yat-sen University. His research interests include social network mining, large scale machine learning, deep learning algorithm, etc.
刘冶(1989–), 男, 中山大学硕士研究生, 主要研究领域为社会网络挖掘, 大规模机器学习, 深度学习算法。

YIN Jian was born in 1968. He received the Ph.D. degree from Wuhan University in 1989. He is a professor and doctoral supervisor at Sun Yat-sen University. He is a senior member of China Computer Federation. His research interests include big data, data mining, etc.
印鉴(1968–), 男, 博士, 中山大学教授, 博士生导师, CCF 高级会员, 主要研究领域为大数据, 数据挖掘。

DENG Zeya was born in 1990. He is a master candidate at Sun Yat-sen University. His research interests include large scale machine learning, social network mining, etc.
邓泽亚(1990–), 男, 中山大学硕士研究生, 主要研究领域为大规模机器学习, 社会网络挖掘。

WANG Zhisheng was born in 1986. He is a Ph.D. candidate at Sun Yat-sen University. His research interests include text data mining, large scale machine learning, etc.
王智圣(1986–), 男, 中山大学博士研究生, 主要研究领域为文本数据挖掘, 大规模机器学习。



PAN Yan was born in 1979. He received the Ph.D. degree from Sun Yat-Sen University in 2007. He is an associate professor at Sun Yat-sen University. His research interests include machine learning, data mining, information retrieval, etc.
潘炎(1979–), 男, 博士, 中山大学副教授, 主要研究领域为机器学习, 数据挖掘, 信息检索。