网络链路预测：概念与前沿

吕琳媛[1] 任晓龙[1] 周 涛[2]
[1]杭州师范大学
[2]电子科技大学

关键词：复杂网络 网络科学 链路预测 可预测性

引言

在网络科学发展早期，研究主要聚焦在发掘不同网络所遵循的普适规律，如小世界规律[1]、无标度现象[2]、社团结构[3]等。随着研究的深入，人们发现宏观指标以及基于宏观量的运算使得个体的特征被“平均化”，一些关键个体的表现被淹没了，导致得到的结论仅在统计上有意义，却难以进行精细的定量化描述。譬如说知道一个网络的稀疏程度、平均距离或者平均簇系数，并不代表就知道了某个具体节点的度和局部结构，因为不同节点之间相差甚远。因此，要想获得真正精细可靠的解释，就必须进行微观层面的深入分析[4]。近年来，网络科学研究逐渐从发掘不同网络宏观上的普适规律，转为更多地从介观层面（如网络的社团结构[5]、层级[6]、群组结构[7]等）和微观层面（如节点[8]和链路[9,10]）研究各类网络的特征[11]。

链路预测 (link prediction) 作为网络科学研究中一个重要且有趣的问题，本质上是从网络链路的微观层面解释网络结构生成的原因，进而帮助我们更好地理解网络所对应的复杂系统的结构生成和演化规律。

什么是链路预测？

作为连接网络科学与信息科学的重要桥梁之一，链路预测处理的是信息科学中最基本的问题——缺失信息的还原与预测[12]。该问题从已观察到的网络结构入手，预测存在但未被观察到，或者未来可能会出现的链路——可以说，链路预测探讨的是复杂网络中的“哈姆雷特”问题：连，还是不连？链路预测不仅能推动网络科学理论的发展，而且具有巨大的实际应用价值。

链路预测有着广泛的应用场景（如图 1 所示）。例如，用于指导生物实验以提高实验成功率[13]，对社交网络中的朋友推荐[14,15]和敌友关系进行预测[16]，电子商务网站上的商品推荐[17,18]，以及通过识别隐藏的链边和虚假的链边对信息不完全或含有噪音的网络进行重构[19]。这些应用场景都是显而易见的。然而有趣的是，链路预测的算法和思想通过一些变换还可以用来解决一些看似没有太大关系的问题。例如，对疾病信号[20]和股票指数的预测[21]。基于链路预测的思想和方法，我们还可以设计基于节点相似性的社团划分[22]以及对网络演化模型的定量化评估方法[23]。链路预测还可以帮助提高节点标签预测的精度[24]，特别是在已知标签节点稀疏的情况下效果显著。这一方法可以应用于判断一篇学术论文的类型，判断一本书的政治态度（中立、自由派或保守派），或者判断一个手机用户是否产生了更换电信运营商的念头。

在一个包含 N 个节点和 M 条边的无向无权网络中测试链路预测算法的效果，需要先将已知的连边分为两部分：训练集和测试集，且这两部分之间没有重合的连边。在使用算法进行预测的时候只能使用训练集中的信息。最常用的测试集的选择方法是在已知连

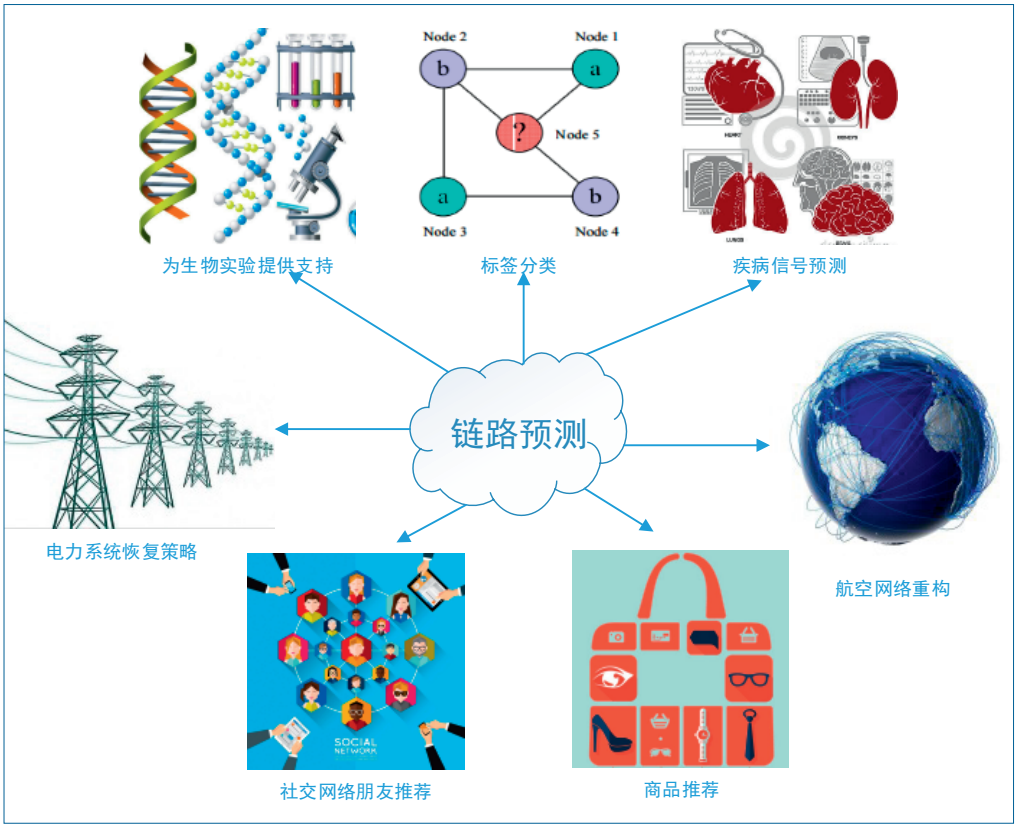


图1 链路预测应用场景示意图（图片来自网络）

边中随机选取一定比例的连边构成测试集。如果网络是含时的，可以按时间顺序将最近产生的连边作为测试集，也可以根据不同的预测目的采用不同方式选取测试集。例如，在考察算法对于小度节点之间产生连边的预测能力的时候，可以选择小度节点的连边组成测试集。那些没有连边的节点对构成“不存在的边”集合，所有不存在的边和测试集中的边构成“未知边”集合。链路预测的最终目的就是利用训练集中的信息将测试集中的连边预测出来。

衡量链路预测算法精确度最常见的指标是 AUC（Area Under ROC Curve，ROC 曲线下面积）和准确率 [12]。AUC 是从整体上衡量算法的精确度，可以理解为在测试集中随机选择一条边的分数值比随机选择一条不存在的边的分数值高的概率。这里，一条边的分数值由预测算法给出，分数值越高表示这条边出现的概率越高。准确率主要考虑预测列表中排在前 L 位的边的预测正确率。AUC 实际上是指 ROC（Receiver Operating Characteristic，受试者工作特征）曲线下的面积。在实际计算中，可以采用抽样比较的方法得到近似值，即每次从测试集中随机选取一条边，再从“不存在的边”集合中随机选择一条，如果前者分数值大于后者就加 1 分，如果两者分数值相等就加 0.5 分。这样独立比较多次后，求得的平均分数就是 AUC。显然，如果所有分数都是随机产生的，则 AUC 约等于 0.5，因此 AUC 大于 0.5 的程度衡量了该算法在多大程度上比随机选择的方法更精确。

在有些文献中也经常使用召回率 (Recall)、假阳性率 (False Positive Rate, FPR)、*F1* 等指标评价算法的性能。召回率指存在的连边中被预测准确的比例，假阳性率等于把不存在的连边误判为存在的连边的比例，*F1* 指标又称为平衡 *F* 分数，定义为准确率和召回率的调和平均数。

链路预测的方法

基于节点属性相似性的链路预测

物以类聚，人以群分。两个节点的属性越相似，就越可能产生联系。由于在线社交数据相对容易获取，所以目前这方面的研究主要集中在社交网络上，

但这些方法都能够很容易地推广到其他网络的链路预测中。在社交网络中，刻画节点的属性相似性，最简单直接的方法就是使用标签。例如，考察两个节点之间的年龄、职业、教育、兴趣、地理位置、性别、信仰等属性的相似程度，对节点对之间产生联系或者节点对之间关系的演化做出预测。例如，对一所大学中4万多名师生一年内电子邮件的交流情况（包括发送者、接收者、时间戳）、个人信息（年龄、性别、科系、年级）以及每名学生/教师的上课情况的研究发现，属性相似的学生/教师之间有更多交流[26]。这一结论在含有1.8亿人和13亿条无向边的MSN在线社交网络上也得到了验证——人们更愿意和与自己年龄相仿、使用相同的语言、地理位置相近的人聊天[27]。当用户标签不全、存在大量缺失时，可以用局部社交网络结构来补充用户的标签[28]。在节点的标签并非显而易见时，可以分析描述节点属性的文本相似性[29]。例如，用网络上用户收藏/发表的内容、参与的话题等信息计算用户之间的话题相似性并进行好友预测[30]。

基于网络结构相似性的链路预测

很多时候，我们很难获得网络中真实可信的节点属性信息。此时，根据所观察到的网络结构来计算节点相似性是一个可行且可信的方法。基于网络结构相似性的方法假设：在网络中，两个节点之间相似性（或者相近性）越大，它们之间存在连边的可能性就越大[31]。与节点属性的相似性不同的是，这里的相似性仅依赖于网络结构。

基于局部信息的最简单的相似性指标是优先连接(Preferential Attachment, PA)。应用优先连接的方法可以产生无标度网络结构。在该网络中，一条即将加入的新边连接到节点 x 的概率正比于节点 x 的度[2]，因此，新边连接节点 x 和 y 的概率正比于两节点度的乘积。因为需要的信息量非常少，所以该算法的计算复杂度较低。

另一个基于局部信息的相似性指标是共同邻居(Common Neighbor, CN)，即两个节点如果有更多的共同邻居，则它们更倾向于连边。在共同邻居的基础上考虑两端节点度的影响，从不同的角度又产生6种相似性指标[10,12]，分别是索尔顿(Salton)指标（也叫余弦相似性）、雅卡尔(Jaccard)指标、索伦森(Sorenson)指标、大度节点有利指标(Hub Promoted Index)、大度节点不利指标(Hub Depressed Index)[31]和LHN-I指标[32]。这一类指标统称为“基于共同邻居的相似性指标”。

在这些指标中，共同邻居对节点对之间产生连边都有促进作用，且不同的共同邻居的贡献是相等的。然而，一般情况下我们会觉得大度的共同邻居比小度的共同邻居贡献小。举个直观的例子，两个人都喜欢看同一部热门电影，这并不能说明两个人的兴趣品味有多相似。相反，如果这两个人同时喜欢一部冷门电影，那么可以说他们至少在这方面具有相似的品味。基于这样的考虑，阿达密克(Adamic)和阿达尔(Adar)提出了AA指标[33]，他们根据共同邻居节点的度为每个节点赋予一个权重值，该权重等于该节点的度的对数分之一。于是AA指标就等于两个节点的所有共同邻居的权重值之和。巧合的是，资源分配指标(Resource Allocation, RA)最终的形式与AA指标如出一辙，而不同在于权重的形式变成了节点的度分之一。大量的实验结果以及后期学者在网络社团挖掘等领域的应用显示，资源分配指标与AA指标表现相近，在精确性上略微胜过AA指标[10,12]。在共同邻居指标的基础上，一些改进算法可以进一步提高预测精度。例如，局部朴素贝叶斯模型可以区分共同邻居对于两个节点之间产生连边的作用是正的(促进连接)还是负的(抑制连接)[34]。

本质上，共同邻居指标可以看成是考虑两个节点间的二阶路径数目的方法。如果在二阶路径基础上再考虑节点间的三阶路径数，就可以得到局部路径指标(Local Path, LP)[35]。显然，局部路径指标可以继续扩展到更高阶的情形。当然，随着路径阶数的增加，计算复杂度也会越来越高，当阶数趋于无穷的时候，这一指标就相当于考虑全部路径的卡茨(Katz)指标[36]。上述的相似性指标大都是基于结构等价(structure equivalence)原则的，莱希特(Leicht)、霍姆(Holme)和纽曼(Newman)基于一般等价性

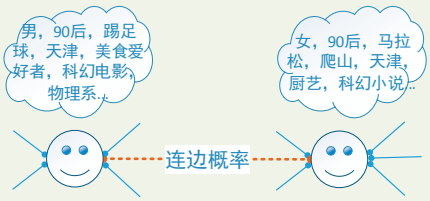
(regular equivalence) 提出了 LHN-II 指标 [32]，假设两个节点的邻居节点之间相似（不要求是同一个节点），则这两个节点也相似。

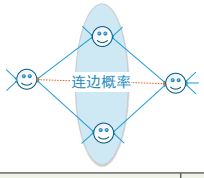
基于全局信息的指标除了卡茨指标和 LHN-II 外，还包括一组基于网络随机游走过程的指标，如平均通勤时间 (Average Commute Time, ACT)[37]、Cos+ 指标（或称“余弦相似性”）[38]、有重启的随机游走 (Random Walk with Restart, RWR)[39]、SimRank 指标 [40] 等。半局部指标包括局部随机游走指标 (Local Random Walk, LRW) 和有叠加效应的局部随机游走指标 (Superposed Random Walk, SRW)[41]，其中 SRW 比 LRW 更强调网络连接的局域性特征。其他全局指标还包括基于网络拉普拉斯矩阵的矩阵森林指数 [42] 以及考虑节点间相似性的可传递性的自洽转移相似性指标 [43] 等。

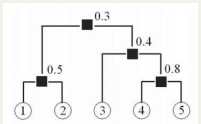
基于似然分析的链路预测

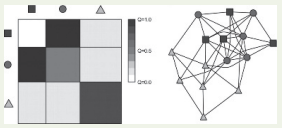
基于似然分析的链路预测的基本思路是：根据网络结构的产生和组织方式以及目前已经观察到的链路计算网络的似然值，并认为真实的网络使得网络似然值最大，然后再根据网络似然最大化计算每一对未连接的节点产生连边的可能性。层次结构模型 (Hierarchical Structure Model, HSM)[6] 假设真实的网络都存在某种层次性，网络的连接则可看作是这种内在层次结构的反映。一个 N 个节点的网络可以用一个包含 N 个叶子节点的族谱树表示，这 N 个叶子节点将由 N-1 个非叶子节点连接起来，其中每个非叶子节点都有一个概率值，则两个叶子节点连接的概率就等于他们最近共同祖先节点的概率值。给定一个族谱树，将网络的似然值最大化，就可以得到非叶子节点的概率值，并由此计算出这一个族谱树所对应的网络最大的似然值。文献 [6] 使用马尔科夫链蒙特卡洛方法对网络不同的族谱树进行抽样，使得每个族谱树出现的频次正比于该树对应的最大的网络似然值。两个节点连边的概率就等于所有抽样出来的族谱树中两个节点连边概率的平均

表1　四类链路预测方法（*表示该方法含有参数）

<table>
<tr><td>基于节点属性相似性的方法</td><td colspan="5"> 

两个节点的属性越相似，两节点间就越可能产生联系。在社交网络中，刻画节点的属性相似性，最简单直接的方法就是使用标签。例如，考察两个节点之间的年龄、职业、教育背景、兴趣、地理位置、性别、信仰等属性的相似程度。</td></tr>
<tr><td>基于网络结构相似性的方法</td><td>偏好连接相似性PA认为两节点连边的概率正比于两节点度的乘积。</td><td>基于共同邻居的指标：
 

CN, Salton
Jaccard,
Sorenson
HPI, HDI
LHN-I
AA, RA</td><td>基于路径的相似性指标：
LP*
Katz*
LHN-II*</td><td>基于随机游走的相似性指标：
ACT, Cos+
RWR*
SimRank*
LRW*, SRW*</td><td>其他指标：
SPM*
矩阵森林*
自洽转移相似性*</td></tr>
<tr><td>基于似然分析的方法</td><td>层次结构模型
 

网络可以用族谱树表示，每个非叶子节点有一个概率值，则两个叶子节点连接的概率等于他们最近共同祖先节点的概率值。</td><td colspan="2">随机分块模型
 
节点可以被分为若干集合，两个节点间连接的概率只取决于节点所在的群。</td><td colspan="2">闭路模型*
 
闭路模型认为封闭环路代表一种局部性，可以根据网络中的封闭回路数定义网络的似然，一条未被观察到的连边存在的可能性等于添加这条连边后网络的似然。</td></tr>
<tr><td>机器学习方法</td><td>基于特征分类的链路预测</td><td colspan="2">基于概率图模型的链路预测</td><td colspan="2">基于矩阵分解的链路预测</td></tr>
</table>

值。该方法在处理具有明显层次结构的网络（如恐怖袭击网络和草原食物链）时具有较高的精确度，但在处理一般性网络时效果并不一定突出。

基于类似的思想，随机分块模型 (Stochastic Block Model, SBM)[19] 假设网络中的节点可以被分为若干集合，两个节点间连接的概率只与相应的集合有关。随机分块模型的效果要好于层次结构模型。同时，该方法还可以剔除网络的错误连边，如纠正蛋白质相互作用网络中的错误连边。

似然分析的更一般的框架是 [12]：给定一个网络系统，特定网络哈密顿量的负指数被统计分配函数归一化后，就得到这个网络出现的似然，一条未被观察到的连边存在的可能性就等于添加这条连边后网络的似然值。闭路模型 (Loop Model) 考虑网络结构形成中的“局部性原则”，并由此定义了网络的哈密顿量。实验表明，闭路模型的预测精度大于层次结构模型和随机分块模型。虽然基于最大似然估计的方法计算复杂度高，但这类方法除了应用于链路预测以外，还带给我们关于网络结构的深刻洞见，比如层次结构模型给出了研究网络层次组织形态的定量化方法。

基于机器学习的链路预测

用机器学习的思路进行链路预测，主要分为基于特征分类方法、基于概率图模型方法和基于矩阵分解方法三大类 [44]。网络中的链路预测问题可以看成机器学习中的分类问题，其中每个数据点对应一对节点之间关系的标记，假定两个节点之间存在连边，则数据点的值为 +1，否则为 -1。特征选取是分类问题中最重要的问题之一，目前研究较多的主要包括基于节点与边的属性特征和基于节点所处网络的拓扑结构特征（如 CN、AA、Katz、最短路径）等 [15,44]。哈桑 (Hasan) 等人 [45] 提取合著网络中科学家研究领域的关键词作为特征，用监督学习中一些常用的分类算法（如决策树、K 近邻法、多层感知器、支持向量机、径向基网络）对缺失的连边进行较为精准的预测，其中以支持向量机方法表现最佳。文献 [46] 分析了 2012 年美国大选期间推特用户发送的推文，定义了一个基于情感特征的特征集合来预测两个用户成为朋友的可能性。

基于概率图模型的链路预测方法使用图模型来表达节点之间的连边概率，根据模型中概率依赖关系，可分为有向无环的贝叶斯网络和无向的马尔科夫网络 [44]。随机分块模型和层次结构模型分别是典型的基于贝叶斯网络和基于马尔科夫网络的链路预测方法。基于概率图模型的链路预测方法有很多，典型的还有文献 [47] 提出的有监督随机游走链路预测算法，将网络结构信息与节点和连边的属性信息结合起来，给每条边分配不同的转移概率。与其他有监督的机器学习方法相比，该算法无须知道网络的特征和相关的领域知识。

在基于矩阵分解的方法方面，文献 [48] 提出了一种基于网络邻接矩阵的代数谱变换的方法，来预测网络中链路的存在性和链路的权值。相比其他几种机器学习方法，该方法要学习的参数少很多，但计算复杂度较高。链路预测问题可以视为邻接矩阵填充问题，并可以通过双线性回归模型将节点和链路的显特征和隐特征结合起来用矩阵分解方法解决 [49]。边缘去噪模型将根据给定关系矩阵求未知边或丢失边的问题变成一个矩阵降噪去噪问题 [50]。文献 [15, 44] 从机器学习视角对网络链路预测进行了详细的介绍。

复杂类型网络上的链路预测

在真实系统中，个体间的关系非常复杂，有时难以用简单无向图精准刻画。比如食物链网络、微博关注网络、交通运输网络、社交关系演化网络等，需要用有向、含权、含时、多层等更加复杂的网络来描述。这些复杂类型网络上的链路预测往往与应用紧密结合，也因此吸引了越来越多研究者的关注。

在有向网络的链路预测中，不仅要考虑连接的存在性还要关注连接的方向。对于方向的预测是极具挑战性的 [51]。除了可以简单地将以往无向网络中的方法拓展到有向网络，还可以设计出专门针对有向网络的方法。基于局部子图结构的预测是常用的方法，例如可以基于社交网络的 9 种三角形闭合结构进行好友

推荐[14]。莱昂 (Leung) 等人[52]对局部结构做了统计，不仅包含三个节点的子图，还含有四个节点的子图，此外还考虑了异质边（即性质不同的连边，如敌友关系）的情况。文献 [53] 提出了有向网络链路预测的“势理论”，认为如果一条连边的出现能够产生更多的可定义势的子图，那么它出现的概率就越大。在一个有向子图中，任选一个节点给定初始的势能，顺边的方向，邻居的势能降低 1；逆边的方向，邻居的势能增加 1。如果此结构内每个节点的势能都是可定义的，那么就称此结构为“可定义势的子图”。实验发现，如果一条有向边的加入可以产生的含有四个节点的双风扇结构越多，那么这条边出现的概率就越大。除了基于局部子图的预测方法，还有一些基于有向网络中局部随机游走过程的方法也可以得到较好的预测效果[54,55]。此外，基于网络结构信息和节点标签信息的机器学习方法也可以进一步提升预测的精度，例如把基于结构的节点相似性作为此节点对的一个特征，并结合其他结构特征构成此节点对的特征向量，然后进行训练学习[56,57]。

含权网络链路预测的研究主要分为两个方面，(1) 权重对于预测连边存在性的作用；(2) 对权重本身的预测。显然，原有的无权网络的相似性指标（如局部指标 CN、AA、RA、PA）和基于路径的指标（如 LP、Katz 和 LHN-II）都可以很自然地扩展为含权的形式[58~60]。文献 [58] 应用含权的 CN、AA、PA 指标在日本的雅虎问答公告板[1]数据中进行实验，发现含权指标的预测效果要好于无权的预测方法。但是在另外一些研究中，研究人员发现了不同的结果。例如，在一些科学家合作网络中含权的 Katz 指标比不含权的 Katz 指标预测效果要差[9]。链路预测的“弱连接”效应表明：在有些含权网络中，权重较弱的边在链路存在性预测中起到的作用比权重高的边还大[59]。含权网络的链路预测同样可以用极大似然模型进行研究[61,62]。

含时网络预测最容易让人想到的是转化为含权网络的链路预测问题[63]，另一种方法是根据网络中的闭合频繁子图以及这些子图之间出现的时间规律进行预测[64]。跨时间片 / 跨层的多层网络链路预测问题也取得了一些进展[65]。这类预测方法将有助于解决网络的节点匹配问题。

可预测性

链路的可预测性是链路预测的一个基本问题。链路预测的精度在一定程度上反映了算法对于网络链路形成的可解释力。然而，不同算法在不同网络中的表现不尽相同。当我们获得一个很差的预测结果的时候，是选择了不恰当的算法造成的，还是因为网络的链路本身就很难预测？好的预测算法会给出网络演化可能机制的暗示。遗憾的是，我们并不知道一个算法是否“足够精确”。比如，针对一个完全随机的网络，“什么都预测不到”可能已经是最好的结果了，然而针对一个非常规则的网络，足够聪明的方法可以达到 100% 的精确预测。知道了一个网络的链路“能够被预测出来”的程度，可以帮助我们判断算法是否已经接近甚至达到预测的上界，是否还有提升的空间。“可被预测的程度”本身也可以看作是网络的一种重要性质。

然而，衡量网络链路的可预测性并非易事。许小可等人[66]通过分析网络演化过程中形成连边的两个节点之间的拓扑距离分布，提出了一种基于此分布概率的方法来刻画基于共同邻居相似性指标的预测上限。这一上界是依赖于算法的。如何计算不依赖特定算法的链路可预测性上界，目前还没有一个好的方法。“结构一致性”指标可以用来衡量网络可被预测的难易程度[67]。其假设网络越是具有某些规律性，就越容易被预测。如果从网络中随机抽取出一小部分连边，网络的特征向量空间受到的影响很小，就说明网络是具有规律性的。在这种思路的基础上，应用类似于量子力学中对哈密顿量做一阶微扰的方法，假定减少或者加入少量连边所产生的微扰只对特征值有影响，而对特征向量没有影响，这

[1] http://chiebukuro.yahoo.co.jp/。

样就可以观察微扰后通过这种办法重构的邻接矩阵和真实邻接矩阵之间的差异。结构一致性指标刻画的就是这种差异。实验发现，网络随机性越高，结构一致性就越差，也就越难被预测，但是随机性并不完全等同于可预测性。此外，结构一致性越强的网络，其丢失的边也越容易被准确地预测出来。

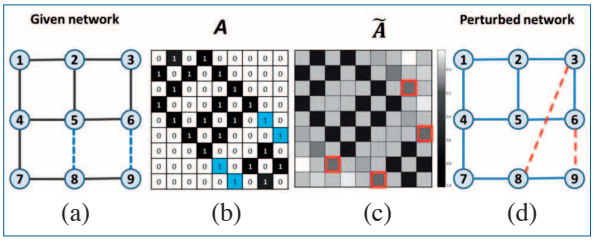


图2 计算网络结构一致性的示意图[67]

图 2 给出了一个计算网络结构一致性的示例。图 2(a) 表示原始网络，其对应的邻接矩阵如图 2(b) 所示。首先，从原始网络中选出一个小的边集合构成扰动集，如图 2(a) 中选择两条边 (5,8) 和 (6,9)，然后用扰动集对剩下的网络进行一阶微扰，得到扰动后的矩阵如图 2(c) 所示，扰动后的矩阵元素不再只是 0 或 1。对所有不存在的边（邻接矩阵中 0 元素对应的节点对集合）和扰动集中的边按照其在图 2(c) 矩阵中对应的值赋值，选出分数最大的两条边添加到不包含扰动集的网络中构成重构网络，如图 2(d) 所示，这个例子中边 (3,8) 和 (6,9) 分数最高。比较原始网络和重构网络，可以看到，两条边中有一条是相同的，因此该网络的结构一致性为 0.5。

借鉴结构一致性的思想，产生了一种名为结构微扰法 (Structural Perturbation Method，SPM) 的链路预测方法。实验表明，这个方法在预测丢失的链路以及甄别虚假连边两方面都优于当前主流的方法，包括层次结构 [6] 和随机分块模型 [19]。

展望

2011 年我们曾在《物理学 A》(*Physica A*) 上发表了一篇关于链路预测的英文综述 [10]，得到来自计算机科学、生物学、药物学、物理学、数学、社会科学等多学科的引用。这些论文不仅从理论上推动了链路预测的发展，还给出了各种各样的可以应用的实际场景。过去提出的一些挑战性问题如今都得到了不同程度的推进。例如，各种算法对不同网络预测能力的深度比较 [12]，利用链路预测建立网络演化机制的比较平台，揭示网络演化的规律 [23]，对网络链路的可预测性问题进行探讨 [67] 以及链路预测在其他网络类型中的理论和应用研究等。

尽管近几年链路预测方向取得了一系列成果，但是仍然存在很多有意思、具有挑战性的问题值得进一步深入研究，包括：

计算网络链路可预测性的上界 尽管文献[67]给出了刻画网络链路可预测性的方法，但是文中提出的结构一致性指标仅能帮助我们比较两个网络的链路可预测性的高低，并不能给出一个可预测性的上界。如何得到一个不依赖于预测算法的链路可预测上界仍是一个悬而未决的问题。

复杂类型网络的链路预测 现有的研究并不系统，如何在复杂结构的网络（例如含时网络、多层网络、相互依赖网络、超网络等）中进行链路预测，还有很大空间可供挖掘。对于含时网络的研究仍属凤毛麟角，我们知道如何应用时间的信息提高存在性的预测精度，但是如何准确预测两节点再次接触的时间仍是一个挑战，人类动力学的研究在这方面会起到重要的作用。

大规模网络的快速算法设计 信息技术的发展使得我们获取数据越来越容易，然而大数据所具有的海量复杂的特征也对信息处理技术提出了更高的要求。如何在超大规模网络上有效利用多维的丰富信息进行快速、准确的链路预测，设计局域化或并行化的算法具有重要意义。

链路预测的应用问题 链路预测已广泛应用于社交网络、生物网络、推荐系统、股市预测等多个领域。相信随着研究的不断推进，会发现更多可以应用的场景，从而体现链路预测在解决实际问题中的价值。值得注意的是，用户的个人隐私保护是实际应用中一个不可回避的问题，如何在分散式在线社交网络中进行链路预测，使得既不侵犯用户的

隐私又能得到精度较高的预测效果，这也是一个值得探讨的问题。

评价指标体系的建设 目前大多数的算法评价都是基于离线的数据测试，但是真实系统中的用户行为往往受到多种因素的影响，因此获得真实系统中用户的反馈对于评价算法的性能更加重要。在复杂类型网络的研究进展中，如何针对这些网络的预测建立更加合理的评价指标也是亟待解决的问题。例如在含权网络的链路预测中，就权重预测的准确性而言，用一些刻画相关性的指标是否合适，还是值得商榷的。

链路预测研究方兴未艾。近年来，涌现出一批具有理论深度的研究成果。特别是一批年轻学者的加入，促使链路预测成为网络信息挖掘和预测领域最具活力的研究方向之一。作为一门典型的交叉学科，链路预测从理念、方法到应用，融合了计算机科学、统计物理学、社会科学、生物学、数学等多个学科的精华。未来希望能在多学科的交叉融合中碰撞出更多有意义、有价值的思想火花，共同推动链路预测在理论和应用方面取得进展。■

致谢：感谢国家自然科学基金11205042、61433014和11222543，浙江省自然科学基金LR16A050001的支持。

吕琳媛

CCF专业会员。杭州师范大学教授。主要研究方向为链路预测、推荐系统和节点排序等。linyuan.lv@hznu.edu.cn

任晓龙

杭州师范大学研究生。主要研究方向为复杂网络链路预测和节点排序等。
renx868@gmail.com

周　涛

CCF专业会员。电子科技大学教授。主要研究方向为统计物理、复杂性科学。
zhutou@ustc.edu

参考文献

[1] Watts D J and Strogatz S H. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440~442, 1998.

[2] Barabási A L and Albert R. Emergence of scaling in random networks. *Science*, 286(5439):509~512, 1999.

[3] Newman M J and Girvan M. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004

[4] Castellano C and Pastor-Satorras R. Thresholds for epidemic spreading in networks. *Physical review letters*, 105(21):218701, 2010.

[5] Fortunato S. Community detection in graphs, *Physics Reports*, 486(3):75~174, 2010.

[6] Clauset A, Moore C, and Newman M J. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98~101, 2008.

[7] Borgatti S P, Mehra A, Brass D J, et al. Network analysis in the social sciences. *Science*, 323(5916)892~895, 2009.

[8] 任晓龙, 吕琳媛. 网络重要节点排序方法综述. 科学通报, 59(13)1175~1197, 2014.

[9] Liben-Nowell D and Kleinberg J. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019~1031, 2007.

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

[11]吕琳媛, 陆君安, 张子柯等. 复杂网络观察. 复杂系统与复杂性科学, 7(2-3):173~186, 2010.

[12]吕琳媛, 周涛. 链路预测. 北京: 高等教育出版社, 2013.

[13]Mamitsuka H. Mining from protein-protein interactions. Wiley Interdiscip. Rev. Data Min. Knowl. Discov., 2(5):400~410, 2012.

[14]Brzozowski M J and Romero D M. Who should I follow? Recommending people in directed social networks. in International AAAI Conference on Web and Social Media (ICWSM), 2011.

[15]Hasan M AI and Zaki M J. A survey of link prediction in social networks. in *Social network Data Analytics*, Springer, 2011: 243~275.

[16]Leskovec J, Huttenlocher D, and Kleinberg J. Predicting positive and negative links in online social networks. in *Proceedings of the 19th international conference on World wide web*, 2010:641~650.

[17]Lü L, Medo M, Yeung C H, et al, Recommender

systems. Physics Reports, 519(1):1~49, 2012.

[18]Ren X-L, Lü L, Liu R, et al, Avoiding congestion in recommender systems. New Journal of Physics, 16(6):63057, 2014.

[19]Guimerà R and Sales-Pardo M. Missing and spurious interactions and the reconstruction of complex networks. PNAS, 106(52): 22073~22078, 2009.

[20]Kaya B and Poyraz M. Unsupervised link prediction in evolving abnormal medical parameter networks. International Journal of Machine Learning and Cybernetics, 2015:1~11.

[21]Chen S, Lan X, Hu Y, et al. The time series forecasting: from the aspect of network. arXiv: 1403.1713, 2014.

[22]Chen Z, Xie Z, and Zhang Q. Community detection based on local topological information and its application in power grid. Neurocomputing, 170:384~392, 2015.

[23]Wang W-Q, Zhang Q-M, and Zhou T. Evaluating network models: A likelihood analysis. EPL (Europhysics Letters), 98(2):28004, 2012.

[24]Zhang Q-M, Shang M S, and Lü L. Similarity-based classification in partially labeled networks. International Journal of Modern Physics C, 21(06):813~824, 2010.

[25]Dasgupta K, Singh R, Viswanathan B, et al. Social ties and their relevance to churn in mobile telecom networks. in Proceedings of the 11th international conference on Extending database technology: Advances in database technology, 2008:668~677, 2008.

[26]Kossinets G and Watts D J, Empirical analysis of an evolving social network. Science, 311(5757): 88~90, 2006.

[27]Leskovec J and Horvitz E. Planetary-scale views on a large instant-messaging network. in Proceedings of the 17th International Conference on World Wide Web, 2008:915~924, 2008.

[28]Akcora C G, Carminati B, and Ferrari E. User similarities on social networks. Social Network Analysis and Mining, 3(3):475~495, 2013.

[29]Navarro G. A guided tour to approximate string matching. ACM Computing Surveys (CSUR), 33(1):31~88, 2001.

[30]Aiello L M, Barrat A, Schifanella R, et al. Friendship prediction and homophily in social media. ACM Transactions on the Web (TWEB), 6(2):9, 2012.

[31]Zhou T, Lü L, and Zhang Y-C. Predicting missing links via local information. The European Physical Journal B, 71(4):623~630, 2009.

[32]Leicht E A, Holme P, and Newman M J. Vertex similarity in networks. Physical Review E, 73(2):26120, 2006.

[33 Adamic L A and Adar E. Friends and neighbors on the web. Socical Networks, 25(3):211~230, 2003.

[34]Liu Z, Zhang Q-M, Lü L, and Zhou T. Link prediction in complex networks: a local naïve Bayes model. EPL (Europhysics Letters), 96(4):48007, 2011.

[35]Lü L, Jin C-H, and Zhou T. Similarity index based on local paths for link prediction of complex networks. Physical Review E, 80(4):46122, 2009.

[36]Katz L. A new status index derived from sociometric analysis, Psychometrika, 18(1):39~43, 1953.

[37]Klein D J and Randić M. Resistance distance. Journal of Mathematical Chemistry, 12(1):81~95, 1993.

[38]Fouss F, Pirotte A, Renders J-M, et al. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation, Knowledge and Data Engineering, IEEE Transactions on, 19(3):355~369, 2007.

[39]Brin S and Page L. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1):107~117, 1998.

[40]Jeh G and Widom J. SimRank: a measure of structural-context similarity. in Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002:538~543.

[41]Liu W-P and L Lü. Link prediction based on local random walk. EPL (Europhysics Letters), 89(5):58007, 2010.

[42]Chebotarev P and Shamis E. The matrix-forest theorem and measuring relations in small social groups. Automation and Remote Control, 58(9):1505~1514, 1997.

[43]Sun D, Zhou T, Liu J-H, et al. Information filtering based on transferring similarity. Physical Review E, 80(1):17101, 2009.

[44]Wang P, Xu B, Wu Y, et al. Link prediction in social networks: the state-of-the-art. Science, China Information Sciences, 58(1):1~38, 2015.

[45]Hasan M AL, Chaoji V, Salem S, et al. Link prediction using supervised learning. in SDM’06: Workshop on Link Analysis, Counter-terrorism and Security, 2006.

[46]Yuan G, Murukannaiah P K, Zhang Z, et al. Exploiting sentiment homophily for link prediction, in Proceedings of the 8th ACM Conference on Recommender systems, 2014:17~24.

[47]Backstrom L and Leskovec J. Supervised random walks:

predicting and recommending links in social networks. In Proceedings of the fourth ACM international conference on Web search and data mining, 2011:635~644.

[48]Kunegis J and Lommatzsch A. Learning spectral graph transformations for link prediction, in Proceedings of the 26th Annual International Conference on Machine Learning, 2009:561~568.

[49]Menon A K and Elkan C. Link prediction via matrix factorization. in Machine Learning and Knowledge Discovery in Databases, Springer, 2011:437~452.

[50]Chen Z and Zhang W. A Marginalized Denoising Method for Link Prediction in Relational Data. in Proceedings of the 2014 SIAM International Conference on Data Mining, 2014:9, 2014.

[51]Guo F, Yang Z and Zhou T. Predicting link directions via a recursive subgraph-based ranking. Physica A: Statistical Mechanics and its Applications, 392(16):3402~3408, 2013.

[52]Leung C W, Lim E P, Lo D, et al. Mining interesting link formation rules in social networks. in Proceedings of the 19th ACM International Conference on Information and Knowledge Management, 2010:209~218, 2010.

[53]Zhang Q-M, Lü L, Wang W-Q, et al. Potential theory for directed networks. PLoS ONE, 8(2):e55437, 2013.

[54]Lichtenwalter R N, Lussier J T, and Chawla N V. New perspectives and methods in link prediction, in Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ' 10, 2010:243, 2010.

[55]Narayanan A, Shi E, and Rubinstein B I P. Link prediction by de-anonymization: How we won the kaggle social network challenge. in Neural Networks (IJCNN), The 2011 International Joint Conference on, 2011:1825~1834, 2011.

[56]Ahmad M A, Borbora Z, Srivastava J, et al. Link prediction across multiple social networks. in Data Mining Workshops (ICDMW), 2010 IEEE International Conference on, 2010:911~918.

[57]Corlette D and Shipman F M III. Link prediction applied to an open large-scale online social network. in Proceedings of the 21st ACM Conference on Hypertext and Hypermedia, 2010:135~140.

[58]Murata T and Moriyasu S. Link prediction of social networks based on weighted proximity measures. in Web Intelligence, IEEE/WIC/ACM International Conference on, 2007:85~88.

[59]Lü L and Zhou T. Link prediction in weighted networks: The role of weak ties. EPL (Europhysics Letters), 89(1):18001, 2010.

[60 Meng B, Ke H, and Yi T. Link prediction based on a semi-local similarity index. Chinese Physics B, 20(12):128902, 2011.

[61]Psorakis I, Roberts S, Ebden M, et al. Overlapping community detection using bayesian non-negative matrix factorization. Physical Review E, 83(6):66114, 2011.

[62]Karrer B and Newman M J. Stochastic blockmodels and community structure in networks. Physical Review E, 83(1):16107, 2011.

[63]Dunlavy D M, Kolda T G, and Acar E. Temporal link prediction using matrix and tensor factorizations. ACM Transactions on Knowledge Discovery from Data, 5(2):10:1~10:27, 2011.

[64]Lahiri M and Berger-Wolf T Y. Structure prediction in temporal networks using frequent subgraphs, in Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on, 2007:35~42.

[65]Oyama S, Hayashi K, and Kashima H. Cross-temporal link prediction. in Data Mining (ICDM), 2011 IEEE 11th International Conference on, 2011:1188~1193.

[66]许小可, 许爽, 朱郁筱等. 复杂网络中链路的可预测性. 复杂系统与复杂性科学, 11(1):42~47, 2014.

[67]Lü L, Pan L, T Zhou. Toward link predictability of complex networks. PNAS, 112(8):2325~2330, 2015.