复杂网络大数据分析中的链路预测：局部路径指标的精度分析

农昂

# 摘要

在我们的生活中，很多研究领域都可以抽象成复杂网络，每个对象都是网络中的节点，而对象之间的关系可以用边来表示。而链路预测，是复杂网络研究的一个重要手段，具有巨大的科研与应用价值。具体来讲，链路预测就是指如何通过利用已经知道的网络节点以及网络结构等信息来预测网络中尚未连边的两个节点之间未来产生链接的可能性，实际上是一种数据挖掘的过程。近年来，该方向的研究获得了很多的成果，学者们也提出了很多优秀的链路预测算法。在本文中，我们主要对局部路径指标的链路预测算法的精度进行分析。通过对六个真实网络进行大量的数值模拟表明，与两个众所周知且广泛使用的指标:公共邻居指标和Katz指标相比，局部路径指标具有较高的准确率和效率。事实上，局部路径指标提供了与Katz指标一样具有竞争力的预测精度，同时比Katz索引需要更少的CPU时间和内存空间，因此局部路径指标是大型网络数据挖掘中潜在实际应用的有力候选。

**关键词：复杂网络 链路预测 局部路径**

# Abstract

In our life, many research fields can be abstracted into complex networks, each object is a node in the network, and the relationship between objects can be represented by edges. Link prediction, as an important means of complex network research, has great value in scientific research and application. Specifically, link prediction refers to how to predict the possibility of future link between two nodes in the network that have not yet been linked by using the known network nodes and network structure information, which is actually a process of data mining. In recent years, a lot of achievements have been made in this field, and many excellent link prediction algorithms have been proposed by scholars. In this paper, we mainly analyze the accuracy of link prediction algorithm based on local path index. A large number of numerical simulations on six real networks show that the local path index has higher accuracy and efficiency than the two well-known and widely used indexes: the public neighbor index and the Katz index. In fact, the local path index provides the same competitive prediction accuracy as the Katz index, while requiring less CPU time and memory space than the Katz index. Therefore, the local path index is a strong candidate for potential practical applications in large-scale network data mining.

**Key words: complex network link prediction local path**

# **目录**

目录

[摘要 1](#_Toc22716)

[Abstract 1](#_Toc11507)

[目录 2](#_Toc4694)

[第一章 绪论 2](#_Toc10183)

[1.1研究背景及意义 2](#_Toc5014)

[1.2链路预测问题研究现状 3](#_Toc5368)

[1.3本文主要研究内容 7](#_Toc9943)

[第二章 相关理论介绍 8](#_Toc16859)

[2.1 图及其类型 8](#_Toc1754)

[2.2 网络的拓扑性质 9](#_Toc7657)

[2.3链路预测理论概述 12](#_Toc28169)

[第三章 实验结果及分析 14](#_Toc10095)

[3.1实验设置及实验环境说明 14](#_Toc224)

[3.2实验数据集 17](#_Toc6863)

[3.3结果与分析 18](#_Toc29863)

[第四章 结论 19](#_Toc24803)

[致谢 19](#_Toc7384)

# **绪论**

本章主要是介绍该课题的研究背景和意义、领域内的国内外研究现状以及本文的组织结构。

## 1.1研究背景及意义

联系，是一种特征，体现了不同个体之间的相互关联，在各种系统中广泛存在。复杂网络是各种真实系统的抽象表示。其中，网络中的节点代表了真实系统中的个体，连边代表了个体之间的相互作用或者关系。随着网络信息技术的飞速发展，在我们的日常生活中复杂网络已经成为非常普遍的存在，如社交网络、人脑神经网络、基因调控网络、生物网络、通信网络、合作网络、万维网等。换句话说，小到能量粒子，大到宇宙中的天体行星，网络都在以其强大的表达能力在影响同时也在塑造着这个世界。能否确保各种复杂网络系统安全可靠的运行，将直接影响到人类的生活和生产活动。因此，复杂网络的研究已经成为许多学科共同关注的研究问题。

链路预测作为复杂网络研究的重要分析手段之一，其再在推动网络科学和信息科学的发展方面有着非常大的作用。在一个真实的网络中通常存在着丰富多样的隐含信息。链路预测就是指通过利用已经知道的网络节点和网络结构等信息，去预测网络中尚未存在连接的两个节点之间未来可能产生新连接的可能性。预测的目标一般来说可以分为两种，一种是在网络中真实存在但还没有被发现的连接；另一种为网络中将来才可能出现的连接。简单来说，链路预测探讨的是复杂网络中不同节点之间是否存在连接的问题。

链路预测问题的研究也因为其重大的实际应用价值，受到了不同背景和不同领域的研究学者们的广泛关注。例如，在生物研究领域的蛋白质网络中，将一个一个的蛋白质抽象成节点，彼此之间的相互关系抽象成边，一般情况下，需要大量的实验才能判断某条边是否存在。然而，目前人们所已经知道的蛋白质间的相互作用还不到0.5%。通过大量的实验来寻找尚未被发现的链接需要消耗非常高的人力、物资和时间成本。所以如果我们能够通过已经知道的蛋白质网络结构，设计出一种足够合理并且准确度足够高的链路预测算法，通过根据算法的运行结果对实验进行指导，就可以实现在提高实验成功率以及降低实验花费的成本的同时，加快揭示该网络的本质的目的。另外，在社会网络中经常会出现信息结构缺失的情况，这时候就可以通过链路预测将整个网络的拓扑结构补全，然后再进行下一步的网络分析。其次，链路预测还能运用在演化网络以及那些能够抽象为网络形式的系统中，对未来的链路的形成进行预测。这对于在线社交网络和电商网络等商业应用的价值非常巨大。比如在社交网络中，可以通过网络的拓扑结构等信息来预测可能产生的链接，从而推荐现实中可能认识的人互加好友，从而提高用户对产品的忠诚度。此种方法也同样适用于电商网络。最后，链路预测在网络中对错误链接的发现，对网络结构的优化，乃至于网络的重组都有非凡的意义，比如在构建蛋白质网络过程中处理一些关系不明朗或者是自相矛盾的数据。

综上所述，作为一项基础的研究，链路预测为人类揭开网络科学本质的面纱奠定了理论的基石，并且极大地推动了我们探索自然科学的进程。同时，它还在生活、生产的各个领域中都得到了普遍的应用，带来可观的经济效益。因此，无论是在科学研究领域，还是工程领域，其研究价值都值得肯定。

## 1.2链路预测问题研究现状

从无标度网络以及小世界网络模型的提出算起，发展到现在，对于复杂网络的研究已经持续了近 20 年。这20 年以来，通过各路学者的不懈努力，让这个原本由物理学家开创的学科日渐壮硕。现在的复杂网络包含着数学、物理及计算机等多个学科的内容，是一门多个领域交叉的复合学科。毫无疑问，随着技术和信息的蓬勃发展，复杂网络已经成为了计算机领域数据挖掘的一个重要来源。而在数据挖掘领域中，链路预测已然是研究的热点方向之一，因此它在计算机领域中的研究也在得到不断的突破，方法大致分类如图 1-2。

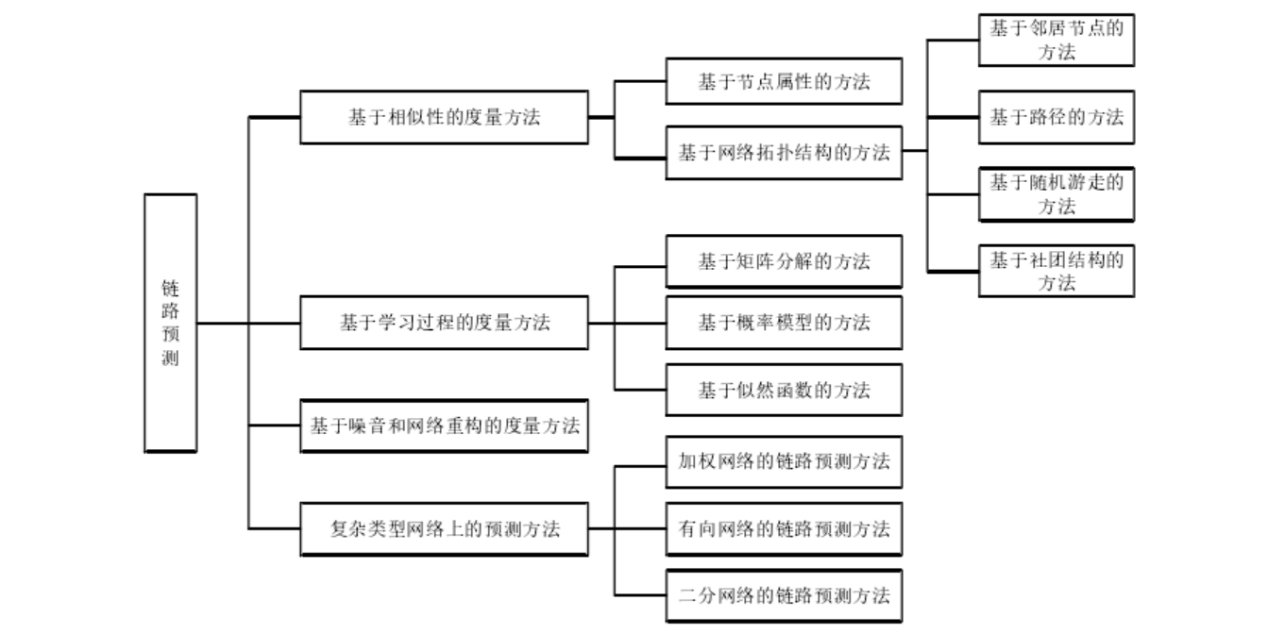


图1-2 链路预测方法分类图

最早期的链路预测研究是基于马尔科夫链和机器学习的思路而展开，主要是通过利用节点的属性信息，具有较为良好的预测效果，但在大多数的情况下，由于受到保密性的约束，使我们想要获取到可靠的节点属性信息面临着巨大的困难。正是由于这样的原因，很多学者不得不去寻找一种更容易实现的方法进行链路预测，而通过不断的思考，学者们提出了基于网络结构进行预测的链路预测方法的研究。相比于节点信息，网络结构的信息我们更容易能获得到，并且具有更好的可靠性，同时此类方法能很好的适用于那些结构相似的网络。一般来说，利用了包括节点或实体的属性以及网络的拓扑两种类型的信息的算法统一被称为是基于相似性的一类算法，它的基本假设是，如果两个节点具有较多的共同点，那么它们之间则更有可能形成链路。若以研究网络范围大小来进行划分，这些算法大致可以分为三类，分别为：（1）基于网络局部结构的预测算法；（2）基于网络全局结构的预测算法；（3）基于准局部结构的预测算法。

首先，基于网络局部结构的预测算法是最直观且相对有效的，此类算法仅仅只是利用节点的邻居信息，拥有较低的计算复杂度，非常适合运用于大规模网络应用。但是受到信息量的限制，它的预测精度往往不太理想，比如共同邻居（Common Neighbors，简称 CN）、偏好连接（Preferential Attachment，简称 PA）、Adamic-Adar （简称 AA）、Jaccard（简称 JC）、Salton和资源分配（Resource Allocation，简称 RA）等相似性指标。相比之下，考虑了整个网络的结构全局结构相似性算法获得了更多的信息，使得预测精度有了很大的提高，但由于需要计算全局的节点，因此计算时间也大大增加了，时间复杂度较高。代表算法有基于一般等价的 LHN-II（Leicht Holme-Newman-II）指标、全局路径指标 Katz、有重启的随机游走指标（Random Walk With Restart，简称 RWR）、平均通勤时间（Average Commute Time，简称 ACT）、SimRank 指标（简称 SimR）等。而基于准局部结构的预测算法，则是可以看作是以上两种算法的中和，它考虑了比局部指标更多的因素，但又不需要计算全局网络拓扑的信息量，成功实现了算法精度以及时间复杂度的兼顾，如局部随机游走指标（Local Random Walk，简称 LRW）、局部路径相似性指标（Local Path，简称 LP）、有叠加效应的局部游走指标（Superposed Random Walk，简称 SRW）。

近期在基于相似性的算法研究方面，取得的研究成果有：贾等人针对引用网络提出一个新的指数：H指数，并用该指数代替了度，采用加权的方式有效衡量了引文的重要性，同时增强了 Sorenson、Salton和 AA三种经典的链路预测方法，并进行大量的数据实验，实验结果表明在引文网络中该方法取得了较为良好的表现。Cannistraci 等人将共同邻居节点与经典相似性指标之间的联系进行了强化，设计出了 Cannistraci 资源分配相似性预测算法CAR（Cannistraci-Resource-Allocation）。周等人基于知识是依靠网络的路径进行传播的这一前提做出假设，将复杂网络中知识的传播机制（Knowledge Dissemination Mechanism）进行了定义。基于这个假设，吕等人又将节点所具有的知识量等价为它的 H 指数，定义了网络中的任意一条真实连边的权重等于由起点传输的知识占终点从它的邻居节点所得到的全部知识的比重，然后再结合节点间的全部路径信息，提出了 KDLP（Link Prediction based on Knowledge Dissemination）全路径预测指标。三元闭包是网络中最小的局部结构，高等人基于此结构，提出了加权的 CN、AA 以及 RA 指标，分别为 TWCN（Common Neighbors with Triadic Closure Weight）、TWAA（Adamic-Adar with Triadic Closure Weight）、TWRA（Resource Allocation with Triadic Closure Weight），对应还有 TWCN\*、TWAA\*、TWRA\*3个具有调节参数的相似性指标。这项工作是通过计算出每一个节点在网络中所占相邻的三元闭包的权重，并再计算节点相似性指标时将该权重加入，提高了预测精度，但时间复杂度却大幅提升，上升到 O（N3）。刘等人从资源交换的角度出发，提出了扩展资源分配指标 ERA（Extended Resource Allocation Index），这个指标是通过在两个端点之间的非共同邻居和公共邻居交互的资源量来测量节点间的相似性。孙等人提出了局部亲和结构指标（Local Affinity Structure，简称 LAS），通过节点的度和其邻居的信息来显示该节点对和它的共同邻居间的亲密关系,并且通过大量的实验证明了算法的可行性。但是这个方法仅仅只是利用了节点的信息，并没有将链路对于相似性的贡献考虑进来。武等人则直接根据节点的聚类系数设计了一个全新的预测指标 CCLP（Clustering Coefficient for Link Prediction），取名为基于聚类系数的链路预测指标，实验证明，再在CN的基础上额外考虑节点聚类系数的方法在预测网络中缺失的链接方面具有较好的有效性，但是它仍然采用的是通用估计，同样是没有考虑到特定预测节点对的局部网络环境。在这之后，武等人对这个算法进行一定的扩展，认为当公共邻居节点的聚类系数相同时，不同节点对的贡献并不一定相同，在此基础上提出了基于节点及边聚类信息的预测大型网络缺失链路的NLC（Similarity Indices based on Node and Link Clustering Information）算法，在计算相似性指标时将节点聚类信息及边聚类信息考虑进来，这种方式更加有效的利用到了共同邻居在在链路预测中的价值，取得了较为不错的实验效果。

最大似然估计类算法是先制定出假定的规则与参数，然后通过将其与已知的结构进行最大概率的结合来预测网络中那些缺失的链路。目前，这类算法主要存在两种，一种是基于网络层次结构信息的最大似然法（Hierarchical Structure Model，简称 HSM），是由 Clauset 等人在 2008 年提出。该方法对两节点间的连接概率是通过二者最近的共同祖先的存在概率值进行判断的。先是用一组 M 个叶子节点，M−1 个内部节点的族谱树来表示一个网络（如图 1-3），再通过调整族谱树来获得这个网络的最大似然值及其对应的所有的内部节点的概率值，该连边的预测概率就是一组族谱树中两个节点所对应的共同祖先概率的均值。另一种是随机块模型方法（Stochastic Block Model，简称 SBM）由Guimerà提出。如图 1-4 所示，在该模型中，集群的密切度是决定两节点连接的关键因素。//todo另一种最大似然方法是闭路模型方法（Loop Model，简称 LM），是由吕等人提出的。寻找出网络中的闭合回路，并将其看成是一种局部性，网络中的似然值通过闭合环路的多少来进行定义，某缺失边存在概率的预测值就是添加该边后网络的似然值。总的来说，基于最大化网络似然函数的这类链路预测算法的优点是可以将缺失的链接识别出来，同时还能够将网络更深层次的特征挖掘出来，但存在的问题是这类算法一般都具有较高的时间复杂度，因为它对网络顶点的个数非常敏感。还有一种是基于矩阵分解的预测算法，该方法主要是通过对网络数据中存在的潜在特征进行学习，然后进行链接预测。先是把网络中的节点投射到一个潜在的空间上，链路的存在概率就由节点在该空间的位置来进行衡量。潜在空间的每一个特征都可以被认为是可能存在的属性，如果两个节点存在越多相似的潜在特征，那么就认为它们之间就更有可能是相似的。从另一个视角出发，将复杂网络的相似性矩阵看成是两个低维特征矩阵的乘积，一个对应基矩阵，另一个对应系数矩阵，将它们的元素都限制为非复数，那么我们也就可以利用非负矩阵分解的算法来进行求解。但这类方法也存在特征数量较难确定的问题。

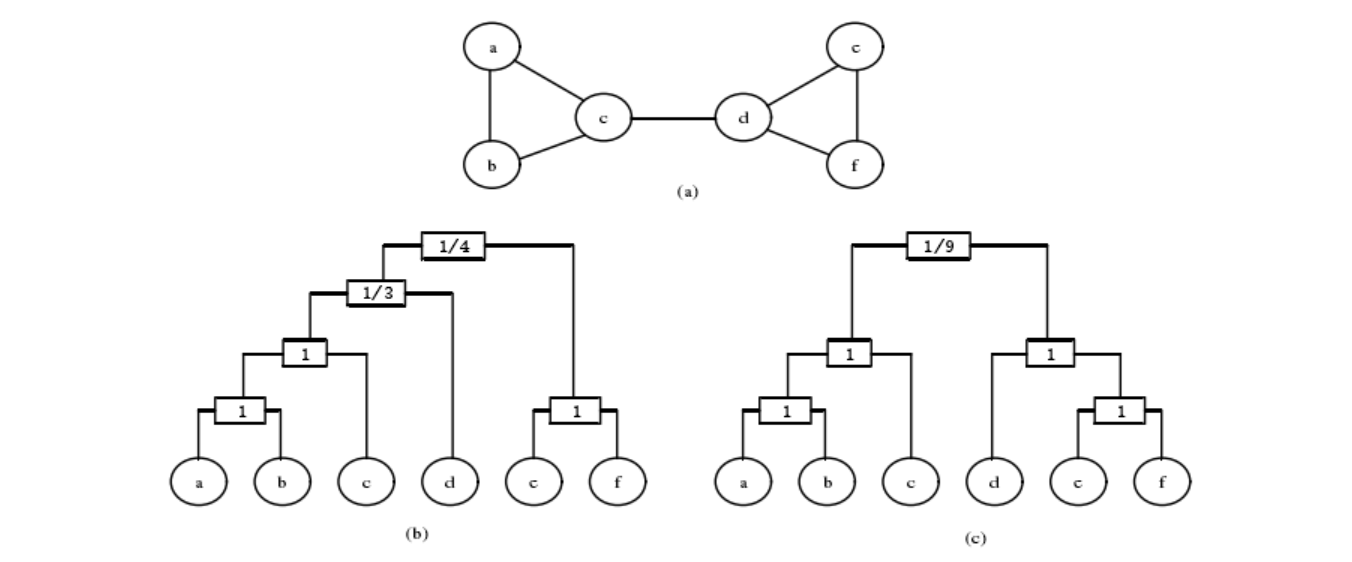


图 1-3 网络（a）与族谱树（b），（c）

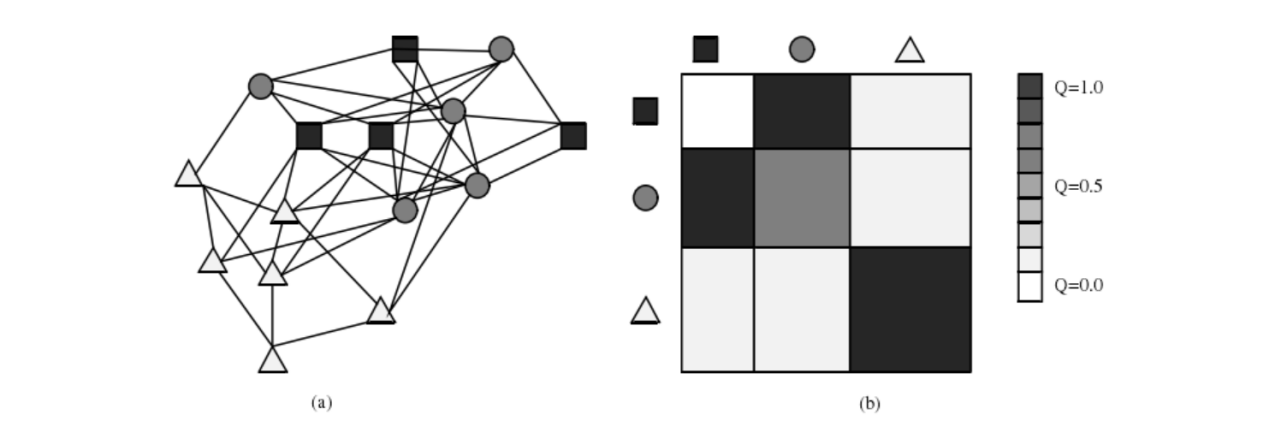


图 1-4 网络的划分（a）与随机块模型（b）

相关还有一些基于网络的噪音过滤及进行网络重构的链路预测方法。由于此类方法能对网络进行一定的噪音过滤处理，使预测指标具有较高准确率的同时还能具备较好的健壮性，因此即使在网络中含有噪音，也能取得较好的表现。如Zhang 等人提出了加权的局部路径预测指标。Zeng等人提出了一种混合共同邻居以及边中心性的预测模型，通过实验表明该预测模型在预测准确率上表现的较为良好，并且在进行网络重构后，网络的结构特性和原来的网络保持一致。吕等人又在扰动网络邻接矩阵的方法基础上，提出了一种结构一致性预测指标（Structural Consistency Index，简称 SCI）。类似的工作还有噪音环境下链路预测健壮性的评价方法和基于噪音过滤的链路预测方法等。

目前，链路预测算法方面的研究更多的是针对于中小型规模的网络，而对大规模网络的链路预测研究还是比较稀缺的。同时，学术界目前的主流研究方向依然是对离散邻接矩阵形式表示的网络来进行链路预测的研究。然而，随着科学技术的飞速发展，大规模信息网络时代即将来临，这钟网络表示方式也出现了越来越多的问题。一方面来说，离散邻接矩阵表示的网络能获取到的信息较少，只能捕获到顶点之间的相邻关系，因此当需要构建当下复杂网络中更复杂的高阶结构关系，比如路径、频繁的子结构等，则会因信息不足而略显薄弱，同时对于体现节点所携带的属性的外部信息方面也存在着较大的不足。另一方面，系统在面对大规模的网络时，需要具备更高的处理器性能和存储要求。基于以上原因，部分学者逐渐对网络表示的方式进行的深入研究，希望可以为链路预测的研究找到一种新的视角和方法。遗憾的是，该方向的研究进展并不是很顺利，到目前为止，大多数的网络表示算法都是和任务无关的实现，在面对针对网络分析任务的研究成果时仍惧挑战，而利用网络表示学习的技术实现链路预测问题的工作更是屈指可数。

## 1.3本文主要研究内容

链路预测的研究存在许多困难。一个是目标网络的稀疏性，这将导致了一个严重的问题，即链路的先验概率通常很小，导致在建立统计模型时有很大的困难。另一个问题是真实系统的巨大网络规模需要高效的算法。然而，作为实际应用中的一个关键因素，计算时间和内存的复杂性尚未得到系统的研究。一般来说，一个算法的精度与其计算复杂度是正相关的，即精度越高通常意味着计算复杂度越高。需要注意的是，如果消耗的时间或内存是不可接受的，那么任何高度精确的算法都将变得毫无意义。因此，设计一个准确又快速的算法是一个很大的挑战，尤其是对于稀疏和庞大的网络。

在本文中，我们主要是对局部路径（LP）指标链路预测算法进行分析研究。对模拟网络和真实网络的大量数值模拟实验表明，该相似性指标同时兼顾精度与效率，其预测精度远高于普通邻居（CN）指标，与Katz指数具有竞争力，并且高效，计算所需的时间和空间远小于Katz指数。特别是，当网络很大时，局部路径指标与Katz指标相比显示出很大的优势，因为计算后者要求CPU时间缩放为网络大小的立方，而计算前者要求线性CPU时间作为网络大小。最后，我们通过在不同领域及不同规模的网络数据集上进行实验，并将实验结果与CN、Kazt算法对比，来证明 LP指标算法在链路预测问题上具有相对的竞争力。

# 第二章 相关理论介绍

## 2.1 图及其类型

网络描述的是对象与对象之间的关系，本质上是一种集合。在进行网络的研究时，一般是将其抽象为图的结构。图的节点就是网络中的对象，图中的边则是各个对象之间的联系。可以说，图是进行复杂网络研究的一种重要工具，能帮助我们对网络的结构进行更好的理解。

通常，网络都可以抽象成图。图在数学上的定义为 G=(V,E)，其中V是由网络中一个一个的顶点组成的集合，E 则是网络中联系的边的集合。假设定义一个网络，在这个网络中存在的节点总数为Y，边的数量总共为X，而节点对组成的全集是P，那么存在有Y=|V|，X=|E|,P=N(N-1)/2。在后面的描述中，本篇文章对网络和图、节点和顶点以及链接和边三组的用词不做任何区分，每个组对应的两个词概念等价。根据连边的有权、无权以及有向和无向，网络所对应的图还可以被分为无权有向图、无权无向图、有权有向图和有权无向图四类。四种不同类型图的示意如图 2-1，它们之间的关系如图 2-2。在这里，我们用权来表示节点与节点之间连接的强弱程度。计算机实际分析中，常用于表示图结构信息的方式主要有两种：邻接表和邻接矩阵，本篇文章主要重点介绍后者。对于给定的含有 M个节点的简单网络，其邻接矩阵为一M阶方阵，用数学表示为 A={bij}M\*M。在矩阵A中第i行第j列的元素bij的定义由网络类型而决定，不同的网络类型会有所区别。

当一个图中的连边没有权重和方向的关系时，这样的图被称为无向无权图，如图 2-1（a），比如现实在线社交网络中用户间的好友关系。这一类网络的邻接矩阵 A 中 bij的定义为：

（2.1）

当一个图中的连边由方向但是没有权重，这样的图被称为有权无向图，如图2-1（b）。该种类型的图表示网络中的个体之间的联系只有方向的区分，没有强弱的区别。比较常见的有B站、微博中的用户关注关系。这个时候bij的定义为：

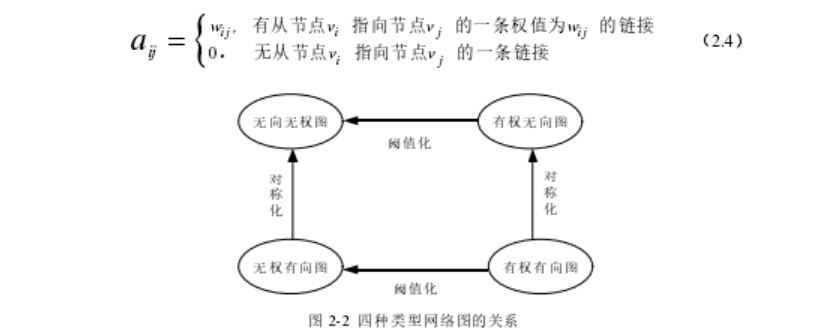
（2.2）

当一个图中的连边存在权重但是没有方向，这样的图被称为有权无向图，如图2-1（c），表示在该网络种节点之间的关系没有方向的区分，只有强弱的不同。科学家合作网络是这类网络的就是这样的网络，两个科研人员之间的连边没有方向的区分，而边的权重则是二者一起合作发表文章数量的体现。网络对应的 bij定义为：

（2.3）

当一个图中的连边既存在权重又存在方向的关系时，这样的图被成为有权有向图，如图2-1（d）。在现实生活中，航空网络就是这类网络。每个机场构成图中的节点，而图中节点间的连边就代表了机场与机场之间的航向，航班的频率就是边的权重。邻接矩阵 A 中 aij的定义如下:

（2.4）



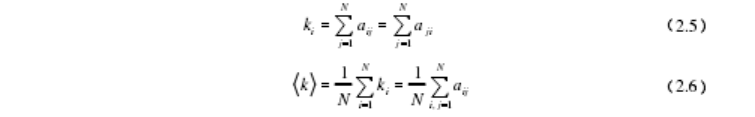
如果没有特别的说明，本文所提到的网络将都是无向无权的网络，并且不考虑网络中节点之间的重边或自环现象。

### **2.2 网络的拓扑性质**

网络的拓扑性质是指不受节点的尺寸、形态、方位、效用以及彼此间的连接方式等因素影响，仅直接相关于网络中的节点数目和链接情况的一类性质。本节仅就部分基本的拓扑性质简要介绍。

2.2.1 度、平均度、网络密度与度分布

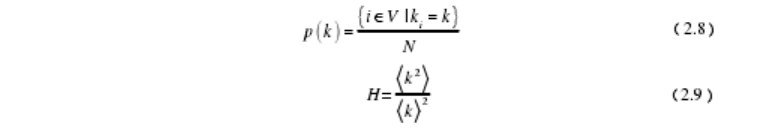
描述网络节点一个最根本也是最重要的属性即为度（Degree）。给定无权无向网络G=(V,E)，节点 vi的度定义为包含节点 vi的边的数目，记为 ki。网络 G 的平均度<k>，则定义为网络中所有节点度的平均值。对于 G 的邻接矩阵 A={aij}N\*N，ki和<k>的计算表达式分别为：



在边具有方向性的网络中，节点 vi的度又被区分为出度和入度。其中，出度表示始于该节点，终于其他节点的边数，而入度的含义与其相反。网络密度以包含 N 个节点的网络中实际存在的链接数 M 与理论上最大可能的链接数的比值来度量，记为



度分布 p（k），定义为网络中节点度为 k 的节点数占节点总数的比例，如公式（2.8）。现实生活中，网络的度分布是多种多样的，代表性的有幂律分布、高斯分布、指数分布等。其中，以具有幂律分布的网络实例最为广泛，如万维网、电影合作网络、电网等。符合这类网络的一个明显特征就是：度非常小的节点占绝大多数，只有小部分节点的度非常大。值得一提的是，衡量网络中节点度分布均匀程度的指标称为度异质性（Degree Heterogeneity），通常记为 H，定义为公式（2.9）。



除了度分布外，联合度分布（Joint Degree Distribution）同样是网络的统计特性之一，其定义为从网络中随机选择端节点分别有 k 和 k'个邻居的一条边的概率，即



### 2.2.2 网络的连通性与路径

若某网络 G=（V,E）中，所有节点对间都至少存在一条路径使其产生联系，则称该网络 G 是连通的，反之，则不连通。其中，网络 G 中的一条路径（Path）p 是指一个节点序列 p={v0,v1,v2,…vl}，且满足（vi,vi+1）∈E（0≤i<l），这里 v0和 vl为路径 p 的两个端点[46]。路径 p 的长度即为从 v0到 vl所经过的节点数加一。 网络中两个节点之间给定长度的路径数目可通过网络的邻接矩阵来表示。对于给定的节点对（vi,vj）,若其存在连边，则 aij=1，表示二者之间存在一条长度为 1 的路径。由于对于节点对（vi,vj）之间长度为 2 的任意路径 p={vi,vk,vj},有 aikakj=1，故节点对（vi,vj）之间长度为 2 的不同路径总数可表示为：



以此类推，节点对（vi,vj）之间存在长度为 l 的不同路径的数量为：

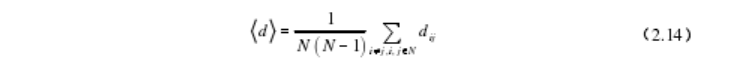


### 2.2.3 直径、平均路径长度与效率

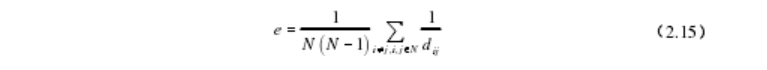
在复杂网络中，将两个节点联系起来的路径通常不止一条。其中，将经过最少的点实现两节点（vi和 vj）之间联系的路径称之为最短路径，其对应的路径长度称为两节点间的距离，记为 dij。网络的直径 D[66]则定义为任意两个节点间距离最大的路径长度，即



对于非完全连通的网络，其直径则由任何两个存在有限距离的节点之间的距离的最大值来表示。 作为网络中一个重要全局属性，网络的平均路径长度<d>[67]，定义为所有节点之间最短路径的平均值，即



由上式可知，当出现两个节点之间不可达的情况时，整个网络的平均路径长度为无穷大。为避免此发散问题，则以网络效率 e[67]来反映网络节点间信息发送的平均效率，定义如公式（2.15）。可见，网络效率正相关于节点间距离，即 d 越大，e 越小。



#### **2.2.4 聚类系数**

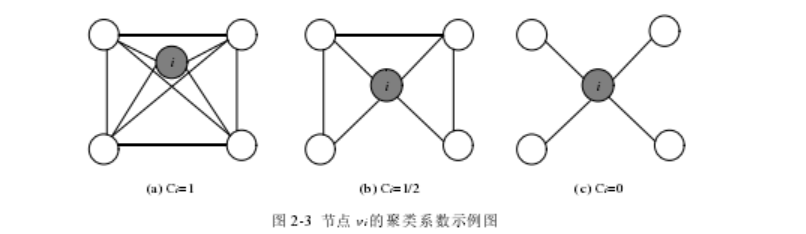
聚类系数（Clustering Coefficient）是用来刻画节点的邻居节点间联系的紧密程度。对于度为 ki的节点 vi，其聚类系数 Ci定义为：



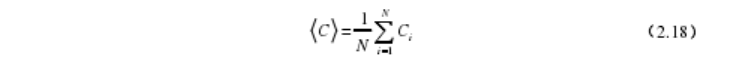
其中，Ei表示节点 vi的邻居节点之间实际存在的连边数，而(ki（ki-1）)/2 则为 ki个邻居节点之间能够产生链接数的理论最大值。 对于以上定义，Watts 等人[62]从另几何图形的角度给出了新的解读，认为节点 vi的聚类系数 Ci也可等价表示为：



可见 Ci∈[0,1]，Ci越高，表明节点 vi邻居节点间的链接越稠密。在不同网络结构情况下，节点 vi的局部聚类系数不同，如图 2-3 所示。



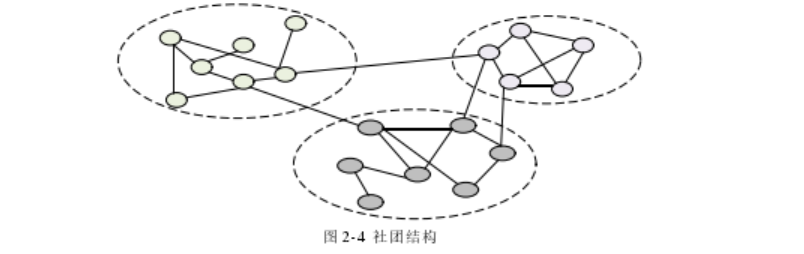
网络的平均聚类系数[66,67]由所有节点的聚类系数的平均值表示，表示整个网络的聚集程度，记为<C>，则



### 2.2.5 社团结构

由网络中的一些节点集聚所形成的结构，被形象地称为社团结构（Community

Structure），如图 2-4 所示。该结构属于复杂网络中观层面的拓扑结构，也是常见且重要的结构特征之一，为分析和理解现实系统提供了独特的视角。 社团结构在诸如社交网络等许多真实场景中都有存在，但理论上至今还没有对其本身的规范定义，只是形成了一个共识，即该结构明显的网络，社团内部的连边相对于跨社团的连边会更为紧密。



## 2.3链路预测理论概述

### 2.3.1 问题描述

考虑一个未加权的无向简单网络G（V，E），其中V是节点的集合，E是链路的集合。不允许多重连接和自连接。对于每对节点x，y∈V，我们分配一个分数Sxy。由于G是无向的，所以分数应该是对称的，比如说Sxy =Syx。所有不存在的连接按照分数降序排列，最上面的链接最有可能存在。本文采用最简单的框架，即直接将相似度设置为得分，所以得分越高，相似度越高，反之亦然。在一些链接预测算法中，分数可能不直接与某个相似性度量相关，而是描述链接的存在可能性，并且在一些其他算法中，分数可以通过目标链接的邻域中的节点对的一些相似性的集成来生成，例如协同过滤方法。

### 2.3.2本文主要算法介绍

在本文中，我们比较了三种相似性指数的预测精度和计算复杂度:CN、Katz指数和一种提出的相似性指数，即局部路径指数LP指数或简称LP。它们的定义和相互关系介绍如下:

1. 公共邻居（CN）指标：

一般意义上，两个节点x和y如果有很多共同的邻居，将来更容易形成链路。对于一个节点x，让Γ（x）表示x的邻居集合。邻域重叠的最简单度量是有向计数：

(1)

其中Q是集合|Q|的基数。很明显Sxy=A2xy，其中A是邻接矩阵，其中如果x和y是直连的，则Axy =1，否则Axy=0。注意A2xy也是连接x和y的长度为2的不同路径数。纽曼在协作网络的研究中使用了这个量，显示了共同邻居的数量和两位科学家未来合作的概率之间的相关性。一些更复杂的指标，如Salton指数、 Jaccard指数、Sørensen指数和Adamic-Adar指数，也可以归类为基于CN的指标。然而，最近，大量的实证分析表明,CN指数比那些复杂的变体表现得更好.因此，我们在这里选择CN作为所有基于CN的度量的代表。虽然CN耗时少，在众多局部指标中表现相对较好，但由于信息不足，其精度赶不上基于全局信息的测度。

1. Katz指数

这种度量是基于所有路径的集合，它直接对路径集合求和，并按长度指数衰减，以给短路径更多的权重。数学表达式如下：

(2)

其中pathsxy<l>是长度为l的所有连接x和y的路径的集合，βl是控制路径权重的自由参数。很明显，βl非常小的路径产生了接近于CN的度量，而长路径时，βl会变得越来越小贡献很小。S矩阵也可以写成（I-βA）-1-I。注意β必须低于矩阵A的特征值最大值的倒数，才能保证方程的收敛性。

1. 局部路径（LP）指标

在CN（共同邻居）指标的基础上考虑三阶邻居的贡献。其定义为：

(3)

其中，S表示相似矩阵，是一个自由参数。显然，当=0时，这个测度退化为CN，如果x和y不直接相连，这就是我们感兴趣的情况，A3xy等于长度为3的不同路径的数量，连接x和y。虽然它比CN需要更多的信息，但它仍然是一种比全局复杂度相对较低的局部度量。

选择这三个指标进行比较，是因为它们都可以归类为路径依赖的相似性，统一形式为，其中对于CN，l = 2；对于LP，l=2，3；对于卡茨，l=1，2，3，...，∞。因为我们只对间接连接的节点对感兴趣，所以Katz指数可以被视为考虑l=2，3，...，∞。

为了测试算法的准确性，观察到的网络链接集合E被随机分成两部分:训练集ET被视为已知信息，而探测集EP用于测试，探测集中的任何信息都不允许用于预测。显然，E=ET∪EP和ET∩EP=∅。在本文中，训练集总是包含90%的链接，剩下的10%的链接构成探测集。

### 2.3.3 性能评价指标

我们使用AUC指标来衡量链路预测算法精确度。AUC可以理解为在测试集中的边的分数值有比随机选择的一个不存在的边的分数值高的概率，也就是说，每次随机从测试集中选取一条边与随机选择的不存在的边进行比较，如果测试集中的边的分数值大于不存在的边的分数值，就加1分；如果两个分数值相等，就加0.5分。独立地比较n次,如果有n’次测试集中的边的分数值大于不存在的边的分数，有n’’次两分数值相等，则AUC定义为：

(4)

显然，如果所有分数都是随机产生的，AUC=0．5。因此AUC大于0．5的程度衡量了算法在多大程度上比随机选择的方法精确。

# 第三章 实验结果及分析

## 3.1实验设置及实验环境说明

虽然真实网络具有复杂的结构属性，如社区结构、混合模式和富俱乐部现象，但作为起点，我们只考虑一个非常简单的模型，为了消除度异质性的影响，我们假设每个节点都有一个相同的度k。在该模型中，每个节点由一个10维向量表征，每个元素在区间（1，1）中随机选择一个实数.该向量表示节点的内在特征，例如对象的属性和人的轮廓。两个节点被认为是相似的，因此如果它们共享许多相似的属性，则很有可能相互连接。因此，我们将两个节点之间的内在相似性定义为对应向量的标量积，即：

(5)

其中是节点x的向量，上标强调这种相似性是内在的，不能在真实系统中观察到。

给定网络大小N和每个节点的度k，这个模型从一个空的网络开始，但是有N个节点，也就是说，每个节点的度为零。在每个时间步长，随机选择一个最小度的节点，一般有一个以上的节点具有最小度。在度数小于k的所有其他节点中，这个选定的节点将以概率1-p连接到最相似的节点，而随机选择的节点以概率p连接。当所有节点都是k度时，此过程将终止。参数p∈[0,1]表示生成链接时的随机性强度，可以模拟为几乎每个真实系统中都会存在的噪声或不合理性。

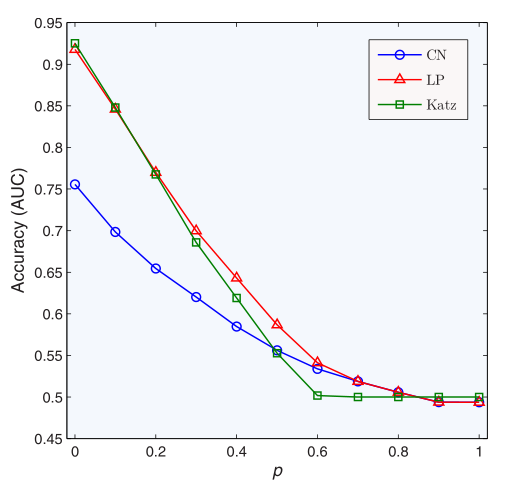


图2-2-1 三个相似性指数的颜色在线预测精度与随机性强度的关系

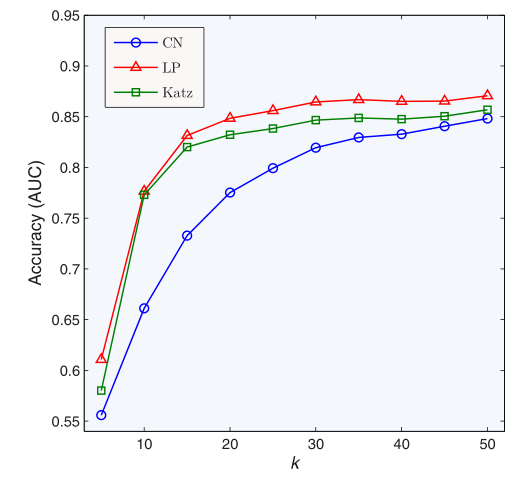


图2-2-2 三种相似性指数的颜色在线预测精度与网络密度的关系

在图2-2-1和图2-2-2中，我们报告了这三个相似性指数的算法准确性的比较。数据点对应于（Katz指数）或（LP指数）的最佳值，以最高精度为准。显然，Katz和LP指数的表现都明显好于简单的CN指数。如图2-2-2所示。当随机性/噪声强度较弱时，LP指数给出的竞争结果与Katz指数相同，而对于高噪声情况，LP指数表现更好。不管相似性指数是多少，一个链接预测算法在更密集的网络能给出更高的精度，这与图2-2-2中观察到的是一致的。在缺乏信息（K比较小），或信息比较多（K比较大）的区域，LP指数的表现略好于CN指数，而在当每个节点的度为真实网络的中间值时，LP指数的表现要比CN指数好得多。

CN索引的性能明显不如LP索引的原因是两个节点对被CN赋予相同相似性的概率很高。也就是说，CN指数的区分度较低，尤其是在相对稀疏的网络中。例如，在N =1000，p=0，k=10的情况下，大约有 5×105个节点对，其中94.01%的分数为零，对于所有非零分数，79.87%的分数为1。如后面所示，实际情况可能更糟，例如，在具有5022个节点的路由器级互联网中，99.59%的节点对被CN分配0分，而对于所有这些非零分，其中91.11%被分配1分。LP指标通过引入包含次近邻的附加信息使相似性更容易区分，从而显著地提高准确性。需要注意的是，如果连接任意两个节点的长度为3的路径的最大数量是Pmax，并且，将给出完全相同的预测。因此，当LP算法的参数不太大时，其预测精度对参数不太敏感。事实上，将设置为一个小的正数，如0.01，可以获得接近最优的精度，通常比实际最优值小1%以下.在表1中，我们将六个实际网络的最优AUC值与通过设置=0.01获得的值进行了比较。为了确定的最佳值，可以先计算邻接矩阵的最大特征值，而的最优值总是小于它的倒数。这样就很容易接近最优。

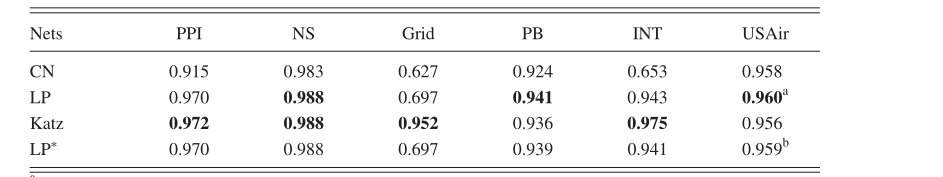


表2-1 三种相似性指数的颜色在线预测精度与网络密度的关系

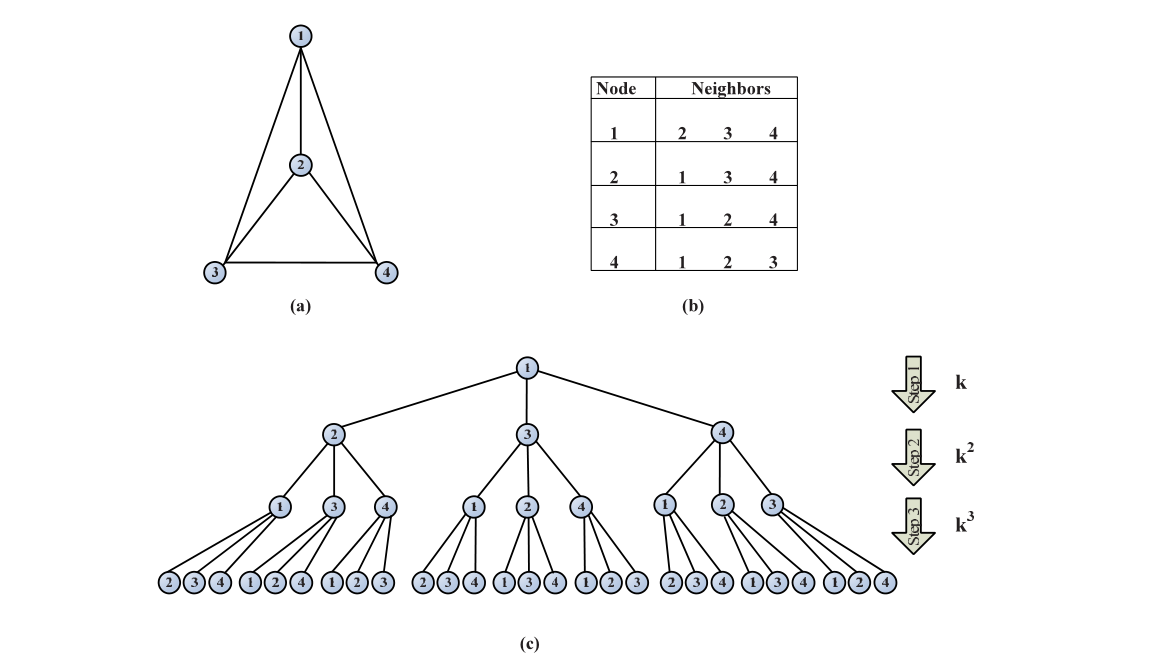


图2-2-3 一个简单的网络示意图

接下来，我们讨论三个相似性指数的计算复杂性。在计算CN索引时，对于由x表示的每个节点，我们首先搜索x的所有邻居，称为步骤1，然后再搜索x的每个邻居的邻居，分别称为步骤2.如果一个节点y在步骤2中出现n次，Sxy=n。由于遍历节点邻域的时间复杂度仅为k，因此计算CN索引的时间复杂度为O(Nk2).类似地，对于LP索引，我们需要做的是更进一步，称为步骤3，分别检查x的每个二阶邻居的所有邻居。如果一个节点y在x的二阶邻域出现n次，在x的三阶邻域出现m次，。因此，计算线性规划指数的时间复杂度为O(NK3)。更进一步进行分析，例如，网络由四个节点组成，如图2-2-3,对于Katz指数，时间复杂度主要由网络节点个数决定，为O(N3)。//todo

除了时间复杂度之外，空间复杂度也是衡量一个算法效率的一个重要考量因素。。在计算CN和LP索引时，所需的内存是O(Nk)的数量级，而对于Katz索引，则是O(N2)的数量级.总的来说，与广泛应用的CN指标和Katz指标相比，LP指标既准确同时也高效。需要相对较少的内存和CPU时间。

## 3.2实验数据集

在本文中，我们考虑了来自不同领域的六个代表性网络:

1. PPI:一个包含2617个蛋白质和11855个相互作用的蛋白质-蛋白质相互作用网络.虽然这个网络没有连接，但它包含92个组件，大多数节点属于巨型组件，其大小为2375。
2. NS:科学家之间的合作网络，他们自己也在发表关于网络5主题。这个网络包含1589个科学家，其中128个是孤立的。这里我们不考虑那些孤立的节点。NS连通性不好。它由268个相连的组件组成，最大相连组件的大小只有379。
3. 三级电网（grid）:美国西部53个州的电力网，节点代表发电机、变压器和变电站，连接它们之间的高压输电线路。该网络包含4941个节点，运行良好连接。
4. 美国政治博客网络（PB）：原始链接是定向的，在这里，我们将它们视为无方向的。PB有1224个节点，巨型组件包含1222个节点。
5. INT:互联网的路由器级拓扑，由Rocketfuel Project 收集.INT有5022个节点，连接良好，而它是一个极度稀疏的网络，平均度只有2.49。
6. 美国航空公司网络:美国航空运输系统的网络，包括332个机场和2126家航空公司。请注意，这里考虑的所有相似性指数，以及除参考文献中报告的优先附着指数之外的那些众所周知的指数。将对位于两个不相连组件中的一对节点给出零分。因此，这里我们只考虑巨型组件，在准备探测集的时候，我们还要确保训练集保持不变，代表一个连通的网络。实际上，每次在移除到探测集的链接之前，我们首先检查这种移除是否会使训练网络断开。表二总结了这些网络的巨型组件的基本拓扑特征。

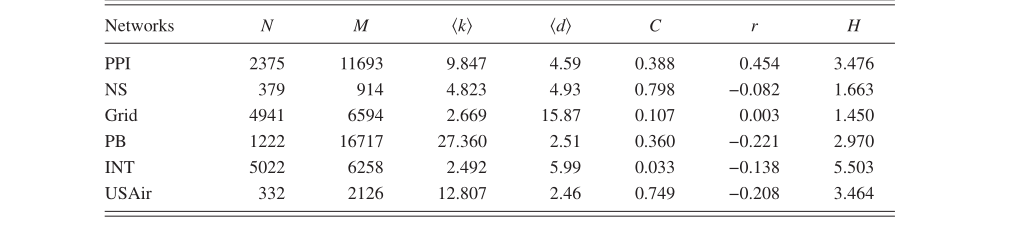


表2-1 三种相似性指数的颜色在线预测精度与网络密度的关系

## 3.3结果与分析

我们将链路预测算法应用于六个真实网络，其精度如表1所示，对应于最高精度的条目用黑色强调。显然，LP指标的表现总是优于CN指标，尤其是对于INT网络，AUC从0.653大幅提高到0.943。除了网格电力网络，LP指数给出的预测和Katz指数一样具有竞争力。电力网格网络是一个强局部化的网络，大多数链路的地理长度较短，因此，网格的拓扑距离<d>=15.87比其他五个示例网络大得多。虽然电力网格网络在地理上是局部化的，但是聚类系数相对较小，并且缺少短循环，因为从工程角度来看，这种循环是冗余的并且效率较低。实际上，在电力网格网络中，当链接被移除时，通常很难找到非常短的路径，例如连接两个端点的长度为2或3的路径。因此，如果链路被移除，仅考虑非常短的路径，CN和LP指数不能细化两个直接连接的节点之间的相关性。此外，我们注意到美国航空网络的最佳值为负。在美国航空网络中，大度数节点密集连接，并共享许多公共邻居。即使没有的贡献，大度数节点之间的链接也被赋予了非常高的分数；因此，附加考虑条件几乎不改变它们的相对位置。考虑两个小的本地机场a和b，这两个机场连接到它们的本地中心机场a’和b’。当然，很多小机场都是a’和b’的公共邻居，a’和b’可以直接相连。如果去掉链接(a，b’)，a和其他节点的相似度都为零。或者，假设h表示一个中心节点，存在（a-a’-h-b’或者a-a’-b’-h）的连接否则，由Sab’（a-a’-h-b’）、由Sab(表示a-a’-b’-y)，以及Sah,这三都是正的。这是由于长度为3的路径的贡献.连接地方小机场和地方中心的环节很多；其中一些被移除，而其他的被保存在测试集中。根据对于上述讨论，由于附加项，移除的链接比不存在的链接具有更低的分数。总之，USAir的非常具体的结构，由枢纽、地方中心和小型地方机场组成的分层组织，使得正数的LP指数比对应于=0的简单CN更差，这也是负数表现更好的原因。

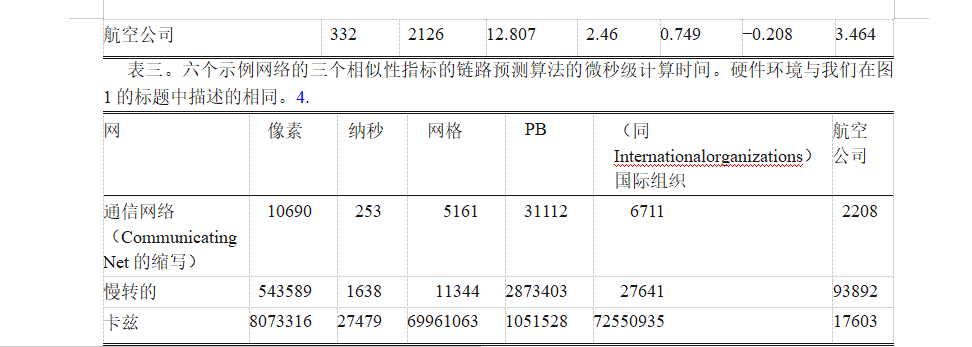


表3 时间对比

表三给出了基于三个相似性指数的链路预测算法的计算时间。显然，CN成本最低。请注意，计算LP指标的计算复杂度对节点的平均度非常敏感，而计算Katz索引的计算复杂度对网络大小非常敏感。因此，与采用Kazt指标的算法相比，采用LP的算法在大规模稀疏网络中具有很大的优势。以INT为例，使用Katz指数的算法运行一天左右，而使用LP指数的算法不到半分钟。由于对计算复杂性的真正挑战总是与大规模的真实网络有关，这些网络大多非常稀疏，所以LP指数比Kazt指数更实用。

# **第四章 结论**

在这篇文章中，我们通过将LP指标与常见的CN指标，Kazt指标进行比较分析，发现LP指指标比Katz指数提供的预测略准确，尤其是在高噪声情况下。我们进一步使用六个有代表性的真实网络来测试三个相似性指数，表明LP指标可以提供与Katz指标相当的精确预测。与Katz指标相比，LP指标需要更少的CPU时间和内存空间，因此更实用。忽略节点度相关性，计算LP指数和Katz指数的时间复杂度分别为o(NK3)和O(N3)。因此，对于巨大的并且稀疏（即非常小的平均度k）的网络，LP的优势是惊人的。

高度准确的预测在实践中非常重要。例如，许多生物网络，如蛋白质-蛋白质相互作用网络、代谢网络和食物网，在实验室或野外发现链接/相互作用的成本很高。如果预测足够准确，而不是盲目地检查所有可能的相互作用，以预先预测已知的相互作用，并专注于最有可能存在的联系，可以大幅降低实验成本。对于其他一些，如网络社会中的友谊网络，很有可能向相关用户推荐尚未存在的链接，作为有希望的友谊的推荐。这些推荐可以帮助用户找到新朋友，从而提高他们对网站的忠诚度。除了实际意义之外，值得强调的是，链路预测的研究还可以提供一些关于结构组织的理论见解。例如，在本文中，关于电力网格网络和USAir的意外结果给出了一些不明显的特定结构特性的证据。

本文只考虑静态网络中的链路预测问题。但是很多真实的网络都是一直在进化的，不同时间创建的链接原则上应该赋予不同的权重。这个涉及时间的链接预测问题很少被研究，当然，值得在将来进行认真的研究.以往相关方向的研究大多只测试真实网络中算法的准确性。这里，我们认为应该使用建模的网络，因为人们可以控制模型中的一些有意义的参数，这些参数在真实网络中不能直接观察到，并且可能存在噪音或不合理的强度。

# 致谢